ACTA PHILOSOPHICA GOTHOBURGENSIA
17

APPLYING UTILITARIANISM

The Problem of Practical Action-guidance

Jonas Gren

ACTA UNIVERSITATIS GOTHOBURGENSIS

It was never contended or conceited by a sound, orthodox utilitarian, that the lover should kiss his mistress with an eye to the common weal.

John Austin, *The Province of Jurisprudence Determined*, p. 97

APPLYING UTILITARIANISM

The Problem of Practical Action-guidance

ACTA PHILOSOPHICA GOTHOBURGENSIA
17

APPLYING UTILITARIANISM

The Problem of Practical Action-guidance

Jonas Gren

ACTA UNIVERSITATIS GOTHOBURGENSIS

Distribution:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222
SE-405 30 Göteborg
Sweden

ISBN 91-7346-508-9
ISSN 0283-2380

Printed in Sweden by
Intellecta DocuSys, Göteborg 2004

ACTA PHILOSOPHICA GOTHOBURGENSIA
ISSN 0283-2380

Editor: Dag Westerståhl

Published by the Department of Philosophy of the University of Göteborg

Subscription to the series and orders for single volumes should be addressed to:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, SE-405 30 Göteborg, Sweden

VOLUMES PUBLISHED

1. MATS FURBERG, THOMAS WETTERSTRÖM and CLAES ÅBERG (editors): Logic and Abstraction. Essays dedicated to Per Lindström on his fiftieth birthday. 1986. 347 pp.

2. STAFFAN CARLSHAMRE: Language and Time. An Attempt to Arrest the Thought of Jacques Derrida. 1986. 253 pp.

3. CLAES ÅBERG (editor): Cum Grano Salis. Essays dedicated to Dick A. R. Haglund. 1989. 263 pp.

4. ANDERS TOLLAND: Epistemological Relativism and Relativistic Epistemology. Richard Rorty and the possibility of a Philosophical Theory of Knowledge. 1991. 156 pp.

5. CLAES STRANNEGÅRD: Arithmetical realizations of modal formulas. 1997. 100 pp.

6. BENGT BRÜLDE: The Human Good. 1998. 490 pp.

7. EVA MARK: Självbilder och jagkonstitution. 1998. 236 pp.

8. MAY THORSETH: Legitimate and Illegitimate Paternalism in Polyethnic Conflicts. 1999. 214 pp.

9. CHRISTIAN MUNTHE: Pure Selection. The Ethics of Preimplantation Genetic Diagnosis and Choosing Children without Abortion. 1999. 310 pp.

10. JOHAN MÅRTENSSON: Subjunctive Conditionals and Time. A Defense of a Weak Classical Approach. 1999. 212 pp.

11. CLAUDIO M. TAMBURRINI: The 'Hand of God'? Essays in the Philosophy of Sports. 2000. 167 pp.

12. LARS SANDMAN: A Good Death: On the Value of Death and Dying. 2001. 346 pp.

13. KENT GUSTAVSSON: Emergent Consciousness. Themes in C.D. Broad's Philosophy of Mind. 2002. 204 pp.

14. FRANK LORENTZON: Fri vilja? 2002. 175 pp.

15. JAN LIF: Can a Consequentialist Be a Real Friend? (Who Cares?). 2003. 167 pp.

16. FREDRIK SUNDQUIST: Perceptual Dynamics. Theoretical Foundations and Philosophical Implications of Gestalt Psychology. 2003. 261 p.

17. JONAS GREN: Applying Utilitarianism. The Problem of Practical Action-guidance. 2004. 160 pp.

ISBN 91-7346-508-9

ACTA PHILOSOPHICA GOTHOBURGENSIA
17

APPLYING UTILITARIANISM
THE PROBLEM OF PRACTICAL ACTION-GUIDANCE

Jonas Gren

AKADEMISK AVHANDLING

för avläggande av filosofie doktorsexamen i praktisk filosofi,
som med tillstånd av humanistiska fakultetsnämnden
vid Göteborgs universitet
framläggs till offentlig granskning
lördagen den 25 september 2004 10.00
i Lilla Hörsalen, Humanisten, Renströmsgatan 6, Göteborg.

Abstract

Title: Applying Utilitarianism. The Problem of Practical Action-guidance.

(Acta Philosophica Gothoburgensia no 17)

ISSN 0283-2380, ISBN 91-7346-508-9

Language: English, 160 pp.

Author: Jonas Gren

Doctoral Dissertation 2004 at the Department of Philosophy, Göteborg University

This dissertation addresses the question of whether act-utilitarianism (AU) can provide practical action-guidance. Traditionally, when approaching this question, utilitarians invoke the distinction between criteria of rightness and methods of decision-making. The utilitarian criterion of rightness states, roughly, that an action is right if and only if there is nothing else that the agent can do that has a better outcome. However, this criterion needs to be supplemented, it is said, with some description of a strategy that allows an agent to reach decisions that approximate the utilitarian idea – a method of decision-making. The main question in the essay is if any such method can indeed be justified on the basis of AU. I argue that the justification of a method of decision-making depends on the extent to which it has two different features: practicability and validity. Roughly a method of decision-making is practicable if an agent trying to adhere to the method will succeed in doing so. A method of decision-making is valid if adhering to the method makes the agent approximate the overall goal of AU. I then proceed by examining whether it is possible to justify a belief to the effect that any of the various candidates of methods of decision-making that have been proposed in the literature have these features. My main conclusion is negative. No proposed method of decision-making can be shown to satisfy these desiderata to a sufficient degree. In the final chapter the implications of this conclusion are examined. Does this mean that we cannot justify a belief in AU? Does it mean that AU is false? My conclusion is that whether or not this shows that AU is false depends on what meta-ethical view is the most plausible one. I also present a tentative way of justifying a belief in AU.

Keywords: act-utilitarianism, action-guidance, maximising expected utility, criterion of rightness, methods of decision-making, practicability, validity, secondary rules of conduct, justification.

Acknowledgements

There are many persons who have contributed to my completing this thesis. I began this undertaking under the supervision of my first tutor, Torbjörn Tännsjö, whose comments has improved much of the thoughts developed in the process of writing. I have learnt a lot from his approach to philosophy. My second tutor, Folke Tersman, has also helped me a lot in the process of improving and refining my ideas. I am indebted to them both.

I would also like to thank my colleagues and the participants of the seminars at the philosophy department at Gothenburg University, e.g. Petra Andersson, Jan Lif, Christian Munthe, Pia Nykänen, Joakim Sandberg, Lars Sandman, Bolof Stridbeck, Claudio M. Tamburrini and Anders Tolland. I am particularly indebted to Bengt Brülde and Ragnar Francén for giving me extensive critique and constructive suggestions, in the final phase of completing this essay.

One person's assiduous and painstaking commenting, creative suggestions, support and willingness to discuss my ideas with me stand out in particular. To my colleague and dear friend Niklas Juth, without whose support this essay would be much worse, I am deeply indebted.

I am also grateful to Angus Hawkins for checking my English. Any remaining mistakes are to be attributed to the author.

Finally, I would like to extend my gratitude to my family, Karin, Lars and Anna Gren, as well as Stina Nordvall, for their support over the years. I dedicate this essay to my wife, and the love of my life, Karin Håkanson, who has stood by me throughout.

Table of content

Chapter I
Introduction

1.0 The problem of practical action-guidance

Imagine that you have just finished a course in moral philosophy at university. As the subject of the course was applied ethics, much of the discussion came to revolve around utilitarianism. These discussions made you appreciate many of the advantages of this theory. Among other things, you noticed its apparent simplicity and scope. You also saw the theory's focus on wellbeing, as the only thing that matters, as a desirable feature. To cut a long story short, you came to accept 'act-utilitarianism', as an account of what makes right actions right. Now, on the first day of your summer holiday, you have just finished breakfast. Suddenly, a question which you have asked yourself on previous occasions, but which never before has seemed so urgent as now, begins to plague your mind: "How ought I to act in practical situations of choice?" Due to your recent acceptance of hedonistic act-utilitarianism, this question appears for the first time to have a definite answer. You realise that you ought to maximise utility. But your initial exaltation quickly changes. Although you are now able to answer *this* question, you realise that another question is at least as important as the former: "Just *how* am I supposed to act in order to maximise utility?"

Which kinds of questions should moral theories answer? There is no uncontroversial answer to this query. Different traditions in ethics take different questions to be the most important ones. Issues regarded to be of great importance in one tradition are regarded as unimportant, peripheral or uninteresting in other traditions. At least for some virtue ethicists, the question of what makes right actions right is of only marginal importance compared to the question of what is a good life. A Kantian ethicist would place great importance on the question of what constitutes a morally worthy motive, a question that a utilitarian would regard as at best a peripheral one.

For a utilitarian, the crucial question is: "What makes right actions right?" It is the answer to this question that *defines* act-utilitarianism (henceforth AU) and that contrasts it with other theories, both those in the consequentialist tradition, and those that belong to other traditions. More specifically, I shall conceive of AU, not as a full-fledged ethical theory, but rather as a general schema representing the *structure* common to a particular family of consequentialist theories. However, it is worth noting already at this early stage that since AU is a particular brand of consequentialism, what will be said about AU in the following may very well apply to other brands of

consequentialism as well. Moreover, other types of moral theories that purport to answer the same basic question as consequentialist theories do may also be affected by the conclusions and arguments put forward in this essay.

There are, however, other questions that we want ethical theories to answer. We want them to be *action guiding*. The importance of ethical theories does not stem exclusively from our want of a general answer to the question of what characterises right actions, but also from a want of action-guidance in practical situations of choice.

> […E]thics is not an ideal system that is noble in theory but no good in practice. The reverse of this is closer to the truth: an ethical judgement that is no good in practice must suffer from a theoretical defect as well, for the whole point of ethical judgements is to guide practice. (Singer, 1993, p. 2)

The example of the student above illustrates what this essay is about: AU and the problem of practical action-guidance. My aim in this essay could roughly be stated as an examination of the relation between theory and practice in relation to this particular theory.

2.0 The main issues

The question that provides the focus of this essay can be stated as follows. Is it possible to justify decisions in particular situations of choice with reference to AU? More specifically, is it possible for an agent that is normally equipped from a cognitive point of view to determine whether or not a particular action satisfies AU's criterion of rightness? Could she ever be (epistemically[1]) justified in believing that an action is right, given AU? I shall say that, if the answer is "no", then AU fails to provide practical action-guidance. Of course, "justified in believing" is a vague phrase, as is "normally equipped from a cognitive point of view". However, one of the main conclusions of this dissertation is that even given pretty weak conditions regarding when a person is justified in believing something, and even given a pretty generous view regarding the cognitive equipment of a normal person, AU does not give practical guidance in the indicated sense.

It seems to me that the question of practical guidance has been neglected in the discussion about AU. Most of the discussion has instead focused on how to formulate the version of AU (conceived of as a criterion of rightness) that is most advanced and refined from a purely theoretical point of view. This is strange, in my view, since practical action-guidance surely is, at least *prima facie*, the Alpha and Omega of normative theories. Indeed, one may ask: What is the point of norms that are not action guiding in practice?

[1] The term "epistemically" is meant to indicate that the justification sought for is the justification held to be a prerequisite of knowledge rather than some form of "pragmatic" justification (justification in terms of some practical end).

Perhaps there could still be such a point. Even if a normative theory cannot be applied in practice there may still, perhaps, be reasons to examine it. Indeed, as I shall argue in chapter V, such a theory might even be justified. Still, one cannot ignore the fact that people turn to ethics at least in part also for something else. They want more than just abstract models that are to be studied for theoretical purposes. Many people, and especially those who are outside of professional philosophy, would surely agree that if we believe that the criterion of rightness is plausible, or even true, then it should have some impact on our decision-making in actual life[2].

I shall approach the problem of practical action-guidance in relation to AU by imagining an agent who accepts AU (an AU-agent) and who is determined to live by her belief. The question is if the decisions made by such an agent could be rationally defensible given her goal[3]. In this context, advocates of AU usually introduce the distinction between "criteria of rightness" and "methods of decision-making". Up till now, AU has been conceived as a criterion of rightness, i.e. as an account of what makes right actions right. However, someone who wants to live up to the ideal set by AU also, it is held, needs a "method of decision-making", i.e. some form of strategy or method for reaching decisions. One way in which I shall address the problem of practical action-guidance is to ask whether any given method of decision-making could be rationally justified on the basis of AU.

Advocates of AU have characteristically argued that the appropriate method of decision-making might not be to "reason as a utilitarian" (a phrase that will be clarified below). Instead, they should reach their decisions through applying common sense rules or certain 'rules of thumb' or something of the like. One of the aims of the essay is to scrutinise these claims, and to consider the ideas about the appropriate methods of decision-making that have been proposed more closely. Sadly, as we shall see, the suggestions made are often imprecise, obscure and sketchy, and do not provide believable answers to the question of practical action-guidance.

To repeat, the question that stands at the centre of the present essay is: Can an agent who is normally equipped from a cognitive point of view ever be justified in believing that a particular action is right according to AU? The main conclusion is that the answer is "No". Where does this leave AU? Does it indicate that we should reject AU, or that AU is false? These questions are addressed in the final chapter. My main conclusion in that chapter is that although the sceptical conclusion means that AU fits badly with some views

[2] Of course it is possible that the 'impact' is that we should not accept AU's criterion, or not let our acceptance of AU's criterion have any impact on how we decide in practical matters. But this is just another kind of impact, although on a higher level of application. We would, however, like to be justified in our belief that one or the other of these possibilities is the best strategy. Cf. Sidgwick (1907), p. 490.

[3] "Rationally defensible" should be distinguished from "morally right". Thus, a decision can be rationally defensible given the aim set by AU even if it is not right according to AU.

about what it could mean for a normative theory to be true, it does not affect other combinations.

3.0 Limitations

In this thesis, my interest is restricted in several ways. I will not engage in any normative debates between AU and some alternative ethical system. I accept AU as an account of what makes right actions right until chapter V, where I discuss the possibility of justifying a belief in AU.

There are many questions concerning AU and practical action-guidance. AU, being a version of consequentialism, can be applied to many different things. As Derek Parfit points out:

> Consequentialism covers, not just acts and outcomes, but also desires, dispositions, beliefs, emotions, the colour of our eyes, the climate and everything else. More exactly, C [consequentialism] covers anything that could make outcomes better or worse. (Parfit, 1984. p. 25)

AU answers the question which actions we ought to perform. It also answers the question of what motives or dispositions we ought to have or cultivate. We ought to have the dispositions and motives that make the outcome best. It answers questions about which laws, punishments, constitution, healthcare system and taxes we ought to have. An important limit to my inquiry should be noted. I will discuss different proposals for methods of decision-making for AU, i.e. what AU has been taken to say about how moral agents ought to decide what to do in concrete situations of choice. These methods include trying to maximise expected utility and following secondary rules of conduct.

One might ask at what level of application AU is most suitable. Should AU be applied at the political[4] level of society or should it (also) be applied at the personal or private level? I will not try to answer this question. In what follows my discussion is posed in terms of how, given the truth of AU, an AU-agent should decide what to do. I limit the investigation by exclusively discussing AU in relation to decision-making of individual moral agents, i.e. persons. I do not discuss the question of how *collective* agents, e.g. groups of persons, ought to act. Of course, this limitation is not innocent, but some limitation is required in order for the subject to be manageable within the scope of a doctoral dissertation.

4.0 The plan of this essay

In chapter II, the main problem is elaborated. I state the act-utilitarian criterion of rightness of actions and discuss the distinction between such

[4] Traditionally, utilitarian writers such as Jeremy Bentham and John Stuart Mill seem to advocate utilitarianism primarily as a theory regulating political decisions, i.e. how society ought to be designed. Utilitarianism conceived of in this way primarily regulate the assignment of individual legal rights and legislation in a society. Also cf. Goodin (1995).

criteria and methods of decision-making. It is argued that the distinction is a valuable and necessary tool for making act-utilitarianism coherent, let alone normatively plausible. The argument takes a negative form, showing that a failure to uphold the distinction will make AU incoherent. Arguments for resisting a direct application of AU's criterion of rightness are presented. It is argued that trying to maximise utility, as a deliberative approach, is not a plausible method of decision for AU. I consider different *desiderata* for any method of decision-making for AU. These desiderata are of three different kinds, i.e. *methodological*, *practical* and *normative*. My main interest, this being an essay on AU and the problem of practical action-guidance, is in the latter two, *viz.* practicability and validity. These two desiderata are presented, developed and defended. This sets the stage for the discussion in chapters to come, i.e. if an AU-agent could be justified in adopting any of the methods of decision-making suggested by the utilitarians given that she wants to approximate the goal given by AU. If a method could be shown to meet these desiderata to a higher degree than any alternative method, then trying to adhere to this method would, to some extent, be justified.

In chapter III, the idea of trying to maximise expected utility is discussed as a possible method of decision-making, or part thereof, for AU. I discuss different possible interpretations of what I call *the deliberative approach*, i.e. trying to maximise expected utility as a method of decision for AU. The conclusion is that there is little or no evidence suggesting that a plausible interpretation of this approach to moral decision-making meets the desiderata stated in chapter II to a satisfying degree. Reasons, presented in the literature on the subject for believing this method to meet the desiderata are presented, but found to be very weak. The method, when it is given a plausible interpretation, probably scores low in terms of the practicability desideratum. Reasons for believing the method to do better in terms of validity are presented and criticised. No utilitarian thinker defends an unrestricted adherence to this approach. Utilitarians do, however, give this method a prominent place in their respective approach to decision-making. The different restrictions on this method suggested by Henry Sidgwick, George. E. Moore, Richard M. Hare and Torbjörn Tännsjö raise questions about the role of secondary rules as a part of a method of decision for AU. This is the theme of chapter IV.

In chapter IV, the role of secondary rules of conduct within a method of decision-making for AU is discussed. The idea is that AU can justify adherence to secondary rules as a practical guide to action. In different forms, all proponents of consequentialism endorsed this idea. Generally, adhering to these secondary rules is probably a relatively practicable approach to decision-making (given that these rules are determinate enough and that they do not conflict with one-another). Their validity, relative to AU is more doubtful though. The different approaches to this issue can be placed on a

scale ranging from giving the rules a more secluded place within the method, to giving them a more conspicuous place. The approach defended by Moore could be called the *conservative* approach to moral decision-making for AU in that it prescribes categorical adherence to (certain kinds of) common sense rules. Hare's 'two level approach' to moral decision-making, as well as Sidgwick's and Tännsjö's method, occupy more moderate positions in between the extreme positions on this scale. The upshot of this chapter is that the more salient the place of secondary rules of conduct is given, the higher the practicability of the method. The restricted deliberative strategy seems to meet the practicability *desideratum* to a greater extent than does the unrestricted deliberative strategy. The validity of the different versions of the restricted deliberative strategy, however, is difficult to evaluate. The conclusion is that good reasons for believing that any, hitherto presented method of decision, meets both desiderata to a satisfying degree, is hard or impossible to come by. Furthermore, there seems to be some tension between the two *desiderata*. They sometimes seem to pull in opposite directions, or so I shall argue.

In chapter V the implications of the conclusions of the foregoing chapters are examined. What has been understood as an internal problem or challenge for AU, in the forgoing chapters, can also be understood as an argument against AU. The difficulties with putting AU into practice have, so to speak, lead opponents of AU to use this 'weakness' of the theory as an argument for rejecting the theory. I will consider several aspects of this argument. One strong and direct version of this argument is the following:

1. If a normative theory cannot help us guide our actions in practical situations of choice, then the theory is false.
2. AU cannot help us guide our actions in practical situations of choice.

_____

3. AU is false.

I will examine both premises of this argument. My conclusion is that there are good reasons for accepting the second premise (even given an interpretation of 'help' and 'guide our actions' which is sufficiently strong as to make the issue interesting). What about the first premise, then? Discussing this premise shifts the perspective, from an internal to an external evaluation, relative to AU. My conclusion is that the relevance of 2 to 3 depends on which account of the grounds of morality one accepts. Given moral realism, the fact that AU cannot help us guide our actions in practical situations of choice does not give us reason to think that AU cannot be true. Given certain versions of moral constructivism, however, it does. In other words, given 2, the claim that AU is the true normative theory squares better with realism than with (those forms of) constructivism. Or so I shall argue anyhow.

That AU can be true is one thing. That it can be justified is quite another. It has been argued that although 2 does not exclude the possibility of AU being true, it does exclude the possibility of justifying a belief in AU. This claim will also be considered in chapter V, where I outline a way of justifying a belief in AU even granted the truth of 2. This defence is the most speculative part of the thesis.

In the Appendix, I consider some other normative theories and their ability to guide actions in practical situation of choice. If lack of practical action-guidance is something that afflicts other theories as well, then this problem does not give us any reason to prefer one of these theories to AU. The conclusion is that the theories I consider do not seem to do much better than AU in this respect.

Chapter II
AU and Practical Action-guidance

In this chapter I start by giving an account of AU. On the basis of this account, I then proceed to recapitulate and elaborate some of the arguments for the impossibility of what will be called a "direct" or "crude" way of applying AU. Surely, it is this impossibility that at least partly explains why no utilitarian thinker, modern or classic, has advocated this particular way of applying AU. In this context, I introduce and defend the distinction between criteria of rightness and methods of decision-making. Then I go on to propose certain desiderata for methods of decision-making.

1.0 AU as a criterion of objective rightness of actions

The most common way to formulate AU in a more precise manner is to give an account of what makes right actions right, i.e. to state a criterion of rightness of actions. The formulation of AU's criterion of rightness that will form the basis of my discussion is:

> AU: A particular action is right if, and only if, there was nothing the agent could have done instead in the situation, such that had the agent done this, the universe (on the whole) would have been better[1].

For reasons of simplicity I will use the phrase 'an action is optimific' as an abbreviation for the more cumbersome expression 'there was nothing the agent could have done instead in the situation, such that had the agent done this, the universe (on the whole) would have been better'. Thus AU claims that an action is right if, and only if, it is optimific, and that it is its optimificness that makes it right. Moreover, every action that is not right is wrong. Notice that I will in the following use the term 'action' so as to denote a particular action[2] that an agent can perform[3] in a particular situation.

---

[1] AU is stated in this way, without determining what it is that has intrinsic value, because my argument will be applicable to all theories with this common structure. All versions of act-utilitarianism, hedonistic-, preferentialist-, and ideal-act-utilitarianism share the features which I will use in my arguments. The criterion is a more unspecific version (in that it does not specify what it is that possesses intrinsic value) of Torbjörn Tännsjö's criterion of rightness of actions in Tännsjö (1998).

[2] A few words on what kind of things 'actions' are might be appropriate. Actions may be conceived of as either abstract or concrete entities. On the former view actions are states of affairs or facts, such as the fact *that* I am drinking coffee, *that* I go to bed. These actions may be performed in different ways (they can have different "versions"). I can drink the coffee slowly or I can drink it rapidly. Another view is that actions are identified with concrete parts of space-time, e.g. the actual movement of my body during a specific time interval. According to this view every action is unique and does not have different versions. I will not take a

As stated now, AU contains several problematic notions all of which need clarification. However, I take 'action', 'agent' and 'could have done instead' in an everyday, non-technical sense[4]. Notions such as 'alternatives' and 'outcomes' will be examined in the course of this investigation, since these notions play a role within the context of formulating methods of decision-making for AU. Even 'better' will, for reasons already stated, be largely left undetermined. In many examples, as well as in my discussion on the possibility of justifying a belief in AU in chapter V, I will assume a hedonistic understanding. Still, the general argument of this essay is applicable also if 'better' is given a preferentialist, or a more ideal understanding[5]. AU implies for every action, past, present or future, whether it was, is or will be right or wrong.

Notice that AU is a version of *act*-utilitarianism[6]. I have stated AU's criterion of rightness of actions in an actualistic form. There is also a *probabilistic* interpretation. AU is then stated in terms of *expected* outcomes, rather than in actual outcomes[7]. Several of the arguments put forward in chapter III against taking the principle of maximising expected utility to be a

---

stand in this controversy though. What I will have to say about AU and the problem of practical action-guidance will apply, *mutatis mutandis*, to both theories. When it comes to methods of decisions, when the agent consider different possible actions and wonders which of the alternatives she should perform, the different alternatives are represented by the agent as abstract, more or less precise descriptions (which, on the 'concretist' interpretation might be taken as descriptions of possible concrete actions). Cf. Carlson (1995) and Tännsjö (1998) for a discussion of this.

[3] There have been some controversies regarding what it means that an action is 'performable'. I will not go into this debate. It suffices for my purposes in this thesis to say that a particular action is performable by an agent in a situation if it holds that the agent would perform the action if she wanted to do so. For a discussion of this notion, cf. Carlson (1995).

[4] I am primarily concerned with intentional actions. Agents are those kinds of entities which can perform actions. 'Could have done instead' is obviously a tricky affair, an analysis of which would involve questions of free will and its relation to determinism. I will ignore this issue. I will take 'could have done (something else) instead' to imply that the agent had an alternative, which she would have performed instead if she had wanted to do so. In discussing methods of decision-making for AU, I will consider different ways of representing the relevant alternatives in a situation of choice. There is no uncontroversial interpretation of 'alternative', however. For a discussion of this, cf. Bergström (1966) and Carlson (1995). I will not take a stand on this controversial issue. My aim in this thesis is not to state and defend the most plausible version of AU. I want to discuss the possibility of justifying methods of decision-making for theories with the common "consequentialist" structure of AU. My arguments are (probably) applicable to all plausible interpretations of this notion.

[5] One comment is appropriate at this point, however. Sometimes a distinction between 'intrinsic' and 'extrinsic' value on the one hand and between 'final' and 'conditional' or 'instrumental' value on the other. I will use the term 'intrinsic' throughout, to refer to the value that according to AU ought to be maximised. In this I follow the writers that I discuss in this essay. In terms of the distinctions mentioned above, I take 'intrinsic' to imply that the value is also 'final'. For a discussion of these distinctions see Brülde (1998).

[6] Another form of utilitarianism is *rule*-utilitarianism. This kind of utilitarianism makes the rightness of an action depend, not on the action's own outcome, but on what would happen if a rule prescribing this action were observed by everyone. If the general observance of this rule were optimific (compared to the general observance of another rule prescribing some other action) the action in question is right. It has been pointed out that justifying trying to adhere to particular methods of decision-making has similarities with this kind of utilitarianism. "[…I]t takes a rule-utilitarian argument to carry us from the premise that it pays in the long run to try to maximise expected happiness, if this premise could be established, to the conclusion that, in a particular case, we ought to do so." (Tännsjö, 1998, p. 24)

[7] Cf. e.g. Jackson (1991), Bennett (1995) and Gruzalski (1981).

reasonable *method of decision* for a proponent of actualistic AU, apply also to these probabilistic interpretations.

It has been suggested that the distinction between the actualistic and the probabilistic version of AU does not matter much:

> In practice, however, there is little difference between actual-outcome utilitarianism and expected-outcome utilitarianism. Because we are not omniscient and never know for certain the exact outcomes of the various actions we could perform, actual-outcome utilitarians will say that the reasonable way for us to proceed is by trying to maximize probable or expected happiness. Doing this may result in our doing the wrong thing, but it is safe to assume that if our actions maximize expected utility, then we will succeed in producing the most happiness over the long run. Expected-outcome utilitarians, on the other hand, say that maximizing expected utility is not only the reasonable way to proceed; it is also the standard of right and wrong. (Shaw, 1999, p. 30)

Shaw suggests that both the probabilistic and the actualistic versions of AU yield the same method of decision-making. This may or may not be true. I suspect that, even if it is not true, much of the following discussion still applies to both versions. In any case, in what follows, I will focus on the actualistic interpretation of AU, and ignore the probabilistic version[8]. Having said this, let us turn to the main line of argument.

2.0 AU and the problem of practical action-guidance

The problem with AU is not that it doesn't tell us what to do. It does give us a specific instruction, i.e. to maximise utility. It is just that it is extremely difficult to determine whether one's proposed actions satisfies this criterion. This problem can be illustrated by reference to the following analogy: We know that one person can correctly be described as "Olof Palme's murderer". We can state a criterion of success for the work of the special task force of the Swedish police currently assigned to solving this case. Their work is successful if, and only if, they manage to catch Olof Palme's murderer. This, however, is not of much help for the detectives working on the taskforce when trying to catch him (or her). They need *sound methods of conducting a criminal investigation*. In a similar way, what we, as moral agents, need is an

[8] There are well-known normative problems with probabilistic AU as a criterion of rightness of action. Sometimes things do not turn out as they were supposed to. Suppose that, against all odds, when I turn on the coffee-machine at work, someone has turned the 'on' button into a trigger for a nuclear bomb, which wipes out the city I live in. Suppose further that it would have been better had the bomb not detonated. What ought we to say about my pushing the 'on' button? Well, as the argument for the actualistic interpretation of AU goes, my action was wrong. Had I (and nobody else for that matter) not pressed the button the outcome would have been better (let us for the sake of the argument imagine that pushing and not-pushing are my only alternatives in the situation). According to the probabilistic interpretation of AU, however, my action was right (we can reasonably suppose). Which theory gives the most plausible answer here? I think that actualistic AU gives the right answer. This is why I discuss this version of AU in this thesis. I will not argue this point further, though.

answer to the question of what we ought to do under a *practically useful description*. The utilitarian criterion of rightness pinpoints all right actions under one description, i.e. it points out the property of right actions that makes them right. The problem is, of course, that this description does not help us to *determine* which of the actions open to us is right. This suggests that to obtain practical guidance, it is not enough to find out what conditions right actions must satisfy. One also needs advice as to how to determine when an action in fact satisfies these conditions. That is, besides a criterion of rightness one needs a "method of decision-making".

The introduction of the distinction between criteria of rightness and methods of decision-making is a theoretical move that raises a large complex of issues and that will be the focus of interest in the rest of this chapter. In the following sections (3.0-3.1) I consider what happens if the distinction is not upheld. The problems that arise provide arguments in favour of upholding the distinction. After this exposition, a short history of the distinction is presented. It is argued that it, or at least its substantial idea, can be traced back at least as far as to the "fathers" of utilitarianism, e.g. Jeremy Bentham and John Stuart Mill. Having come that far, I will then move to comment on some consequences of upholding the distinction. Having separated the criterion from methods of decision-making, I turn to formulate *desiderata* that a method of decision-making ought to live up to in order for agents to be justified in trying to adhere to it. Let us begin with considering different problems that arise if we do not uphold the distinction between criterion of rightness and method of decision-making.

3.0 Direct applications of the criterion of rightness

This section deals with an extreme version of what has been called a "direct" or "crude" way of applying AU[9]. It is time to characterise what I will call an *AU-agent*, a character that will be employed in the discussions to come. This agent has the following characteristics: She accepts AU (as a criterion of rightness)[10]. Furthermore, she wants, and is prepared to do what it takes[11], to realize as much positive over negative utility as she can, i.e. she wants to approximate the overall goal of AU. One might ask why I characterise the agent in this way. It could be maintained that this is especially problematic due to the fact that this thesis deals with *practical action-guidance*. There is, probably, no actual human agent that satisfies these characteristics. Well, the

[9] Cf. Hare (1981), p. 36.

[10] Of course, as has been pointed out many times, perhaps it would be for the best if utilitarianism were not accepted. Accepting utilitarianism might have bad consequences. I will not explore this possibility. Suppose that this is true. Then designing a method of decision-making, without accepting utilitarianism, yet designing it in such a way that it makes agents trying to adhere to the method perform actions satisfying AU, surely puts the designer in a somewhat awkward situation. I will assume that at least when designing methods of decision-making, accepting utilitarianism will not have detrimental effects.

[11] This agent is such that she is *prepared*, given that she would believe this to be the best way of maximising utility overall, to become a 'pure do-gooder'. Cf. Parfit (1984), p. 27.

rationale for taking an AU-agent as the point of departure is this. In examining the possibility of justifying a method of decision from the point of view of AU, we must avoid begging any questions relative to AU. "AU cannot help us guide our actions because we refuse to believe that AU is true!" or "AU cannot help us guide our actions because we refuse to do what AU tells us to do!" constitutes poor reasons for believing AU incapable of providing practical action-guidance. If AU cannot even help guide the actions of an AU-agent, AU seems even more incapable of helping guiding the actions of less 'cooperative' agents.

Suppose that an AU-agent, call her *A*, decides to try to do what AU requires of her. Any one of the following goals can then be taken to be consistent with A's beliefs and desires:

1. To try to perform optimific actions.
2. To try to perform actions that maximise expected value.

Both these approaches have certain things in common, besides being possible goals for A when she decides what to do. Each of them have to be supplemented with an account of how to deliberate, i.e. some account of how alternatives are to be identified and how outcomes are to be estimated or determined. For example, on approach 1, A tries to determine which action is optimific and then tries to perform it. The process of determining which alternatives are open to A in a situation and the estimation of what their different outcomes, as well as their values are, is commonly referred to as 'calculating' or 'deliberating'. This picture has become something of a caricature of how utilitarian moral agents make their moral decisions. It is this 'calculating'-approach that critics of utilitarianism often suppose that utilitarian agents are always committed to. As a reply to this claim, supporters of AU will point to the distinction between criteria of rightness and methods of decision. The possibility of providing such a reply is certainly one part of the story that explains and justifies the popularity of the distinction among utilitarians. However, this in no way exhausts the reasons one may give in its favour.

Consider the second goal stated above. Entertaining this goal, in practical situations of choice, seems to be a distinct approach. In chapter III, different accounts of this approach will be examined.

However, let us return to the problems associated with the idea of applying AU directly and to the corresponding reasons for maintaining the distinction between a criterion of rightness and a method of decision-making, especially to problems that can be subsumed under the general problem of deliberation.

3.1 Regress arguments

In his (1971) paper, R. E. Bales presents a number of 'strong' arguments directed against act-utilitarianism. These arguments are 'strong' in that they go beyond the mere practical difficulty of trying to figure out what actions are optimific and, instead, turn on theoretical problems resulting from these practical difficulties. One of these arguments–of the 'strong' type–runs as follows:

> We consider a case […] in which decision is postponable. For the sake of simplicity, we begin by supposing that two acts, *A* and *B*, are open to the agent. Which should he perform? If the agent is a consistent act-utilitarian, the argument goes, he will estimate and compare the probable consequences of *A* and *B* and perform the one with the better probable consequences. In brief, he will calculate. But the act of calculating is itself an act which the agent may or may not choose to perform. Thus, a third act, *C*, the act of calculating, has entered the picture. Shall the agent, then, simply perform *A*, or shall he perform *B*, or shall he perform *C*? If the agent is a consistent act-utilitarian, these alternatives, too, provide an occasion for calculating, and a fourth alternative presents itself, *D*, which is the act of calculating the probable consequences of *A*, *B*, and *C*. But of course *D* is an alternative itself subject to calculation, and the agent is caught in a vicious regress. (Bales, 1971, p. 258)

Bales favoured way out of this predicament is to distinguish between criteria of right action and methods of decision-making. The argument establishes, not that act-utilitarianism fails as a criterion of rightness, but that the method of decision-making here considered, i.e. trying to assess directly the utility of every different alternative, fails on behalf of giving rise to an infinite, and vicious, regress. There are several versions of this type of *regress* argument, of which Bales mentions two. Any sound method of decision-making for AU must be able to counter or avoid these arguments to be acceptable.

The second of Bales' regress arguments is this:

> […B]ecause at any given time uncountably many acts, both non-trivial and trivial, are open to us, we may deliberate more or less indefinitely about which of the acts to perform. We agree that indefinite deliberation would be absurd, but if the act-utilitarian must deliberate about whether to cut deliberation short–and apparently he must, if he is to justify cutting deliberation short on act-utilitarian grounds–he is, again, caught in a vicious regress. (Bales, 1971, p. 258)

Another way to make much the same point is to argue that for deliberation to be justified in the first place, it must be (believed by A to be) optimific. If deliberation is to be permissible for A in a situation, deliberating would have to be optimific for A in this situation. Remember that according to AU, an action is permissible (or right) if, and only if, it is optimific. But what reason

could an agent have for believing that deliberating is the optimific thing to do in a particular situation? Apparently, to obtain such reasons, the agent must try to determine whether deliberation would be optimific or not. But, once again, trying to determine whether deliberation is optimific or not would be to deliberate. Again, the regress reappears.

Another way in which the regress might start is to ask what method of decision one ought to use, according to AU, when deciding which method of decision one ought to use in practical situations of choice henceforth. Whichever method one decides to use it seems that, in order for us to be justified in using this method, the choice must have been governed by a (justified) method of decision-making. But the existence of such a method presupposes our access to yet *another* (justified) method of decision. Once again, we are back in the regress.

Yet another 'strong' argument could be stated in the following way: From a practical, as well as from a normative point of view, it is not evident that we ought to consider all our situations of choice, *as* situations of choice. However, given AU, it seems impossible to determine, in a non-arbitrary way, just *what* situations of choice we ought to be considering. Which situations we ought to consider seems to be determined by AU, but in a way that is of no help to us; 'Consider the situation which it is optimific to consider!' does not suffice to help us. This leads to a problem in practical decision-making that may be illustrated in the following way.

At 19:28 I wonder whether I should watch the 19:30 news on TV or go to the cinema (which starts at 19:45, a fifteen minute walk from where I am). In this example it is evident that if I want to make a choice between these two alternatives, I can deliberate for only two minutes. After that I no longer face the same situation of choice. This poses no particularly interesting problems, as long as AU is not brought to bear on the case. But if it is, the situation changes. What seemed to me a reasonable representation of my (subjectively) interesting alternatives may no longer be plausible. These alternatives are interesting and subjectively relevant just because of my wish to decide which of them I want to pursue. However, if I ask myself what I morally ought to do (according to AU), these alternatives might not be the only relevant alternatives. There are other things that I can do. I could write on my thesis, go to a restaurant, go visit my grandmother, etc.

On top of this, I might also question why I should at all try to find out what to do at the particular time I am considering. Why am I considering what to do at 19:30? Why not 19:31, 19:40, 00:00, the next morning, the next year? What situation and what time we should consider, it seems, would have to be determined in a non-arbitrary way. The only way to do this seems to be through an appeal to AU. But, once again, in order to do that, I must do this in one specific situation rather than another and the question then becomes what situation should *that* be? Once again we seem to be back in the regress. We

need a method of decision-making that would help us answer these questions without giving rise to these problems.

As I hinted in the outset of this section, all regress-arguments gain their strength from what Bales refers to as the 'weak' arguments from impracticability, i.e. arguments exploiting the apparent fact that human beings are unable to obtain the knowledge (or justified beliefs) needed for identifying the right action according to AU[12].

4.0 Criteria of rightness and methods of decision-making

Let us examine the distinction between criteria of rightness and methods of decision-making more closely. The previous section pointed to problems with failing to uphold this important distinction. These problems constitute the background for the discussion to come. I start with a short history of the distinction. After that I discuss the significance of the distinction, i.e. its justification as well as its importance. Part of the justification is, as we have already seen, that the distinction helps to evade the accusation of incoherence that plagued the "crude" application of AU.

4.1 The origins of the distinction

One could perhaps argue that having to introduce a distinction between a criterion of rightness, on the one hand, and a method of decision, on the other, one has somehow presupposed that two different things was one and the same[13]. As will become apparent in the following, I think hardly any utilitarian thinker has made *this* particular mistake.

It is not easy to determine exactly when this distinction first appeared in the literature of utilitarian ethics and I shall not try to do this. I will only point at some claims made by early utilitarian thinkers that suggests that, in some form, this distinction is acknowledged early in the history of utilitarianism. If we interpret the distinction in a highly inclusive sense, a predecessor of its modern form appears already in 1789, in the writings of Jeremy Bentham. Consider the following passages:

> By the principle of utility is meant that principle which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question […]. (Bentham, 1987, p. 65)

[12] As Niklas Juth has pointed out to me, this implies that an omniscient AU-agent would not be trapped in these regresses. She would not need to deliberate in order to determine what she ought to do; she just *knows*, and consequently does it.

[13] "My reply is that if act-utilitarians have conflated logically independent procedures, that of providing an account of right-making characteristics and that of providing a decision-making procedure, they have conflated them, and someone needs to point that out."(Bales, 1971, p. 264)

In order to determine the 'tendency which an action appears to have', the following process is suggested:

> V. *Process for estimating the tendency of an act or event*. To take an exact account then of the general tendency of any act, by which the interests of a community are affected, proceed as follows. Begin with any one person of those whose interests seem most immediately to be affected by it: and take an account,
>
> 1. Of the value of each distinguishable *pleasure* which appears to be produced by it in the *first* instance.
> 2. Of the value of each *pain* which appears to be produced by it in the *first* instance.
> 3. Of the value of each *pleasure* which appears to be produced by it *after* the first. This constitutes the *fecundity* of the first *pleasure* and the *impurity* of the first *pain*.
> 4. Of the value of each *pain* which appears to be produced by it *after* the first. This constitutes the *fecundity* of the first *pain*, and the *impurity* of the first *pleasure*.
> 5. Sum up all the values of all the *pleasures* on the one side, and those of all the pains on the other. The balance, if it be on the side of pleasure, will give the *good* tendency of the act upon the whole, with respect to the interests of that *individual* person; if on the side of pain, the *bad* tendency of it upon the whole.
> 6. Take an account of the *number* of persons whose interests appear to be concerned; and repeat the above process with respect to each. *Sum up* the numbers expressive of the degrees of *good* tendency, which the act has, with respect to each individual, in regard to whom the tendency of it is *good* upon the whole: do this again with respect to each individual, in regard to whom the tendency of it is *bad* upon the whole. Take the *balance*; which, if on the side of *pleasure*, will give the general *good tendency* of the act, with respect to the total number or community of individuals concerned; if on the side of pain, the general *evil tendency*, with respect to the same community.
>
> VI. *Use of the foregoing process*. It is not to be expected that this process should be strictly pursued previously to every moral judgement, or to every legislative or juridical operation. It may, however, be always kept in view: and as near as the process actually pursued on these occasions approaches to it, so near will such process approach to the character of an exact one. (Bentham, 1987, p. 87-88)

This passage is interesting because it states both an incomplete[14] method of decision and because it seems to acknowledge that using this method or process is not definitive of the right action. This is, I think, evident from the fact that this process should not be strictly pursued previously to every moral judgement. It should be noted that, for the 'procedure' mentioned above to be (a part of) a method of decision, we must assume that the 'moral judgement'

[14] It is incomplete in that it does not tell us what to do when the 'foregoing process' ought not to be followed.

mentioned in the passage figure in the normative judgements preceding the practical situation of choice. That the 'partisan of the principle of utility' makes a moral judgement (as to whether the proposed action is right or wrong) in the situation of choice and then acts accordingly.

When the origins of this distinction are discussed in the literature, it is often claimed that the distinction first appeared in an explicit form in the writings of Sidgwick. But as the above passage suggests, it seems that even Bentham acknowledged and made use of at least a very similar idea. After Bentham, John Stuart Mill, in his *Utilitarianism* from 1863, is even more explicit in maintaining the distinction. In *Utilitarianism*, Mill writes:

> Whatever we adopt as the fundamental principle of morality, *we require subordinate principles to apply it by*: the *impossibility* of doing without them, being common to all systems, can afford no argument against any one in particular […]. (Mill, 1987, p. 297. My italics.)

Obviously Mill was aware of the distinction and believed it to be both important and necessary to uphold it. Henry Sidgwick posed the distinction in terms of the criterion of rightness and the different motives an agent should have. Thus he writes in *The Methods of Ethics*:

> [...I]t is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim: and if experience shows that the general happiness will be more satisfactorily attained if men frequently act from other motives than pure universal philanthropy, it is obvious that these other motives are reasonably to be preferred on Utilitarian principles. (Sidgwick, 1874, p. 413)

And further:

> […W]e are now in a position to consider more closely to *what method of determining right conduct the acceptance of utilitarianism will practically lead*. (Sidgwick, 1874, p. 460. My italics.)

These passages from Bentham, Mill and Sidgwick point to the same distinction that modern utilitarian thinkers have adopted.

Derek Parfit has made much of the same points in other terms. When he discusses C (his formulation of consequentialism) he makes a distinction between what actions we would have most reason to do, on the one hand, and what dispositions we ought to have, on the other. This way of construing C made it possible to defend it against different counterarguments. C could be defended against claims to the effect that it could not account for agent

relative values[15]. That C can account for something close to agent-relative values is shown by the example of 'blameless wrongdoing', i.e. cases where we act from the best set of motives possible for us, yet perform actions that are sub-optimal. It can be argued that because we are justified in having the best set of motives, C could be said to justify some of the dispositions or motives that agent-relative theories would have us foster.

In chapter IV, the two-level approach to moral thinking presented by Hare, will be seen to serve much of the same purpose as do this distinction. Moore and Tännsjö also acknowledge this distinction.

4.2 Examples of methods of decision-making

What is a method of decision? Should it be conceived of as a manual or a handbook, which the agent is supposed to consult before acting? Or is it a stable habit that one acquires and lives in accordance with? It seems that how we should conceive of a method of decision *for AU*, is problematic. We ought not to try to give the notion of a method of decision-making a precise connotation independent of the normative theory that it is a method of decision-making *for*. Because our adoption of the handbook-understanding and our adoption of the habit-understanding of a method of moral decision-making could affect the outcome of our actions, the understanding we give to the concept will not be normatively neutral. I will examine what has been proposed by utilitarian thinkers as possible candidates of methods of decision-making. But I will not offer necessary and sufficient criteria that have to be met in order for something to be a method of decision-making.

A method of decision-making can be taken to play two important roles. First, a method of decision-making may provide a *procedure* for reaching an answer to the question of what one ought to do in a situation of choice. It tells the agent how she should proceed in reaching her decisions. Secondly, A method of decision-making can also issue a *prescription*, telling the agent to perform a particular action (identified by a certain description).

In order for the reader to get a tentative idea about what could count as a method of decision-making, I will give a few examples. One, rather extreme, method is that of using a dice or a coin in taking one's decision[16]. If heads comes up, then the agent decides to perform an action, *a*. If tails comes up, then the agent decides not to. Of course, this method is silent as to what the

[15] Cf. Parfit, 1984, p. 31-35. For an interesting example of the use a utilitarian can have of the distinction between criterion of rightness and methods of decision for defending utilitarianism against objections such as the one from utilitarianisms alleged inability to account for personal integrity, cf. Brink (1986).

[16] Cf. Reinhart, L "The dice man". A travelling program from *The Discovery Channel* also exploited the idea of using a dice in taking one's decision. The host and narrator ask different persons on the street for places of interest to visit. He assigns each option to a number on the dice, rolls the dice and decides to go to that particular place. By assigning options to numbers in the way that he does in this show, on this method only six alternatives are relevant!

agent takes to be the two relevant alternatives. And the specification of the alternatives is important for the (imagined) reasonableness of this method.

Another possible method of decision is to try, in every situation, to adhere to conventional, or common sense, rules of conduct without further ado. This method is not very precise, of course. But, surprisingly enough, it is not that much more imprecise than some of the proposals for method of decision suggested by some of the utilitarian thinkers discussed below! A third example would be to try to maximise expected utility. I will discuss this approach in chapter III. Hopefully, these examples suffice in order for us to get an idea of what a method of decision-making is. In fact, I think that we all have a rough, intuitive, idea of what would constitute a method of decision, i.e. some sort of principled way of approaching decision-making in practice.

There is a sense in which a method of decision-making for AU would govern every situation in which the agent can make a choice. This, in turn, implies that there are no situations of choice which are morally "neutral", i.e. situations where the criterion does not imply that an action is right or wrong. AU is, in this sense, constantly demanding[17]. Of course, even given AU, two situations or states of affairs may be morally neutral in the sense that they are indifferent in terms of realized value. If one action had one of these situations as its outcome and an alternative action had the other, then they may both be a right action. However this is not a very interesting sense of the notion of 'morally neutral situation'. Instead, by this notion I mean situations of choice in which morality does not give any answer concerning the normative status of actions. The existence of such situations seems to be accepted by our common-sense conception of morality[18]. Let us consider an example of this. For common-sense morality, a lifetime spent in the dark corridors of a library reading Hesse and Dostoyevsky is not viewed as morally wrong. This is so, despite the fact that the reader probably could have done something better if she had devoted her life to something else, such as working towards relieving pain and starvation in a poor country, than this seemingly honourable study of

[17] Cf. Kagan (1989), p. 2.

[18] Cf. Kagan (1989). According to Kagan, this conception of morality, which he calls 'ordinary morality', comprises two broad features. The first is a belief in the existence of *options*. "On the view of ordinary morality, then, I am permitted to favor my interests, even if by doing so I fail to perform the act which leads to the best consequences overall. Since the agent is given the option of performing (or not performing) acts which from a neutral perspective are less than optimal, we may call such permissions *agent-centered options*, or more briefly, *options*. […] This is not to say that no sacrifices at all are required by ordinary morality, but they *tend* to be rather modest and limited. […] Of course, I am *free* to make such sacrifices if I choose to–and morality encourage me to do so–but these acts are not required of me […]. The second broad feature of ordinary morality is that it lays down certain strict limits on our actions–forbidding various types of acts *even* if the best consequences overall could be achieved only by performing such an act. I may not murder my rich uncle Albert in order to inherit his wealth. […L]et us call such limits *agent-centered constraints*, or more briefly, *constraints*. The second feature of ordinary morality, then, is the belief in the existence of constraints." (Ibid. p. 3-4) In chapter IV I will discuss the use that utilitarian writers make of 'common sense rules of conduct'. They do not, however, present a precise meaning of this notion. What they say is therefore bound to be rather imprecise. I will return to this point later on.

the arts. Perhaps common sense would still regard one of these life-projects to be *morally better* or more *praiseworthy*, but this is another matter.

AU, on the other hand, implies that if the agent does not perform the optimific action in a given situation, her action is morally wrong. This can be what is usually described as an action, but it can also be what is, by common standards, referred to as an omission. The important thing is that, whatever we do or do not do, our acts and omissions are both in the same moral category when it comes to moral evaluation, at least on the most fundamental level. AU, as an objective criterion of rightness of actions, applies to every action no matter if we describe it as the omission of some other action or in its own right. If some situations should be considered 'neutral', in the sense that the agent is 'morally free' to abstain from optimific actions, the justification for this would have to be provided by a method of decision-making condoning this. Of course, this method would have to be supported by AU as the best method of moral decision-making.

4.3 Consequences of the distinction

There is no direct way of using AU's criterion of rightness in moral deliberation or decision-making. Some utilitarian writers seem to have assigned little importance to this problem. Smart seems to think that the relation between criterion of rightness and methods of decision-making is rather straightforward:

> Moreover I have said that act-utilitarianism *is meant to give a method of deciding what to do* in those cases in which we do indeed decide what to do. (Smart, 1973, p. 44. My italics.)

> The utilitarian criterion, then, is *designed to help* a person, who could do various things if he chose to do them, *to decide* which of these things he should do. (Ibid. p. 46. My italics.)

Smart takes the utilitarian criterion to 'give' a particular method of decision. The criterion is said to be 'designed to help a person' to decide what to do. Things are more complicated than this, though.

The "crude" method of trying to perform (only) optimific actions is not a viable strategy for an agent who believes that AU's criterion of rightness is true. The crude method, if it involves some process of deliberation with the aim of determining the optimificness of alternative actions, leads to a vicious regress. This conclusion has different consequences.

The previous discussion illustrated the necessity of taking the relation between criterion of rightness and decisions to be more indirect or sophisticated. Failing to do this made the resulting way of making decisions open to regress arguments. We must establish a connection or "bridge" between AU's criterion of rightness of actions and a method of decision-

making. In succeeding with this we have to proceed from two points, so to speak. On the one hand AU's criterion of rightness imposes some restrictions as to what should count as a plausible method of decision. On the other hand, what we want out of a method of decision-making imposes other restrictions. We want methods of decision-making that provide us with practical action-guidance. We want these methods to issue prescriptions under descriptions that help us *guide* our actions, descriptions that we can use in determining what to do. AU's criterion of rightness of actions does not tell us what to do in a way that is practically helpful for us. Furthermore, we need reasons for believing that the adherence to a particular method leads to an approximation of the overall aim that AU gives us, i.e. the maximisation of utility.

Summing up the problems of this more direct approach to moral decision-making, the prospects of applying AU when trying to determine what one ought to do in practical situations of choice is bleak. But perhaps we could learn something from these problems. Are there ways of circumventing some of the problems? To answer this question, I will use the preceding discussion in setting up desiderata that a reasonable method of decision-making should meet. Before doing this I shall formulate my main question.

5.0 The main question

The central question of this essay is: Can an agent who is normally equipped from a cognitive point of view ever be justified in believing that an action is right according to AU? As has already been pointed out, in response to this question, advocates of AU usually introduce the distinction between "criteria of rightness" and "methods of decision-making". They say that, even if AU cannot be applied "directly", it can give us reason to adopt some specific method of decision-making, that in turn leads us to form particular decisions. In what follows, I shall scrutinise these claims. More generally my question is:

To what extent is it possible for an AU-agent to justify her trying, from now on, to adhere to a particular method of decision-making?

I have to limit this question in several ways. I am thus not interested in trying to determine which method of decision an agent ought to try to adhere to according to AU, *period*.[19] We have to make our question relative to *agents*, *time* and the *degree of success in adhering to the method*. Let me comment on these points in that order.

[19] A method of decision is never best period. It is *very* unlikely that there is a particular way of making ones decisions that is the optimal method for every one at every time. However, it seems as if the utilitarians I discuss in this essay assume that there are *general features* common to all reasonable decision-making procedures, at least for agents believing in, and wanting to approximate, the overall goal of AU. These utilitarians writers defend methods that have a common structure, then.

Firstly, different methods of decision are probably optimal for, or relative to, different agents. (Cf. My discussion of Hare's *prole* and the *archangel*, in chapter IV, (5.0-5.4). I have already characterised the *AU-agent.* I use this fictive agent in order to evaluate different methods of decision for AU[20]. This agent is normally endowed when it comes to intellectual capacities, ability to predict outcomes of different alternative actions, and so on. Remember that she differs from most us in two important ways, however. The first is that she believes in AU. I will take this to imply that she sympathises with the overall goal of AU, namely that the more realized total positive utility over negative utility, the better. The other way she differs from us is that she wants, she is morally motivated to further, as far as she can, AU's overall goal. She is prepared to make this goal her overriding[21] concern. I have assumed this in order not to beg any questions relative to AU. I do not take this to imply that she is forced to adopt a "crude" method of decision. She has no desire to perform as many right actions as possible, she wants (total) utility to be maximised. And she is prepared to do what she can, and what she believes, to be most conducive to this end. She also wants to be able to justify her trying to adhere to her method of decision.

Secondly, I am interested in the question of which method of decision-making an AU-agent ought to use in her future decisions[22], from now on. It is this kind of practical action-guidance, which is the most interesting for an agent trying to apply AU.

Thirdly, we have the already mentioned degree of successful adherence to the method. Because this essay is about practical action-*guidance*, I am interested in the ability of methods to help agents in their decision-making. In order to do that, the agents should be able to adhere to them. Let me illustrate. Consider the following two questions:

1. Which method of moral decision-making ought an AU-agent *try* to adhere to in making her decisions (from now on)?
2. Which method of moral decision-making ought an AU-agent *adhere* to in making her decisions (from now on)?

AU might be taken to imply that an AU-agent ought to *try* to adhere to a specific method of decision, i.e. a method that is such that if the agent tries to adhere to the method her trying to do so makes her approximate the overall

[20] In real life, and for real agents, different methods may be optimal for different agents depending on which specific characteristics they possess. In this essay, my discussion is a general one so I will not explore this any further.

[21] Cf. Hare (1981).

[22] One thing should be noted here. 'Her future decisions' is not determinate. It could be interpreted as the future decision situations which the agent (i) will be in, or (ii) all possible situations that the agent could be in. Adopting a method of decision will place the agent in situations where she would not be if she had accepted another method of decision. I will not elaborate this point any further. This is a very perplexing and important problem. This complication should be kept in mind in what follows.

goal of AU. AU might also be taken to imply that an AU-agent ought to adhere to (another) method of decision, i.e. a method that is such that the agent adhering to this method makes her approximate the overall goal of AU. These questions are distinct, then, but they are interrelated.

We want methods of decision to be such that if an agent tries to adhere to them, she will succeed. In succeeding to adhere to the method of decision the agent is trying to adhere to, the method is *guiding* her actions. That adhering to a method *would* maximise utility *if* she managed to adhere to it, but that as a matter of fact she will seldom or never succeed in adhering to the method, is irrelevant from this point of view. Such a method does not provide her with practical action-guidance.

Assume that a method of decision-making is such that when the agent *tries* to adhere to it, she *succeeds*. This, surely, is a desirable feature in a method of decision-making. This is what they are for.

The AU-agent, as characterised in this essay, is motivated to try to approximate AU's overall goal. In order to do that, she will need a method of decision-making. Accordingly, it would be a desirable feature of a method of decision-making if adherence to the method would make the AU-agent approximate the overall goal of AU. An AU-agent tries to adhere to the method because she believes that she will succeed in adhering to the method and that adhering to this method is the best way for her to approximate the overall goal of AU. But her beliefs should also be epistemically justified[23]. Is it possible for her to *justify*[24] such a belief? She needs to form a justified belief as to the degree to which she will succeed in adhering to the method, given that she tries. We would also need to form a justified belief as to how close the agent comes, if she adheres to the method, to meeting the overall goal of AU. The form of justification that is relevant here is *instrumental*, i.e. it consists of providing good reasons for believing that a particular method of decision-making is an effective *means* for the agent to approximating the overall *end* of AU, i.e. utility maximisation.

Let us turn to a discussion of different good-making characteristics of methods of decision-making.

---

[23] Notice that, when talking about beliefs being justified, I always refer to *epistemic* justification, i.e. the kind of justification of a belief that is held to be necessary for knowledge.

[24] A note on justification. What we want, ideally, is justified *true* belief. But this sets too high a standard for justification. Because these beliefs are about complicated empirical states of affairs, we have to set our standards of justification accordingly. But as Ragnar Francén has pointed out to me, if justification does not imply truth, then, the argument goes, AU is not action guiding. Even if an AU-agent is justified in her belief that she ought to do X, this might not be what she ought to do according to AU. Justification only implies a high probability of being true. This is (probably) true, but I take it to be irrelevant. This justification would still give the AU-agent reasons for acting. Of course, it would not guarantee that an AU-agent's actions *satisfies* AU, but her actions could plausibly be claimed to be supported by AU.

6.0 Desiderata for methods of decision-making

Let us consider the notion of a 'method of decision'. I will give it a little more precision. The interpretation I shall offer is not, of course, entirely innocent, but is quite compatible with how the utilitarians that I discuss later on seem to understand the notion. By a method of decision-making, I mean a set of rules about how to reach decisions in concrete situations of choice. As I pointed out above, this should not be taken to imply a very high level of systematisation or formalism. A plausible minimum requirement is that the method is determinate enough as to be recognised by the agent as a distinctive method of decision. Adhering to a method of decision implies that one uses a particular way of making one's decisions.

What do we want from a method of moral decision-making? There are familiar desiderata that are commonly recognised in scientific theories. I will call these desiderata *methodological*. I will comment on some such desiderata, but they are of minor importance in what follows. They will only help me in setting up two other types of desiderata for methods of decision-making. According to this categorisation, there are three kinds of desiderata for methods of decisions for AU:

1. Methodological
2. Practical
3. Normative

As will become apparent below, there are no waterproof bulkheads between these kinds of desiderata. In one sense all three desiderata are *methodological*. My reason for dividing them under different headings is that my primary interest is with the latter two. This is because my interest in this essay concerns AU in relation to practical action-guidance. From this perspective, the latter two desiderata are the most interesting. They have a more direct bearing on my investigation to come. Introducing a separate entry for 'methodological desiderata' serves the purpose of delimiting them from the latter two main desiderata. Let us start with the first kind.

6.1 Methodological Desiderata

The methodological desiderata can be divided into two main classes. The first contains the *necessary* characteristics and the second contains the *desirable* characteristics. I start with the former class.

A minimal necessary condition is that it must be *psychologically possible* to use the method. This condition is of course rather indeterminate. Presumably, different agents will have different abilities, which make some courses of actions possible and others impossible. The important thing here is that the method satisfies the dictum "ought implies can". I will make the assumption that all the different methods of decision that I discuss in this

essay are such that it is psychologically possible to adhere to them. Remember that I am discussing a rather unusual agent, the AU-agent, which is prepared to make greater sacrifices for the sake of morality, than ordinary persons. Of course, this is not uncontroversial. Some ethicists have put forward arguments to the effect that, for example, the method of maximising expected utility would make our personal integrity deteriorate[25]. For the sake of argument I will, however, give these methods the benefit of a doubt, and make this controversial assumption.

Secondly, a method is *ceteris paribus* to be preferred to another if adhering to the method does not make us revise it. That is, it satisfies what Gilbert Harman called the ideal of a 'rational equilibrium'[26]. Harman presents his "rational equilibrium test" without extensive elaboration of it:

> In a less extreme idealization an action might count as rational if it results from or is at least in accord with methods of decision that are rational for one to use. […] When is a procedure rational in this sense? One necessary condition is that a rational procedure be in "rational equilibrium" so that the ideal following of the procedure would not lead one to modify it. Occasionally circumstances will make it salient that certain advantages can be gained by changing one's procedure in some way, and until the change is made one's method of decision making will not be in rational equilibrium. (Harman, 1986, p. 112)

That a method of decision must live up to this requirement seems very plausible. It could even be claimed that before a 'method' satisfies this desideratum, it is not really a distinctive method of decision, but rather a 'method' under construction. However, meeting this desideratum is a question of degree. It is *necessary* that a method meet the desideratum *to some extent*, in order to be *a* method. Furthermore it seems *desirable* that a method satisfies the criterion to a high (or even maximal) degree. Let us now turn to the desirable characteristics.

First we have *scope*; a method is *ceteris paribus* to be preferred to another if it is applicable in more situations. A method of moral decision-making ought to give guidance to the agent in every situation of choice the agent encounters (and perhaps even hypothetical situations). This is at least the *ideal* conception of a good method of decision. That great scope is a good thing in a method is, I think, rather uncontroversial. One might claim that because AU is, in a fundamental sense, constantly demanding, a method of decision, if it is to be validated by AU, would need to cover all situations, i.e. it would have to have maximum scope. By taking a method to involve a

[25] Cf. Williams in Smart (1973), pp. 108. Peter Railton argues that this approach leads to 'alienation' with regard to the claims of morality in Railton (1984). This approach is also taken to lead to moral schizophrenia, i.e. between reasons and motives of actions, within the moral agent in Stocker (1976).
[26] Cf. also Tännsjö (1998), p. 38.

default mode one can assure that the method has maximum scope. There is, however, a problem here. It seems to be a good characteristic of a method of decision-making that it gives more *specific* and *determinate* answers to the question "What should I do now?" The answers should leave relatively little space for ambiguity. One problem is that the more specific the principles are, the lesser their scope as well as their simplicity. That is, the more specific a principle is, the fewer are the situations in which it is applicable[27]. This is an instance of counter proportionality. To single out *the* right action, under a useful description, a principle may have to be so specific that it is only applicable in *that* particular situation. A method that provides the agent with a different principle for each and every situation is obviously not of any use. No one (except perhaps God) could keep *that* many principles in his head. Therefore, there seems to be a tension between scope and specificity. Furthermore, fewer principles seem to require that the principles are more general, as opposed to more specific. In chapter IV we will discuss whether, and if so to what extent, the methods proposed by the utilitarians I discuss, involve a default mode of decision-making. The scope of their proposed methods of decision depends on this question.

Secondly, we have *reliability*. A method of decision is reliable to the extent that it prescribes similar ways of making decisions in similar situations[28]. There is an intra-personal, as well as an inter-personal kind of reliability. There are several reasons for regarding intra-, as well as inter-personal, *reliability* as desiderata for methods of decision. Intra-personal reliability is a sign of the methods being in rational equilibrium. Inter-personal reliability is important when it comes to co-operation. The decisions of different agents sometimes, perhaps even most of the time, need to be foreseeable to other agents. In different co-operative ventures we need to predict other peoples' responses, and reliability in our methods of decisions can help us do that[29]. These considerations seem to speak in favour of preferring a reliable method to an unreliable one, at least *ceteris paribus*. This can also be said to make the method more valid, because successful co-operation might be necessary in some situations for the individual agent to realize the most value.

Thirdly, the method should also *handle the imperfect epistemological situation of different agents* in a reasonable and consistent manner. This would involve some way of discounting, for practical purposes, far reaching

---

[27] Cf. Hare (1981), p. 35-36. "A principle which is going to be useful as a practical guide will have to be unspecific enough to cover a variety of situations all of which have certain salient features in common." (Ibid. p. 36)

[28] According to this definition, (with it obvious problems in characterising 'similar'), a method of decision can be reliable despite the fact that the action performed by the agent trying to adhere to the method is not actually prescribed by the method. It is enough that agent's fail in the same way, and in the same circumstances, in order for the method to *be* reliable. However unlikely this is, it is a genuine possibility.

[29] I will not deal directly with the problem of coordination in this essay. Game-theory and cooperative principles not discussed in this essay will obviously become relevant for this larger inquiry, but for reasons of space I will not examine this wider question any further.

outcomes of actions of which we do not have any reason to believe that either the good or the bad consequences would outweigh the other.

6.2 Practical and Normative Desiderata

Let us turn now to the remaining two kinds of desiderata. One desideratum concerns the relation between the method of decision and the agent and the other concerns the relation between the method and AU's criterion of rightness[30]. These two desiderata can be introduced by the following considerations. To be justified in trying to adhere to a specific method of decision in her future decision-making the agent must be able to justify the following two conditions:

1. That she, when trying to adhere to the method, also succeeds in adhering to the method.
2. That adhering to the method, from now on, makes her approximate the overall goal of AU.

If the AU-agent believes that she will fail to adhere to the method, then she is not justified in trying to adhere to the method. The extent to which the agent believes that her attempts to adhere to the method will be successful, influence the extent to which she has a reason to try to adhere to the method[31]. Furthermore, if the AU-agent believes that adhering to the method would not make her approximate the overall goal of AU, she would have no reason trying to adhere to it.

That a method of decision provides guidance to the agent means that it will provide the agent with an answer to the question of how the agent ought to come to her decisions under descriptions that are useful to the agent. A method of decision is not providing practical action-guidance to an agent if she fails to adhere to the method, i.e. fails to meet the method's demands. We are evaluating methods of decision-making, not the actual actions performed by the agent. The outcomes of the actions performed by the agent will, of course, influence the evaluation of the method, but only indirectly. When the AU-agent succeeds in adhering to her method of decision her actions are supported, or governed, by her method. These actions are particularly interesting from the point of view of practical action-guidance in relation to AU. What is the justification for performing these actions? Because this essay

[30] Indirectly, that is. It is the overall goal of AU, that utility is maximised, that is of importance here.

[31] There is a possible loophole here, though. Perhaps the agent could still justify her trying to adhere to the method, despite of the fact that she believes that she will fail, if she can justify a belief to the effect that her failing in the situations where her failures will approximate the overall goal of AU. Of course, it *could* be the case that the outcomes of the actions performed by an agent *trying*, but always failing, to adhere to a method of decision, would, by pure chance or as it happens, maximise utility. But this, it seems, is not relevant on the level of practical action-guidance. The agent would here be 'objectively', but not subjectively, justified in trying to adhere to this method. It is subjective justification that is important relative to the question of practical action-guidance. Cf. also 6.3 and 6.3.1 below.

is about practical action-guidance, the issue of justification of one's actions and of one's method of decision are important. The question is not merely whether AU condones or prescribes trying to adhere to a particular method of decision, but rather if trying to adhere to the method is rationally defensible given the goal set by AU. Whether this is the case, I submit, depends on if the AU-agent has reasons to believe that (i) her trying to adhere to this method will make her succeed in doing so and (ii) that her succeeding in doing so makes her approximates the overall goal of AU.

I will elaborate on both kinds of desiderata. The former desideratum will be referred to as *the practicability desideratum* and the latter *the validity desideratum*. Let us begin with the former.

6.3 The Practicability desideratum

Robert L. Frazier has argued that a method of decision for AU ought to be, what he calls, 'practicable'[32]. A method of decision is 'practicable'

> […w]hen we are able to use it. It need not be easy to use, nor must it be possible in extraordinary situations. But in ordinary situations, even if it takes considerable effort, we must be able to apply the method. Applying it must be practically possible. (Frazier, 1994, p. 45)

I will take 'able to use it' and 'able to apply' to imply that if an agent tries to use the method, then there is a high probability that she will succeed. To try to do something involves a belief that one has a (reasonable) chance of succeeding. I will also take it to admit that a method can meet the desideratum in different degrees. As was mentioned above, this desideratum concerns the relation between the agent and the method of decision. The practicability desideratum links a method of decision (for AU) to its ability to guide the agent in making her decisions in practical situations of choice.

Holly M. Smith, in her (1988) article has discussed similar questions as the ones I raise here. She considers the question whether utilitarianism is 'usable'. She distinguishes two senses in which a principle may be usable, the *internal* and the *external* sense:

> A. Agent S uses principle P as an *internal guide* for deciding to do act A if and only if S chooses to do A out of a desire to conform to P and a belief that A does conform.
> B. Agent S uses principle P as an *external guide for* deciding to do act A if and only if A conforms to P, and S does A out of a desire to conform to P and a belief that A does conform. (Ibid. p. 92)

[32] Frazier (1994).

(A) and (B) corresponds to my distinction between 'trying to adhere to' and 'adhering to'. Smith goes on defining 'usable':

> […W]e may say that a moral principle is usable for making a decision on a particular occasion just in case the agent is able (then) to use it in the sense of (A) or (B). (Smith, 1988, p. 92)

Is AU usable in any of these senses? Smith discusses what she calls the *Problem of Error* and the *Problem of Doubt*. Let us begin with the former:

> This problem afflicts agents who can reason in the requisite manner, i.e., who can deduce a prescription for action from their moral principle. But some empirical premise they invoke, in order to infer that the act is prescribed, is false, and their conclusion is false as well. The act they believe to be right is not in fact prescribed by the principle. Consider a politician who wants to follow utilitarianism in deciding whether to vote for a flat-rate tax or a progressive tax. The politician believes, falsely, that the flat-rate tax would maximize happiness, and so decides to vote for it. But instead the progressive tax would maximize happiness; only it satisfies utilitarianism. (Smith, 1988, p. 94)

Smith argues that AU might well be usable in the internal sense, "since [the] agent chooses to perform an action out of a desire to conform to his or her principle and a belief that the chosen action does so conform." (Smith, 1988, p. 94) AU is not, however, usable in the external sense according to Smith since invoking AU would not 'lead' the agent to perform the action prescribed by AU[33]. But why is this so? In this case the agent invoked a false premise, so Smith's point hold for this case. But what about the following scenario: Reconsider the politician, but suppose that the flat-rate tax actually is optimific. For the sake of argument, we suppose further that the politician holds the balance of power in the vote. Call her voting for this tax, A. The following might hold true, then:

(i) A conforms to AU (A is optimific)
(ii) Agent S does A out of a desire to conform to AU (She wants to perform the optimific action)
(iii) Agent S believes that A conform to AU (She believes that A is optimific)

According to (B) no evidence, empirical or whatever, need to be invoked in order for the principle to be 'usable' in the external sense. AU's criterion of rightness might well be 'usable' in this sense for an extremely 'lucky' and naïve agent. She might, out of pure luck, manage to perform many optimific

[33] Cf. Ibid. p. 94.

actions and because of her naïve disposition, she just believes that they are optimific as well. The problem is of course (iii). Smith explicitly chooses to state her definitions in terms of 'beliefs', not in terms of '*justified* beliefs'. In determining the practicability of a method of decision-making, we should appeal to justified beliefs. This is important if the agent is to be justified in trying to adhere to a method of decision-making. It does not suffice, in order for a principle to be practicable, that the principle is 'usable', in the above-mentioned sense.

The second problem is the Problem of Doubt:

> In this kind of case, the decision-maker cannot even engage in the requisite mental processes: he or she lacks the empirical premise necessary to connect the principle to any act, and so cannot come to believe of any act that it is prescribed by the principle. For example, the politician may feel uncertain which tax would maximize happiness […]. Neither of these decision-makers[34] can assent to an empirical premise stating that some particular act has the right-making characteristic specified by his or her principle. Hence no prescription can be deduced from those principles. (Smith, 1988, p. 94-95)

According to Smith, this makes utilitarianism, in this situation, 'unusable' for the agent, in the *internal* sense. But why is this so? The answer is that the agent does not believe that the action conforms to utilitarianism, and this is a necessary condition for satisfying (A). But *if* the politician just happened to believe that the action would conform to utilitarianism, then the principle *would* be 'usable' for her on this occasion. This is not plausible. The notion of 'usability' does not capture what we are after when we want a principle to be action guiding. An agent's beliefs as to what a principle prescribes must be *justified* in order for us to be warranted in taking the principle as our guide.

I stipulate the following definition of 'practicability':

The degree of practicability of a method of decision-making, M, for an agent, A, is determined by M's *success-value* for A.

'M's success-value for A' depends, in turn, on the extent[35] to which A's attempts to adhere to M will be successful. Let us also say that a method is "optimally practicable" for A if there is no alternative method that is more practicable for A.

Now, the idea is that (optimal) practicability in the indicated sense is a *desideratum* for methods of decision-making; it is regarded as a good-making

[34] Smith speaks also of a deontologist in her example, but this is irrelevant here.

[35] This condition may obviously be spelled out in different ways. However, what view we take more specifically on this issue does not affect the main arguments that I pursue.

characteristic of a method of decision-making. That is, if A has reason to think that M is more practicable than M' (relative to A), then she has a (defeasible) reason to prefer M to M' as a method of moral decision-making. It would be futile to *try* to follow a method of moral decision-making, in the hope that the intentional use of the method would constitute a systematic way of making decisions, if the method were seriously lacking in terms of practicability.

The practicability of a method of decision-making is agent relative. This means that one and the same method can be highly practicable for one agent but score low in terms of practicability for another agent. However, in what follows I will concentrate on features that tend to influence the degree of practicability in roughly the same way for most agents. Accordingly, I will sometimes speak of *the* practicability of a method. The relative character of these claims should, however, always be kept in mind.

What does it mean to say that 'A's attempts to adhere to M will be successful'? Does it mean that A performs the actions that the method prescribes? As has already been pointed out, the output of a method of decision-making *can* be a *prescription.* That is, the method may then straightforwardly tell the agent to perform an action that falls under a certain description, e.g. "Tell the truth", "Help people in need" etc. and then an attempt to adhere to the method is successful only if she actually performs the action in that situation. However, the method of can also provide a *procedure* for reaching a decision. In such a case, if an agent has reached a certain decision, this can be viewed as a successful attempt to follow the method only if it is the result of actually having followed the procedure. It is not enough, for instance, that the agent has reached the decision she *would*, contrary to fact, have reached if she *had* followed it. In other words, the method must *in fact* have guided the agent.

In any case, according to my definition of practicability it is enough, in order for a method to be practicable, that the agent succeeds in adhering to the method when she tries to do so. But in order for the agent to be justified in trying to adhere to the method, she must have reason to think that it is sufficiently practicable[36]. The extent to which an agent has justified beliefs about the practicability of her method of decision-making will become important when I discuss the general justificatory question, i.e. the question in what sense and to what degree an AU-agent can justify her trying to adhere to a method of decision-making relative to the overall goal of AU. This question also involves the question of the validity of the method, which will occupy us below.

Perhaps we can make (rough) estimates of the degree of practicability of different methods of decision. There seem to be paradigmatic cases with

[36] That is, the agent must believe that she will have a reasonable chance to succeed in her attempt to adhere to the method.

which we can compare different methods. On the one hand we have "methods" with a *very* high practicability. A method consisting of only one rule of prohibition, such as "Never take active measures to instigate genocide", is an example of this kind. In *trying* to adhere to this "method", an agent is likely to succeed. It also seems very easy to possess justified beliefs about whether one has adhered to the method or not. Further down the scale we have a method consisting of only two norms, "Never kill a person" and "Never tell a lie". It seems relatively easy to possess justified beliefs as to whether one has violated these norms or not[37]. Furthermore, a method of decision containing only these two principles, in lexical ordering say, seems highly practicable. If an agent tries to adhere to this method, she is still likely to succeed.

On the other hand we have "methods" like: "Never move any part of your body however small the movement might be". In trying to adhere to *this* method, an agent is (presumably) guarantied to fail. An even more extreme "method" would be: "Always act in such a way that insures that you never have existed"[38]. This "method" is obviously impossible to adhere to. An important question is whether it is possible to justify a belief to the effect that the methods suggested by advocates of AU scores any better in terms of practicability than does these last two methods. As we have seen, it is very difficult to justify a belief to the effect that the method of intentionally trying to satisfy AU's criterion of right actions in every situation of choice, scores better in terms of practicability, than the latter two "methods". Later on we will consider other approaches to moral decision-making to see how they "score" in this respect.

As has already been pointed out, it is of course *possible* that a certain method of decision is such that agent's attempts to adhere to the method are bound to fail, but that every failed attempt happens to satisfy AU's criterion of rightness of actions. It seems as if AU would imply that agents ought to try to adhere to this method. That this method is highly impracticable, it could be argued, is irrelevant from the point of view of AU. This could be taken to be an argument for not regarding practicability as a desideratum. If so, is it sound?

---

[37] At least this holds in many circumstances. But even these principles fails to be practicable on some occasions due to their vagueness "A principle may be so vague that it sometimes leaves the moral status of actions indeterminate. Consider a principle which states that killing persons is wrong, but fails to clarify whether 'persons' includes early human fetuses or not. Then no one can use this principle in deciding whether or not to obtain an abortion, since she cannot tell whether or not abortions are prohibited." (Smith, 1989, p. 112) Many norms will suffer from this kind of vagueness.

[38] The last two methods violate the dictum "ought" implies "can", but this is irrelevant here. They are only presented here as the limiting case of maximally unreliable methods.

6.3.1 Why is practicability a desideratum?

The intuitive idea behind this desideratum is that it seems to be a prerequisite for principled intentional moral decision-making, i.e. for acting in a systematic and principle-governed way[39]. A method of decision-making is supposed to be a means to this end.

In order to see why practicability is a desideratum for methods of decision-making, consider the claim that *trying* to satisfy AU's criterion of rightness in a *direct*, or 'crude' manner, might not be a good way of actually satisfying AU's criterion of rightness. Part of the justification of the practicability desideratum is that methods of decision ought to be action *guiding*, i.e. they ought to give practical guidance, in order to be of any interest to moral agents in concrete situations of choice. That is precisely their *point*.

Moreover, consider the objection stated in the previous section, according to which there may be a method such that we maximise utility precisely by failing to adhere to it when we try. This would perhaps be an apt remark if we could be justified in believing that a method has the pertinent features. But if we could be justified in believing this, is not this a reason to modify this method in such a way that incorporates these mistakes, i.e. that the modified method would prescribe the 'mistakes' in the situations in which they are the appropriate thing to do. For every method that makes the agent succeed through failing, it is possible to reformulate it so that it succeeds through succeeding[40]. It seems that we are not justified in believing that we have found such a method. Because we have no reason to believe that the best method of moral decision-making would fail to meet the practicability desideratum in this particular way, this argument is of little relevance to the present inquiry[41].

Is it possible to have justified beliefs to the effect that one method is (probably) more practicable than another? There are methods of decision for which this is possible. Whether or not this is true of any of the proposed methods of decision for AU is discussed below.

6.4 The Validity desideratum

A method of decision-making can have different degrees of validity relative to AU's overall goal of maximising utility. I will argue that the overall goal of a method of decision for AU is that adhering to the method enables the agent to approximate the ideal of AU, i.e. to realize as much intrinsic value (and as

---

[39] That an agent's behaviour is governed by principles does not imply that the agent is supposed to be a rigid "rule worshiper". The point is that there should be a rationale behind the actions.

[40] This assumption is controversial: "There might be cases in which no *correct* description of the act would elicit its performance. (Consider the familiar finger game in which the fingers of both hands are entangled in such a way that one becomes confused as to which fingers belong to which hand. In these circumstances, wanting to *straighten the first finger of one's left hand* will elicit straightening the first finger of one's *right* hand, but no accurate description of this act will elicit it.)" (Smith, 1989, p. 129)

[41] Or at least I will assume that it is, for the sake of argument. The reader should remember that what is said below is conditional on this assumption being justified.

little negative intrinsic value) as possible. This overall goal of a method of decision can be approximated to different degrees by different methods.

Why not define the validity of a method of decision-making in terms of the number of right actions performed by an agent adhering to a method of decision. That is, someone might suggest that the most valid method of decision is the one that makes the adhering agents perform the maximum number of right actions. At least in the context of AU, however, this is not a good criterion.

Consider two different methods of decision, M and M'. Suppose, rather unrealistically, that if an agent adheres to M, 99% of her actions would be right, and only 1% wrong. On the other hand, if the agent adheres to M' she does not perform any right actions at all. However, she always manages to perform the second best action in every situation of choice, i.e. the action that realizes the second greatest amount of intrinsic value (let us assume that there is only one such action in each situation). Now, M need not be the best method of decision. It is possible that adhering to M' realizes most intrinsic value, and is thus plausibly preferred on utilitarian grounds. To see this we can imagine that one of the wrong actions prescribed by M causes an insurmountable amount of pain, e.g. a global nuclear war. From this it emerges that the number of right actions is not, given AU, a good criterion for singling out the most valid method of decision-making. The sheer *number* of right actions seems completely irrelevant for an advocate of AU.

Another argument is this[42]. Whether an agent is disposed to follow a certain method of decision-making is not only likely to affect her decisions, but also which situations she finds herself in. An agent who is a pure "Do-gooder", to use Derek Parfit's phrase (an agent who is disposed always to do what she thinks maximises utility), might find herself in entirely different situations than those that, say, a loyal friend (an agent who sometimes benefits her friends at the cost of the overall good) will face. This means that even if a pure Do-Gooder, in every situation, picks the alternative with the best outcome, it might have been better, from the point of view of AU, if she had been a loyal friend. For the situations a loyal friend is confronted with may be such that picking, say, the second-best alternative in some of them may have better consequences than picking the *best* alternative in the situation that a pure Do-Gooder finds herself in. For example, other people may be more willing to cooperate with a loyal friend than a pure Do-Gooder. In other words, a method of decision-making that allows the agent always to pick the right action may, for all we know, place the agent in situations which are such that even if she performs only optimific actions relative to these situations, there might still be another disposition, that will in the long run place her in *different* situations in which she does not perform optimific actions relative to

[42] This argument was suggested to me by Folke Tersman.

these situations, but the agent might still realize a greater amount of utility if she is disposed in this latter way. This may affect the total amount of utility realized by the agent.

In the above-mentioned article by Robert L. Frazier, he comes close to making the mistake that we have considered here. A method of decision is 'validated', according to Frazier,

> […w]hen we have good reason to believe that it gives the *correct* answer to questions for which it is a decision procedure. (Frazier, 1994, p. 45. My italics.)

It is important to keep these matters separated. A method of decision, M, need not, in the relevant sense, be more validated relative to AU if it gives the 'correct' answer on a greater number of occasions than another method M'. But Frazier's notion of 'validation' can be reformulated. I think that the only acceptable criterion by which we ought to compare different practical methods of decision-making for AU is the amount of intrinsic value that the adherence to the methods makes agents realize. It is the *overall* value of adhering to the method that is of importance, not the value realized on every particular occasion. Adherence to the method which make the agent realize the most positive intrinsic value over negative intrinsic value is to be preferred. This method is the best candidate, on utilitarian principles. This is why "approximating the overall goal of AU" will be used in the formulation of the validity desideratum[43]. I stipulate the following definition of validity:

The (degree of) validity of a method of decision-making, M, for an agent, A, depends on the extent to which A's adhering to M makes A approximate the overall goal of AU.

Accordingly, a method of decision, M, satisfies the *validity desideratum* to a higher degree than another method M', for A, if and only if, M has a higher validity than M'. M is thus preferable to M', because validity is a good-making characteristic of methods of decision-making.

It should be noticed that, in talking about "degrees", I do not want to give the impression of more preciseness than there in fact is. In theory, I do think that a precise measure of validity, that allows one to compare different

[43] Smith seems to be after something close to my use of the term 'validity', when she discusses the use of auxiliary decision rules in (1988): "If an agent is able to make an indirect inference from her moral principle, via an auxiliary rule, to a prescription for action, we may say that the principle is indirectly usable for that agent. But what, exactly, is required for an agent to count as having the ability to make such an indirect inference? In particular, must the agent believe of some auxiliary rule that it *is the most appropriate auxiliary rule to use* when she is unable to use the moral principle itself?" (Smith, 1988, p. 97. My italics.) One plausible sense in which an auxiliary decision rule might be said to be appropriate, is when the agent, trying to adhere to this rule, approximates the overall goal of AU (to a higher degree than any competing rules would make her do so).

methods in this respect, can be constructed. But I have not developed any such measure. It is my firm belief that we do not need it in order to claim that we have little reason to think that any of the methods of decision-making that have been proposed in the literature is valid to any significant extent.

In any case, the idea behind the validity desideratum is that a method should be an effective means for the agent to approximate the overall goal of AU. An agent is justified in her belief that the method is valid to the extent that she is justified in her belief that she approximates the overall goal of AU through her adherence to the method. Of course, an agent can fail in adhering to a method. The best she can do is *trying* to adhere to the method. This is the reason for why we also need the practicability desideratum. Both practicability and validity are thus ideals or *desiderata* that can be satisfied more or less for different methods of decision-making. For each proposed method the satisfaction of these desiderata is a matter of degree. Any method of decision can be more or less practicable, as well as more or less valid.

The above reasoning enables us to set up a criterion of adequacy for a justification of trying to adhere to a method of decision-making:

An AU-agent would be justified in trying to adhere to a particular method of decision to the extent that she can justify a belief to the effect that she (i) succeeds in adhering to the method if she tries and (ii) that her adherence to this method she would approximate the overall goal of AU.

6.4.1 Why is Validity a desideratum?

What rationale or justification can the validity desideratum be given? The chief reason is, of course, AU. More specifically, it is the axiological part of AU, i.e. the idea that the more realized intrinsic value, the better. The desideratum is, so to speak, a principle stating that a method of decision-making is better than another when the agent adhering to the method comes as close as possible in maximising the sum total of realized value. This is the direct reason for the validity desideratum.

Can we have justified beliefs about the validity of a method of decision? Is it possible to justify a belief that the method is highly valid relative to AU or that a method, M, is more valid than a method, M'? I will return to these questions in chapters III and IV. Suppose that the conclusion is negative. Does this make the validity desideratum useless or misplaced? Is this a sign that this kind of critique of utilitarian methods of decision is groundless, i.e. if we cannot establish *how* valid a method of decision is, we cannot establish that it *is not* valid, and consequently that validity is a useless desideratum? Or does this inability tell against the possibility of justifying any method of decision for AU?

I think that it is implausible to let this speak against the desideratum. That it is impossible to determine a method's validity, does not mean that it cannot

be valid, but it precludes the AU-agent from justifying her actions through an appeal to the validity of her method of decision. The method stands in need of justification. The desideratum is a natural consequence of AU's criterion of rightness (or overall goal). The AU-agent wants to justify her method of decision-making by appealing to AU. And this sort of justification is what is needed in order for the agent to be justified in using the method based on her belief in AU.

6.5 The relation between practicability and validity

How can these desiderata be used in evaluating different methods of decision for AU? How determinate will the conclusions of these evaluations be? In my assessment of different methods of decision I will proceed as follows. After a characterisation of the method to be evaluated I will try to form a judgement as to the degree to which an agent is able to justify a belief to the effect that her trying to adhere to the method satisfies the desiderata. It is not possible to quantify these conclusions, saying for example that method M is twice as practicable as the method M'. I will only be able to point to general features of these methods that tell for or against their practicability and validity respectively.

Smith presents an argument to the effect that there exist, for every agent, a practicable and valid method of decision for AU[44].

> For each action that is of a type identified as significant by M is simultaneously of *many* other types. In every case, at least one of these other types is one which the agent could unerringly ascribe to the act. Suppose, in a particular case, that the act prescribed by M is also of (M-irrelevant) type T, the agent correctly believes that she has an act of type T available to her, and she knows how to perform this act. Then an instruction in this case to perform an act of type T would lead the agent to perform the act prescribed by M. […] Of course, which act-type correlates in this way with the M-prescribed type, and is also such that the agent knows how to perform it, will vary from case to case, depending on the circumstances and the agent's beliefs. Thus there will be no simple rules to substitute for M. But the rules of M* may take the form of an extended list of prescriptions to perform individual actions. Each prescribed action would be described in terms of an act-type having the feature that if the agent tried to do an action of that type, he would perform the act actually required by M in those circumstances. Such a list might appear as follows: at ten o'clock, empty the dishwasher; at quarter past ten, pay one's bills; at eleven o'clock, balance one's checkbook; and so forth. Presented with such a list, the agent could follow it an so do everything required of him by M – even though he might not believe any of these acts to […] maximize utility and so would not perform them if

[44] Smith talks of a two tier moral system in which 'M' consists of principles that provide right- and wrong-making characteristics of actions, and where M* "consists in rules that are to be used for actual decision-making." (Smith, 1989, p. 119) I will take M* to be the equivalent of a method of decision for AU.

> he were instructed instead to act so as to maximize utility. Such lists could be relativized to each agent. An agent armed with a suitably designed list of this sort, and morally motivated, would perform each of the actions prescribed by M. So appropriate systems M*, of a peculiar kind, do exist. (Smith, 1989, p. 129-130)

As Smith acknowledges, this is not of much help when it comes to practical action-guidance. Although such a method can be said to exist, this is of no practical help to human AU-agents in need of practical action-guidance:

> First, although for each M an appropriate M* exists, there is no reason to believe that anyone knows, or could find out, what the content of any appropriate M* is. Certainly, the decision-maker herself cannot determine what the content of the appropriate M* is, for the decision-maker could only determine this if she knew what M requires in each particular case. […] Indeed, the kind of rules that M* requires – an extended list of acts – is not simple enough to be learnable in advance by any person of normal intelligence. […] Thus the difficulties in actually implementing this solution appear overwhelming. (Smith, 1989, p. 130)

> The same empirical misinformation that plagues their application of M, now prevents them from seeing that M* is the correct code by which to guide their actions. They can apply M*, but they cannot see that it, rather than some alternative, is justified as an action-guide. (Smith, 1989, 131)

So, the sense in which such a method exists, is the same sense in which there might be said that there exists methods for identifying the winning lottery ticket[45]. The 'method' of decision-making consisting of the prescription of marking number 2,8,13,14,17 and 29 on a ballot in a lottery where this combination is the right one, also 'exists'.

Methods of decision-making suggested by utilitarians themselves exhibit different degrees of validity and practicability respectively. A valid method need not be highly practicable. For example, a method consisting of only one prescription, i.e. that the AU-agent ought to perform only optimific actions, is unlikely to meet the practicability desiderata, for a human AU-agent, to a satisfying degree. A highly practicable method need not be valid. For example, a method consisting only one prescription, i.e. do whatever you feel like doing at all times, is unlikely to be valid relative to AU. We have no good reasons for believing it to meet the validity desideratum. So the relation between a method's practicability and validity is a contingent one.

7.0 Concluding remarks

I have argued that trying to perform (only) right actions, or actions which the agent believes (with good reasons) are right according to AU, is not a

[45] Cf. Bergström (1996) p. 81-83.

reasonable goal or ideal. In confronting a concrete situation of choice, we can never assemble enough evidence so as to determine which alternative(s) satisfies AU's criterion of rightness. Any application or use of AU must take a more indirect form. In chapters to come I will go through different suggestions of such forms. Different methods of decision will be examined. Can the criterion of rightness of actions make any particular method of decision reasonable for an agent to try to adhere to? Can AU validate any particular method? Are some methods more practicable than others? Are some methods more valid than the others? These questions will be addressed in the investigation to come.

Chapter III
Maximising Expected Utility

1.0 Introduction

In the present chapter I explore the principle of maximising expected utility. This principle can be interpreted in different ways and can play different roles in different contexts. In one form or another, it has formed a part of every one of the methods of moral decision-making that have been proposed by the major utilitarian thinkers, including those discussed in this essay[1]. Indeed, it is often seen as the most natural approach to decision-making for defenders of AU. It is therefore, of obvious interest for anyone who is interested in the possibility of applying AU in practice.

Utilitarians turn to the principle of maximising expected utility when they realise that AU cannot be applied directly:

> The solution under examination is most often proposed by those act utilitarians who hold objective act utilitarianism to be the correct theoretical account of right-making characteristics, but recognize that inadequate information sometimes (perhaps always) renders it useless for actual decision-making. These theorists conclude utilitarianism must be supplemented by *auxiliary* or *subordinate* rules designed for ease of application when knowledge is scant. The idea is that the agent is to apply one of these auxiliary rules if he cannot apply utilitarianism itself. For example, it is often suggested that when the agent cannot decide which act would maximize happiness, he should follow a rule prescribing the act with the maximum *expected happiness*. He can decide which act satisfies this rule, even if he cannot decide which act satisfies utilitarianism itself. And deciding what to do by reference to this rule is a rational second-best strategy when one's underlying goal is to comply with utilitarianism. (Smith, 1988, p. 95-96)

The aim of this chapter is simply to examine such claims, and to ask more generally whether the principle of maximising expected utility provides any help to the advocate of AU in the context of practical decision-making.

2.0 Maximising expected utility

What does the principle of maximising expected utility say? Call a possible outcome of an action *a*, *O*. For every *O*, there is a probability (either assigned to *O* by the agent, P, or implied by P's beliefs) that *O* occurs, given that *a* is performed. *O* is also assigned a utility by P (or such an assignment is entailed

[1] Cf. Tännsjö (1998), p. 34-35, Smart (1973), p. 46-48, Hare (1981), p. 133.

by P's system of commitments). The probability and the utility of *O* are associated with numerical values[2]. To arrive at the expected utility of *a*, we proceed in the following manner. For every *O* of *a* the product of its utility and probability is calculated. The sum of these products is the expected utility of *a* (for P in S). We can now define what it means for an action to maximise expected utility:

> An action, *a*, maximises expected utility, for an agent, P, if, and only if, none of the alternatives open to P in a situation, S, has a greater expected utility for P in S.

This definition can be used in different ways. For example, it can be used when stating a criterion of "subjective rightness". That is, we can say that an action is subjectively right if, and only if, it maximises expected utility[3]. In what follows, when I speak of maximising expected utility as a *criterion*, this is what I have in mind. Several utilitarian thinkers have exploited the concept of subjective rightness or similar concepts. For example, J. J. C. Smart defends maximising expected utility as a criterion of 'rational' action:

> According to the act-utilitarian, then, the rational way to decide what to do is to decide to perform that one of those alternative actions open to us (including the null-action, the doing of nothing) which is likely to maximize the probable happiness or well-being of humanity as a whole, or more accurately, of all sentient beings. The utilitarian position is here put forward as a criterion of rational choice. (Smart, 1973, p. 42)

And Tännsjö gives the following criterion of subjective rightness, or SUR:

> SUR: An action (performed by a person who believes that AU[4] is true) is subjectively right if, and only if, the person who performs it believes that it maximises expected happiness. (Tännsjö, 1998, p. 34)

SUR refers to the agent's beliefs. In order for an action to be subjectively right the agent must *believe* that it maximises expected utility. Smart's

---

[2] These assignments of probabilities must satisfy the fundamental axioms of the probability calculus.

[3] The following point is important to emphasise already at this point. The fact that an agent 'probably maximises utility' is not the same thing as an agent 'maximise probable utility'. Phrases like 'Likely to produce the best results' are ambiguous. That an action is likely to produce the best results does not necessarily mean that the action maximises the greatest product of probability and utility. Suppose an agent has two possible actions from which to choose in a situation. The first action, A, has a .51 chance of maximising utility (things turn out good), but a .49 risk of *minimising* utility, i.e. leading to a moral disaster (things turn out bad). B has a .9 chance of realizing almost as much utility as A (things turn out good), and a .1 risk of producing only slightly less than that. Under one of the interpretations above we should go for A, because this is the action which has the greatest chance of being right, because performing B ensures that utility is not maximised. This is the sort of situation that the idea of maximising expected utility is thought to remedy, making us shun too risky alternatives.

[4] This is Tännsjö's particular hedonistic version of AU.

'criterion of rational choice' does not refer to the agent's beliefs. It is natural that the criterion of subjective rightness refers to the beliefs held by the agent, because of its connections to a method of decision-making. Things are different with the notion of maximising expected utility, however. The fact that an action maximises expected utility for an agent need not be taken to imply that this is something that the agent knows or believes. An action can *have* a certain expected utility *given* the agents beliefs, without the agent believing that the action has that particular expected utility. Furthermore, an action can have a certain expected utility for an agent given her beliefs even if the agent does not have any beliefs, or even notions, of its expected utility at all. It suffices that the agent's beliefs *entails* an answer to the question of how much expected utility is connected with a particular action.

What is the point of having a criterion of subjective rightness of actions? The idea seems to be that it can form a part of an appropriate method of decision-making for AU by introducing a kind of interim, or "stand-in" goal that for the advocate of AU. What the notion tries to catch is what it would be 'rational' or 'responsible' for an agent to do in concrete situations of choice:

> […W]e want to say that, in general, it is responsible for a person to perform a subjectively right action. (Tännsjö, 1998, p. 34)

The point of this criterion is, in the face of the bleak possibility of trying, intentionally and deliberately to satisfy AU's criterion of rightness, to serve as a kind of working aim. We are not justified in believing that trying, directly, to perform *right* actions is a sufficiently practicable method of decision-making, but trying to perform subjectively right actions is taken to be a 'viable goal'[5]. This criterion takes into account such things as the fact that the agent has only limited knowledge of the different outcomes of her actions, of the facts in the situation, of her alternatives and so on.

The idea that one should try to perform 'subjectively right' actions, when this means (as I shall assume) that they maximise expected utility, shall in what follows be called the *deliberative approach* to moral decision-making. Certainly, this approach seems, at least *prima facie*, to be a natural one, given AU. If maximising utility is the goal of morals, then trying to maximise *expected* utility, given our cognitive limitations, seems to be a rational approach. The deliberative approach to decision-making is the primary focus of this chapter.

Utilitarians commonly think that attempting to perform subjectively right actions is a part of an appropriate method of decision-making but not the whole of it. Sometimes the agent should try to determine which alternative has the greatest expected utility and act accordingly. At other times, she should not calculate but rather apply certain rules of convention, or

[5] Cf. Tännsjö (1998), p. 36.

'secondary rules of conduct' such as: "Don't tell lies", "Do not commit murder" etc.

I shall call such "blended" approaches *restricted deliberative strategies*. Restricted strategies raise a question that I will call "the problem of demarcation": when is it appropriate to calculate and when should one rather follow the secondary rule? The problem of demarcation, as well as the *restricted deliberative strategy* in general, will be discussed in the next chapter.

3.0 The deliberative approach

Let us begin, then with the unrestricted deliberative strategy, in which the principle is supposed to be applied to all decisions[6]. As I just said, this is a 'straw-man'. No utilitarian thinker has defended such strategy to decision-making. But the discussion serves the purpose of illuminating and clarifying some important aspects of the idea. How should we construe the pertinent method of moral decision-making?

Consider the following:

> But what does it mean more exactly to say of a person that he or she tries in a situation to maximise expected happiness? It must be understood roughly in the following manner. This person tries to represent to himself or herself all the alternative ways in which he or she can possibly act in the situation. He or she tries to make a list of such alternatives that exhausts all possibilities, and such that all the alternatives are mutually exclusive. He or she tries furthermore to form opinions about possible outcomes, were he or she to act in these possible ways, and he or she tries to form opinions about the value of these outcomes. He or she then attempts to find out which one among these possible alternatives is associated with the greatest weighed sum of values and probabilities, and tries to act in this manner. (Tännsjö, 1998, p. 36)

In accordance with this, we may say that *the deliberative approach* consists in six steps. An agent should:

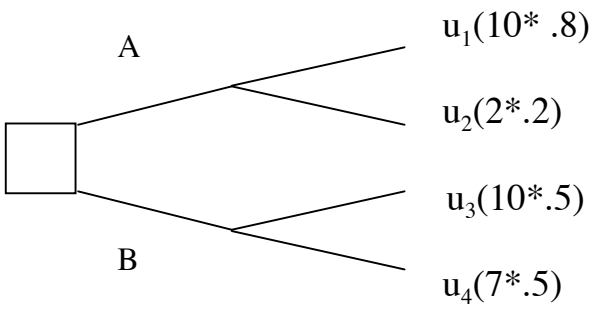
1. Determine the relevant alternatives[7].
2. Determine the possible outcomes of the different alternatives.
3. Assign utilities to the different possible outcomes.
4. Assign probabilities to the possible outcomes of each alternative.

[6] This is a simplification. Unless we want to find ourselves trapped in a regress, the principle cannot be used when deciding which situation of choice to consider. The aim is rather to maximise expected utility relative to some, independently characterised, decision-matrix. This qualification should be kept in mind.

[7] There is a problem here as to how the agent is supposed to determine whether to look for more alternatives or stop with the ones she got. I will return to this issue. This problem is a serious one. No plausible suggestion on how this problem should be solved has, to my knowledge, been presented.

5. Calculate the expected utility of the different alternatives by summing up the products of utility and probability of every relevant alternative.
6. Perform the action associated with the greatest product of probability and utility.

The situation of choice can be represented by the following figure, where A and B are two alternative actions:



The expected utility associated with A is 8.4 (u_1+ u_2) and with B 8.5 (u_3+ u_4) According to the principle of maximising expected utility, B ought to be performed. This account raises certain important questions. The approach requires that the agent identify her alternatives, and assign numbers to utilities and probabilities. In what follows, I shall assume that it is meaningful to ask if these assignments are true or justified, which in turn seems to presuppose, e.g. that utility can be measured, and that states can be ranked with reference to the utility realized in them on a ratio scale. It is well known that this presupposition faces certain deep theoretical problems, as does the assumption that the agent's beliefs about her alternatives can be true or justified[8]. However, I shall simply grant that all these problems can be solved. My aim in this chapter is to evaluate the deliberative approach given that the theoretical problems are solved.

Adopting the deliberative approach, just means that one goes through steps 1-6 of the above list. The application of a principle of rational choice such as the principle of maximisation of expected utility requires a representation of the situation, presumably a matrix with a representation of the alternatives, the utility of each possible outcome and an estimation of the probability that the utility would be secured. How do we obtain such a representation? This question depends upon, among other things, what sort of conditions it must satisfy. The final step in this process is to apply one of the principles of

[8] What does it mean that such beliefs are correct? The concept of "alternative" has been widely discussed, and the various ideas about their properties and nature have been proposed. Cf. Bergström (1966) and Carlson (1995). However, much of this discussion is irrelevant to my present concerns.

rational choice to the matrix and perform the action that the principle prescribes.

Trying to adhere to the deliberative approach means that one is *trying* to maximise expected utility. A further question is whether the deliberation results in the agent succeeding in determining which of the alternatives that satisfies the principle of maximising expected utility. Here, it seems, much can go wrong. In line with my definition in chapter II, the degree of success, over time, in satisfying the criterion will determine the *practicability* of the method. But what does it mean to succeed? This will depend on the particulars of the approach. As I said above, there are several possibilities.

3.1 Five steps of deliberation

To repeat, there are six steps that the deliberating agent has to go through if she uses the *deliberative approach*:

1. Determine the relevant alternatives.
2. Determine the possible outcomes of the different alternatives.
3. Assign utilities to the different possible outcomes.
4. Assign probabilities to the possible outcomes of each alternative.
5. Calculate the expected utility of the different alternatives by summing up the products of utility and probability of every relevant alternative.
6. Perform the action associated with the greatest product of probability and utility.

Let us for the moment ignore the final step[9]. Each of the others allows for different interpretations, depending on the input we require the agent to have, i.e. depending on the conditions we impose on her beliefs, assignments and calculations. Do the probability-assignments have to be justified in some substantial sense? Do the assumptions about which alternatives she faces have to be correct? Different answers to questions such as these give rise to distinct versions of the principle. I shall say that these versions are located on a *spectrum* ranging from no idealisation, when any kind of input would do, to full idealisation, when we impose strict conditions on the appropriate input. Since I am not taking a stand on the question of what is intrinsically valuable, and since my arguments are directed at the structure, rather than the substantial content of AU, I will discuss points 2 and 3 under the same heading. Point 5 is a consequence of 1-4, and I will only comment on it briefly.

[9] I will ignore 6 because it is not part of the deliberation process.

3.2 Interpreting the "input": The spectrum

At the one end of the spectrum, with no idealisation, we have the *Pure subjectivist approach*, and at the other we have the *Ideal subjectivist approach*. Neither of these approaches represents a plausible way of interpreting the steps 1-5 above for normally endowed agents. Therefore, I shall also be considering a middle position, which I call the *Reasonable subjectivist approach*. I think some version of this approach is the most plausible one to take if one is to succeed in justifying ones method relative to AU[10]. I shall begin with describing the two extreme positions and turn to the Reasonable subjectivist approach only after setting the stage for the discussion to come. These positions will initially only be presented briefly. They will be elaborated in more detail only after their suggested interpretations of the inputs of deliberation (1-5 above) have been presented. These positions help to illustrate the fact that one can place low as well as high standards on the relevant inputs of the deliberation process.

3.2.1 The Ideal subjectivist approach

At one end of the spectrum, we impose rigorous conditions on the inputs. In the ideal subjectivist approach only justified true beliefs are allowed. In order for the agent to have succeeded in following the method, the relevant alternatives, outcomes, utilities and probabilities that have figured in her deliberation must be the 'real' or 'actual' alternatives, outcomes, utilities and probabilities. In other words, if she deliberates in accordance with the five steps, and then reaches a decision that does not maximise expected utility given the alternative she actually has, then she has failed to adhere to the method. She has also failed in adhering to the method if she is unable to acquire the beliefs and make the assignments of the requisite kind. In order to give this idea some content, let us introduce the concept of a *cognitively and conatively ideal agent.* This agent has played a crucial role in arguments used by utilitarian thinkers. R. M. Hare called it the *Archangel*[11], to whom we are going to return to in chapter IV.

The ideal agent is taken to possess the following characteristics. She is omniscient[12]. She knows every empirical fact as well as every evaluative or

---

[10] As it turns out, however, the demand for justification probably makes the resulting method score rather low in terms of practicability. This will become apparent below.

[11] Cf. Hare (1981).

[12] Before entering into a discussion of this approach, I shall make a brief comment on the issue of determinism. If determinism is true, an omniscient being would not have to use probability assignments. She would know the outcome of every action. And she will be able to perform only optimific actions if that is what makes her meet AU's overall goal. She would not even need a "method of decision-making". If determinism is false, however, even an omniscient being would have to assign probabilities to outcomes. Her omniscience would be limited. She would know the possible outcomes of every action and the probability it has of being realised. An omniscient being would know if determinism is true or false. Ordinary human agents, however, do not know this. The ideal subjectivist approach sets rigorous demands on the agent. This means that if their probability assignments are to be 'correct', i.e. representing the actual probability of the different outcomes of an action, these agents would not know how to make their assignments. *Because* we

normative fact. We have assumed that AU is true, so this is also something that she knows. Moreover, she wants to approximate AU's overall goal, i.e. to maximise utility.

Actual human agents are far from omniscient. Many of their beliefs are, presumably, false and unjustified. The Ideal subjectivist approach sets up a 'filter of relevance' on the input to be used in deliberation by human agents. Beliefs held by human agents are allowed in deliberation only when they are justified and true. False beliefs are not allowed in the deliberation. Human agents do not know which of their beliefs are true and justified. This is not a problem, however. The ideal subjectivist approach is only characterised as an illustration of one end of the spectrum.

In the discussions of the relevant alternatives, outcomes, utilities and probabilities below, I will use this ideal to give an ideal account of these notions. It is time to turn to the other end of the spectrum.

3.2.2 The Pure subjectivist approach

Here, the deliberative approach is given a pure subjectivist interpretation. The relevant alternatives, outcomes, utilities and probabilities are those that the deliberating agent actually takes them to be. One version of the Pure subjectivist approach is Bayesianism[13]. According to this method:

> To deliberate is to evaluate available lines of action in terms of their consequences, which may depend on circumstances the agent can neither predict nor control. […] In the Bayesian model the agent's notions of the probabilities of the relevant circumstances and the desirabilities of the possible consequences are represented by sets of numbers combined to compute an *expected desirability* for each of the acts under consideration. The Bayesian principle for deliberation is then to *perform an act which has maximum expected desirability*. […] The Bayesian model may be as applicable to the deliberations of a knave or a fool as to those of a good and wise man, for the numerical probabilities and desirabilities are taken to be subjective in the sense that they reflect the agent's actual beliefs and preferences, irrespective of factual or moral justification. […] In the simplest cases the number of possible acts that the agent believes are available to him is finite, as is the number of possible circumstances that he regards as relevant to the outcomes of the acts. The first stage of deliberation is then an analysis of the situation into *acts*, *conditions,* and *consequences* which can be summarized by a *consequence matrix*[…]. […] Entries in consequence matrices are best considered as notes made by the deliberating agent to help determine the numerical desirabilities of situations which he expects to arise if he performs one or another act

seem required to determine whether determinism is true or false in order to make the probability assignments required, normally endowed agents cannot satisfy the demands of the ideal subjectivist approach. This makes it uninteresting as a method of decision for AU. This approach would set to high a standard of justification, standards that are inappropriate in this context. This approach is only presented as a limiting case.

[13] I take as my point of departure R. C. Jeffrey's characterisation of this method in Jeffrey (1965). I will not go into meticulous detail here, because my points are of a general nature.

> under various conditions. […] To compute the expected desirability of the act, multiply corresponding probabilities and desirabilities, and add [the products together] […]. Having done this for each row [of the matrix], select for performance one of the acts for which this sum of products is the greatest. (Jeffrey, 1965, p. 1-6)

On this approach the beliefs, utility- and probability assignments of the agent are taken at face value. I will argue that this is not a plausible interpretation of the deliberative approach. What about a middle position, then?

3.2.3 The Reasonable subjectivist Approach

The Pure subjectivist approach needs to be complemented. This approach imposes too low standards on the input used in deliberation. Since this approach takes the agents beliefs and utility- and probability assignments as given, we have no reason to think that the validity of this approach would be high for normal AU-agents. The positive feature of the method is that it is (probably) more practicable than more idealised approaches for human AU-agent. (I will elaborate this below.)

One way to improve the Pure subjectivist approach is to require that the beliefs and assignments constituting the "input" of this approach are *justified*. This transforms it into the *Reasonable subjectivist approach*. This approach is a middle position. Superficially, it seems to escape the drawbacks of the Pure subjectivist approach (with its lack of validity relative to AU) as well as those of the Ideal subjectivist approach (with its lack of practicability relative to human agents).

The Reasonable subjectivist approach requires that the beliefs and evaluations of the agent that are relevant to her decisions are justified, that they are supported by reasons and evidence. But justification is a matter of degrees. Accordingly, we may imagine a continuum of methods, from a high degree of justification, down to the lowest, where agent's beliefs are taken at face value. When we impose stricter demands on the beliefs of the agent, having them to meet higher standards of justification, new sources of error arise. If the demands are very high then actually and *intentionally* maximising expected utility is not something that an agent with normal cognitive abilities is very likely to succeed in doing.

To find the appropriate compromise might accordingly not be easy. On one suggestion, in order for an agent to have followed the deliberative approach, she must have formed her utility and probability assignments after careful consideration (in a cool hour)[14]. But such an idea raises serious problems.

[14] The question of what standards the input should meet is hinted at by Tännsjö: "[…S]hould we also require that the probability assessments and value assessments of this person hold a certain quality? Should we require that her or his various assessments can be fitted into a 'reflective equilibrium'? This would probably mean that the concept [subjective rightness] will be without application. Who has ever held a complicated set of opinions satisfying such rigorous requirements? But we should perhaps at least require that this person has

How careful ought the agent to be? How much time and effort ought she to spend? The obvious answer is of course "As much as it takes in order to approximate the overall goal of AU!" But this does not help much. To 'carefully consider' a situation of choice is to deliberate. A method of decision-making ought, assuredly, to help an agent approximate the overall goal of AU, but this answer provides no practical guidance to the agent. On the other hand, every other answer would seem to be arbitrary relative to AU. Determining, in a non-arbitrary way, just how much 'careful consideration', i.e. how much time and effort the AU-agent is justified in spending on deliberation, seems very hard. This problem seems to set off the regress-problems of 3.1 in chapter II once again. I will return to a similar problem in chapter IV, which is there called *the problem of demarcation*.

4.0 Determining the alternatives

In any case, let us go back to the five steps. In the first step, the agent is supposed to reach a conclusion about her alternatives. What conditions must this conclusion satisfy? The Ideal, the Pure and the Reasonable subjectivist approaches all give this question different answers. Before discussing these answers, I will comment briefly on the notion of an 'alternative-set'.

As I have stated it, AU's criterion of rightness does not really refer to 'alternatives'. But if an agent is to use the deliberative approach as a method of decision-making she must make a *representation* of her alternatives. She must determine the *relevant* alternative-set, relative to which she is to maximise expected utility. An action maximises expected utility only relative to such a set.

Lars Bergström, in his (1966), has discussed the notion of 'alternative-sets'. According to Bergström, an action may be a member of different alternative-sets. In practical deliberation, determining the agent's *relevant* alternative-set is crucial. *Which* set one ought to consider, then, becomes pressing because an action may maximise expected utility relative to one set, but not relative to another. Bergström presents a number of suggestions of criteria of relevance for alternative-sets (relative to his version of consequentialism)[15].

One kind of criteria suggested by Bergström is 'personalistic'. According to 'personalistic' criteria of relevance, the alternative-set for an agent in a situation is determined through references to characteristics of an agent:

> […I]n order to determine the alternatives in a given case we might take somebody's intuitions or perceptions as a criterion or point of departure. Criteria which are constructed in this way may be called "personalistic criteria". For example, it might be held that the alternatives open to P in

done her or his best (in the situation) to consider relevant information, and to process it in a consistent manner. She or he should also have done her or his best to make an appropriate representation of the alternatives (whatever that may mean)." (Tännsjö, 1998, p. 35)

[15] Cf. Bergström (1966).

> S are those actions which P is aware of and believes that he has it in his power to perform in S. (Bergström, 1966, p.47)

This fits my purpose in this essay[16]. According to the Ideal subjectivist approach the relevant alternative-set contains all those actions that an *ideal* AU-agent is aware of. Since the agent is omniscient this set can be said to be *complete*. Obviously, it also satisfies the formal constraints, whatever these are taken to be, of being an alternative-set, because an ideal AU-agent does not make formal mistakes. This set will also be *normatively* plausible relative to AU. An ideal AU-agent does not fail to represent any of her alternative actions. Thus, the set is normatively plausible because it is complete.

The theoretical possibility of constructing these alternative-sets is not of much use for AU-agents that are not cognitively ideal, but possess cognitive abilities just like ordinary human agents, when *they* want to represent their alternatives in order to try to maximise expected utility. They will not be able to know which actions this complete set contains. These alternative-sets are not part of a sufficiently practicable method of decision-making for human AU-agents.

Let us turn to the Pure subjectivist approach. According to this approach, the relevant alternative-set contains all those actions which the agent *actually* is aware of. This could be interpreted in such a way that it is a necessary and sufficient condition for an alternative to be a 'pure subjective alternative' that it is represented in the agent's own decision-matrix. If the relevant alternatives are taken to be alternatives that the agent acknowledges as possible, the principle of maximising expected utility could be restated thus:

An action, a, maximises expected utility for P if, and only if, there is no alternative a' such that (i) a' is performable by P, (ii) a' is taken by P to be performable by P, and (iii) a' has a greater expected utility for P.

The Pure subjectivist approach has inevitable *objective* connections to the world. For example, according to this approach, the relevant alternative set is the set that the agent takes to be relevant, i.e. the alternatives believed to be performable, by the agent in the situation. But this might include alternatives which are actually not performable by the agent in the situation. If these alternatives figure in the calculation, that might distort the result. The agent may come to believe that an alternative maximises expected utility even if this is not so.

Bergström argues that the criteria of relevance of this pure subjectivist kind are 'rather arbitrary'.

[16] Other types of criteria may be more plausible if the notion of alternatives figure in the statement of a criterion of rightness. But I am speaking of methods of decision-making here. Here, some 'personalistic' criteria seem to be more relevant.

> If people have different perceptions of the alternatives in a given case, it is hard to see why the agent's perception should always be decisive. Moreover, in some cases there may actually be strong normative reasons for saying that the subjective alternative-set is *not* relevant; the criterion may lead to unacceptable normative conclusions […]. As far as I can see it would be normatively unreasonable always to rely upon the agent's perceptions, partly because the agent may be stupid, unimaginative, and so on. (Bergström, 1966, p. 48)

It seems hard to find reasons for the agent, within a subjectivism of this kind, to try to search for more information Why is it sometimes reasonable to think more carefully about these issues? Within this kind of subjectivism it is only reasonable to search for other alternatives, presently overlooked, if the agent's original alternative set already contains the alternative of *searching for other alternatives*. An unimaginative and narrow-minded attitude on the part of the agent can thus be a source of problem. If an agent believes that she has only two options in a situation: *a*: To kill Fred and *b*: To kill Tom (she does not believe that she can refrain from killing in the situation), then she ought to do either *a* or *b* in the situation. This holds regardless of the fact that she can refrain from killing (in the sense that if she tried, she would in fact succeed). A defender of this kind of global subjectivist maximisation of expected utility can be content with the question "Ought I to do *a* or *not-a*?" If the agent believes that she has stronger reasons for *not-a*, then all she has to do is refrain from doing *a* and has thereby done the "right" thing, i.e. the thing that the method prescribes. *How* she did *not-a* need not be relevant here (this depends on whether she believes that not-a can be performed in different ways).

The Ideal subjectivist criterion of relevance of alternative-sets imposes too high demands on the AU-agent's cognitive abilities to be of any practical use to human AU-agents. Realistic AU-agents are far from omniscient. The pure subjectivist criterion, on the other hand, requires too little. This criterion, because it simply takes the agent's own perceptions about which alternatives she is facing at face value, is both arbitrary and may lead to normatively misleading alternative-sets. Is it possible to give an account of the relevant alternative-set that occupies a middle position? Bergström presents a suggestion:

> For example, it might be suggested that *A* is the relevant alternative-set for *P* in *S* if, and only if, *A* is an alternative-set for *P* in *S*, and for every ideal observer *O*, if *O* had carefully considered the question of what alternatives are open to *P* in *S*, and if *O* had at least as much information as anybody else about *P*, *S*, and the consequences of different actions performable by *P* in *S*, then *O* would have been willing to accept the

> view that the members of *A* are the alternatives open to *P* in *S*. (Bergström, 1966, p. 48)

In order for this to be a plausible criterion for a less than ideal AU-agent, given the fact that we are after a criterion that will help the agent guide her actions, the ideal observer *O* ought not to be much different than the AU-agent herself. Could this be taken as a plausible middle position for the Reasonable subjectivist approach? According to the Reasonable subjectivist approach the determination of a relevant alternative-set ought to be governed by good reasons. Should we be satisfied if the agent has 'carefully considered the question' and has 'at least as much information as anybody else' about the facts of the situation and the consequences of the different actions? *How* carefully should the situation be considered? It seems that an AU-agent with normal fantasy and normal capacities of imagination can go on and on finding ever new alternatives. If an agent faces an innumerably number of alternatives, then there seems to be insurmountable problems with this approach. A finite human agent cannot represent an innumerable number of alternatives and cannot therefore set up a decision-matrix. But determining just *how* careful and how much time and effort she ought to spend trying to determining her relevant alternatives, in a non-arbitrary manner, seems very difficult. Again, the only plausible non-arbitrary answer to this problem seems to be "The optimific amount of time and effort!" but then again this answer does no help the agent in guiding her actions in practice.

A passage from Bergström (1966) seems to hold even for this criterion of relevance:

> […I]n practice *we can probably never know what the alternatives are* in a given case (for it seems that we can never have a complete knowledge of all the actions which are performable by a given person in a given situation and of all the consequences of these actions[…].) (Bergström, 1966, p. 54)

Determining the relevant alternative-set according to the ideal subjectivist standard makes this task highly impracticable for human agents. Human agents do not possess the cognitive abilities required in order to make this procedure sufficiently practicable. The pure subjectivist solution to this problem is arbitrary. Taking (human) agents actual beliefs about what their relevant alternative-sets are at face value is often liable to be seriously misleading. Single-mindedness, bias, wish-full thinking and lack of imagination are just some of the factors that would be allowed to influence the determination of the relevant alternative-sets. Obviously, these sets run a serious risk of being construed in a normatively misleading way. And this might threaten the validity of this approach. The Reasonable subjectivist approach also runs into trouble. It is very unlikely that human agents will be

able to justify their choice of taking a particular alternative set to be the relevant one. Deciding the question of just *how* careful they would have to consider different possible alternative-sets in order to come to a decision as to which set is the relevant one, will inevitably be arbitrary. And it is hard to see how AU's criterion of rightness might be of any help here.

4.1 Moore's criterion of relevance

There are utilitarians that have tried to confront the question regarding how to determine one's alternatives within the deliberative approach. Moore was aware of the problem; "…no one has ever attempted to exhaust the possible alternative actions in any particular case." (Moore, 1903, p. 199) Moore's attempt towards a solution to this problem is to limit the scope of the inquiry to *alternatives likely to occur to any one*, i.e. taking only alternatives of which any agent is likely to think of as the relevant set of alternatives to consider for practical purposes[17]. Let us examine this "[…] humbler task which may be possible for Practical Ethics." (Ibid. p. 199)

> With regard to these they [the ethical philosophers] may possibly have shewn that one alternative is better, *i.e.* produces a greater total value, than others. (Ibid. p. 199)

This is a radical change in approach. Moore is careful to insist that answering this question will not tell us what our duty is. All the same, "[…]this question, limited as it is, is the utmost, to which[…] Practical Ethics can hope to give an answer." (Ibid. p. 204)

Moore tries to justify this limitation with an additional argument, besides our cognitive shortcomings:

> […A]nd since we may also know that, even if we choose none of these, what we shall, in that case, do is unlikely to be as good as one of them, it may thus tell us which of the alternatives, among which we *can*[18] choose, it is best to choose. If it could do this it would be sufficient for practical guidance. (Ibid. p. 201)

Through the introduction of an assumption to the effect that if we do not choose one of the alternatives likely to occur to any one, we are unlikely to do as good as we would do by choosing *one* particular alternative of this set, Moore is able to argue that we can know what would *probably* be the best alternative to choose, of *every possible* alternative. To my knowledge, Moore never gives any reason for believing that what we shall do if we do not choose

[17] Ibid. p. 199.

[18] Moore is talking here about one specific sense of 'can' which implies that an agent 'can' perform an action only if the idea of performing the action *might* occur to her. Moore is rather unclear here, but it seems that he means to say, furthermore, that an action *might* occur to an agent only if the idea of performing the action is such that it is likely to occur to anyone. (Cf. Ibid. p. 200-201)

one of the alternatives likely to occur to anyone, 'is unlikely to be as good as one of them'. This bold assumption appears to be based on a very optimistic picture of moral agents or their situation of choice. What is the reason for believing that our 'alternative set' (in this specific sense) is special in this respect? Without any argument for this assumption, I see no reason to accept it[19]. Why should the relevant comparison be with what we would, in fact, do if we did not choose one action from our favoured set of alternatives? The relevant comparison is how close our choice approximates the AU standard. Even if it is true that one of our alternatives is better than any of the other, we do not know which of our alternatives it is.

Even if we, for the sake of argument, accept what Moore says here, this is hardly 'sufficient for practical guidance'. 'Alternatives likely to occur to any one' is vague. It could be interpreted in different ways. Normal agents presumably view situations of choice in somewhat different ways, acknowledging different alternatives in similar situations. Different interpretations can exploit this fact to different degrees.

One possible interpretation of the proposition would make the relevant set relatively small; the relevant alternative set is the set of which it is true that any normal (or representative) agent is likely to either have thought, is thinking or will think of *all* the alternatives in the set in relevantly similar situations.

Another interpretation would make the relevant set relatively large; the relevant alternative set is the set of which it is true that any normal (or representative) agent is likely to either have thought, is thinking or will think of *at least one* of the alternatives in the set in relevantly similar situations.

It is doubtful if this even makes any sense, but suppose that it does. Then it would probably mean that alternatives such as 'start a world war', 'to use all of my spare time working for needy and homeless people in my hometown' and 'devote all of my future life to strive towards world peace' are excluded from consideration in almost every situation of choice.

But from the point of view of AU, we would need a reason for this limitation. What justifies us in considering just this particular kind of alternatives? On what grounds are we to count alternatives falling outside this restriction as irrelevant? I can see only one reason that would be in line with or get support from AU for doing this. It is that agents, as was mentioned above, somehow 'automatically' are such that they, as a matter of fact, have alternative-sets which contains a 'sufficiently' good alternative, i.e. sets that are such that if an agent uses these kinds of alternative-sets, the effects of

[19] Perhaps someone would like to defend a biological/evolutionist ground for this assumption, arguing that the ability to automatically discriminate among alternatives in this way, has a high survival-value. I see no reason for believing that *this* ability would have any survival-value whatsoever. Remember that 'the best' is here understood in terms of maximum intrinsic value in the universe, and this seems to have very little to do with the survival and successful reproduction of agents' genes.

doing so would help the agent approximate the overall goal of AU, in the long run. This empirical claim stands in need of just as much justification as the claim for which it is a reason in the first place.

A further point should be noted though. One of the *desiderata* of chapter II, was practicability. According to this desideratum, an AU-agent has reason to prefer a method of decision to another if the AU-agent trying to adhere to the method succeeds to a greater extent than trying to adhere to the other method. Moore's way of solving the problem of relevant alternatives seems to influence the practicability of the subsequent method of moral decision-making, i.e. the restricted deliberative approach. If the relevant alternatives are restricted in some way, reducing them from an uncountable number to a much lesser number, then it seems that it would, at least *prima facie*, increase the AU-agent's chances of maximising expected utility relative to this alternative set. Moore claims that this way of reducing the relevant alternatives in the situation would leave the deliberating agent with only a few alternatives to consider in situations of choice:

> It [ethics] may, however, hope to decide which among *one or two* such possible actions is the best: and those which it [Ethics] has chosen to consider are, as a matter of fact, the most important of those with regard to which men deliberate whether they shall or shall not do them. (Ibid. p. 201. My italics.)

Even if this were true, i.e. that Moore's simplifying strategy had this effect and thereby increased the practicability of the method, it can also be taken to threaten the validity of the method by excluding some alternatives as irrelevant that, for all we know, might sometimes be the actions that make the agent approximate the overall goal of AU. The restriction will make whatever support the method of decision might get from the criterion of rightness rather weak. We have no good reason for believing that this restriction will give the agent an alternative-set which includes an action which would enable the agent to choose in a way that would be a good approximation of the overall goal of AU. This is a problem, but I will not pursue this point any further. The indeterminate character and the validity problem of this restriction should be kept in mind.

5.0 Determining the outcomes

Let us turn to the second and third steps of the method sketched by the deliberative approach: To identify the outcomes of one's alternatives, and to determine their utilities. There is no consensus regarding how 'outcome' should be understood. Some opt for a causal analysis and others for a counterfactual one. I believe, however, that the arguments presented in this subsection can be applied regardless of what stance we take here. Nevertheless, for simplicity's sake, I will henceforth use the language of

counterfactual analysis. I will take the outcome of a particular action to be the totality of what happens if the action is performed and is such that it would not have happened if the action had not been performed. I will, furthermore, assume that every action has a determinate outcome. I will not present any argument for this, but the assumption seems close at hand for a proponent of AU[20].

We are good at predicting the effects of many of the events that we examine. But in making these predictions we examine the effects with regard to criteria of relevance which we set up relative to a specific and restricted question. For example, we can examine the effects of smoking in terms of contracted lung-cancer. Based on empirical studies of the statistical correlation between smoking and lung-cancer in some population we are able to predict that an increase in smoking by a certain rate in the population will result in an increased rate of lung-cancer. This ability of ours does not take us far when it comes to prediction the *total* outcome of an action though. No one would even begin trying to determine the *total* effects of a populations smoking habits. The kinds of factors that can be relevant to the level of well-being of sentient beings are very numerous. Obviously, the problem increases when the processes we are trying to predict involve, at least in part, the actions of other human beings.

Notice that these problems are relatively independent on which view on utility we presuppose here. If utility is given a hedonistic or preferentialistic interpretation, many of the same problems will arise. Suppose, which sounds plausible, that the sum total of hedonistically valuable states of affairs as well as the amount of satisfied or dissatisfied preferences are determined by the fact that certain empirical facts obtain. Then it seems that we have to be justified in our beliefs as to which states of affairs are brought about by our different actions, in order for us to be justified in our beliefs as to the degree of preference satisfaction or hedonistically valuable states of affairs that the situation "contains". Hedonic states (other than our own), as well as preference satisfaction (sometimes even including one's own), is not directly open to observation. We have to make the detour over the empirical facts of the matter to get the information we want regarding these issues. But the natural facts that determine these 'evaluative facts' are shattered over vast areas of space and time and are often very hard to gain justified beliefs about.

This involves, in addition to the more general problem of foreseeing the outcome of a particular action and its alternatives, the much discussed problem of intra- and interpersonal utility comparison. We can assume, for the sake of the argument, that the different *theoretical* problems have been solved. These problems are different depending on what theory of value one

[20] If one does not accept this assumption, the outcomes of actions are not determinate. This means that the normative status of the alternatives may not be fixed. For the sake of argument I will assume that the assumption holds true.

defends. For hedonism, providing a common hedonistic zero-point and taking the smallest noticeable (or sub-noticeable) difference of wellbeing as our common denominator, can be taken to solve the theoretical problem of comparison[21]. For preferentialism, we can, for example, suppose that every preference-holding subject entertains preferences that could be compared in a 1:1 ratio in terms of a common standard unit and a common zero level point. *If* this could be done, enabling us to compare the preference satisfaction as well as the preference frustration of all preference-holding subjects and to sum them up in an overall measure of preference satisfaction, then we could be taken to have solved the theoretical problems of interpersonal utility comparison. The *practical* problem, however, seems insurmountable whichever theory of value we consider. To do an actual calculation with any reasonable accuracy, one would need to have justified beliefs about how an action would affect the level of well-being, or preference satisfaction, of everyone affected by the outcome of that action[22]. The sheer amount of data required to be processed, in quantitative terms, seem to present an insurmountable obstacle for agents with normal cognitive capacities.

Even if we had the time to *question* all affected parties[23], what reasons do we have for believing that they would give a correct report[24], or even tell the truth? The problem even arises within ones own person. How many noticeable (or even sub-noticeable!) differences in wellbeing would a day by the sea cause in me, as opposed to a day working with this thesis? The possibility of actually performing the required comparison in practice is bleak, to say the least.

We are dealing with *prediction* of events in the future rather than *descriptions* of events that have already occurred. Making such predictions is often very hard. What is the outcome of my going to the cinemas tonight? I might meet an old friend if I do. But then again I might not. Perhaps I will happen to meet the friend if I go to a restaurant instead? Perhaps this will make him glad, or sad. What is the difference in terms of my own and others well-being if I go to the cinema as compared with if I go to the restaurant? If I, instead of spending the money on personal amusements, where to send them to Oxfam, how would this affect the balance of well-being in the universe? What if I send the money to MSF instead? What would happen then? From the perspective of an agent trying to decide what to do, every outcome considered by her is associated with uncertainty of the same kind, although the degree of uncertainty can vary.

---

[21] Cf. Tännsjö (1998), p. 69-70.

[22] Of course, to determine the normative status of that action, one would need the corresponding beliefs regarding the affects of the alternatives to that action too.

[23] Animals could not even be questioned, so regarding actions affecting non-humans the problem seems even greater.

[24] This is hardly something that could be verified.

The Ideal, the Pure and the Reasonable subjectivist approaches give different criteria of relevance for the outcomes to be represented in the deliberation process. According to the Ideal subjectivist approach, the relevant outcomes to consider in deliberation are the actual outcomes of the alternative actions. Remember that in order to live up to the Ideal subjectivist approach, the AU-agent must reach the decision that maximises expected utility given true beliefs about her alternatives, utilities and so forth. If there is genuine uncertainty 'out there' in the world, i.e. if there are different possible outcomes of an action, then this is something that the ideal AU-agent knows. And she knows the probability of each of these possible outcomes. If, on the other hand there is one and only one thing that will happen if the action is performed, then the ideal agent knows *this*. An ideal AU-agent will make only true (and complete) representations of the outcome(s) of her actions. The relevant representations are thus, under the Ideal subjectivist approach, the true representations. Once again, an AU-agent would have to know whether determinism is true or false in order to make a correct representation of the outcome(s). Human AU-agents do not know if determinism is true or false, and cannot therefore know that their representations are correct. Ideal AU-agents, on the other hand, have no problem in assigning utilities to the outcome(s) that satisfy the criteria of relevance of the Ideal subjectivist approach. An ideal AU-agent would know the utility of each of the outcome(s) she is considering.

On the Pure subjectivist approach, the relevant outcome-representations to figure in deliberation are the agent's actual representations of the outcomes that she believes that her alternatives have. Thus if someone *believes* that Thalidomine is harmless to use during pregnancy, her outcome-representation (of the choice of taking or not taking the medicine) will not include any representation of the outcome of giving birth to a disabled child.

The general point is that despite the fact that one is doing the best one can (in this case trying to maximise expected utility) things can go really bad. The agent might easily fail to approximate AU's overall goal. She might fail because the beliefs entertained by the agent about the likely outcomes of actions are false. The possibility of making this type of mistake when forming beliefs about outcomes seems quite significant. Ignorance, wishful-thinking and misinformation are only some of the sources of such mistakes. The complexities and nature of the type of facts of which the outcomes consist make it even harder. Here the same problem that became pressing in the last section, *motivating* a search for new alternatives, resurfaces. Motivating why it sometimes seems to be a good thing to try to represent (take account of) other possible outcomes of one's actions, which has been overlooked by the agent, is hard under the Pure subjectivist approach. According to the Pure subjectivist approach, it is the agent's beliefs about the outcomes that are

relevant in her deliberations. And these, of course, can and probably will often be off the mark.

Demanding that the relevant outcome representations are true, as the Ideal subjectivist approach does, sets too high a standard for AU-agents with cognitive abilities like a normal human being. But taking the relevant outcome representations to be the *actual* representations of the AU-agent, as the Pure subjectivist approach does, seems to set the standards to low. The AU-agent might often entertain beliefs about the possible outcomes of alternative actions that are due to false beliefs, wishful-thinking or other biases. Is it possible to provide a position between these two extremes?

According to the Reasonable subjectivist approach, the relevant outcome representations ought to be based on justified beliefs about possible outcomes of alternative actions. An outcome should only be considered relevant to include in the deliberation process if the AU-agent has good reasons for believing that it is a possible outcome of one of her alternative actions. Furthermore, justified beliefs are also needed about every possible outcome of the alternative actions in the agents relevant alternative-set. Is it possible to have justified beliefs about how much utility the outcome of an action contains? What is needed according to the Reasonable subjectivist approach is justified beliefs about, e.g. the different levels of well-being or the amount of preference satisfaction of all sentient beings affected by the actions being, or not being performed. It is hard to see how an AU-agent, with normal cognitive abilities, could gain justified beliefs about the total effects of the possible outcomes of all alternative actions in a situation.

5.1 Sidgwick's restriction

Some utilitarians have tried to deal with this problem. For example, Henry Sidgwick makes the following suggestion:

> [W]e may perhaps reduce the calculation within manageable limits, *without serious loss of accuracy*, by […] neglecting the less probable and less important contingencies; as we do in some of the arts that have more definite ends, such as strategy and medicine. For if the general in ordering a march, or the physician in recommending a change of abode, took into consideration all the circumstances that were at all relevant to the end sought, their calculations would become impracticable; accordingly they confine themselves to the most important; and we may deal similarly with the Hedonistic art of life. (Sidgwick, 1874, p. 131-132. My italics.)

And further:

> It is, however, conceivable that by induction from cases in which empirical measurement is easy we may obtain generalisations that will give us more trustworthy guidance than such measurements can do in

> complicated cases […]. I am willing to hope that this refuge from the difficulties of Empirical Hedonism[25] may some time or other be open to us: but I cannot perceive that it is at present available. There is at present […] no satisfactorily established general theory of the causes of pleasures and pain […]. (Ibid. p. 178)

Needless to say, the 'most important' clause in the first quotation is highly question begging. Knowing which circumstances are the most important *presupposes* that something similar to a complete calculation has already been done. And the manoeuvre is supposed to bypass precisely this. Further, the analogy seems erroneous. The goals of marching and medicine are naturally restricted in a way that the hedonistic goal is not. A doctor does not, and indeed probably ought not to, deliberate as to how curing her patient from her illness will affect the amount of overall realized utility. The doctor should try to cure her patient's illness if this is what her patient wants. When it comes to ordinary illnesses, of which the doctor has the means and the knowledge required to cure her patient, the criterion of "rightness", or of "successful treatment" as well as the method by which to accomplish it, are well established[26]. The same could be said about strategy. The successful general is probably the one who calculates only a few battles ahead. He sets for himself goals like 'winning the battle', or 'promoting the chances of victory in the war'. Given these restricted and determinate ends, different lines of conduct can be evaluated with sufficient accuracy for his purposes. On the other hand, a hedonist has the whole world's happiness as his end and a restriction of consideration may lead him grossly astray.

Furthermore 'neglecting' less probable contingencies seems to go against the spirit of the deliberative approach. Neglecting to take into account the risk, even if it is very slim, of a meltdown in a nuclear power-plant, presumably, is often a bad idea.

5.2 Moore's guidelines

Furthermore, Moore suggests three concrete guidelines which he thinks the agent should use when trying to maximise expected value[27]. These principles are supposed to be used as some sort of guidelines when an agent is deliberating as to which action she ought to perform[28].

---

[25] This method is, roughly, what I have call the deliberative approach.
[26] Cf. Griffin (1994) p. 181.
[27] The reader is to keep in mind that this is *my* interpretation of the idea.
[28] Sidgwick introduced similar restrictions: "For the practical application of this theoretical impartiality of Utilitarianism is limited by several important considerations. In the first place, generally speaking, each man is better able to provide for his own happiness than for that of other persons, from his more intimate knowledge of his own desires and needs, and his greater opportunities of gratifying them. And besides, it is under the stimulus of self-interest that the active energies of most men are most easily and thoroughly drawn out: and if this were removed, general happiness would be diminished by a serious loss of those means of happiness which are obtained by labour; and also, to some extent, by the diminution of the labour itself. For these reasons it would not under actual circumstances promote the universal happiness if each man were to

1. "That a lesser good, for which any individual has a strong preference […], is more likely to be a proper object for him to aim at, than a greater one, which he is unable to appreciate." (Moore, 1903, p. 215-216)
2. "[…I]t will in general be right for a man to aim rather at goods affecting himself and those in whom be (sic! [he]) has a strong personal interest, than to attempt a more extended beneficence." (Ibid. p. 216)
3. "Goods, which can be secured in a future so near as to be called 'the present,' are in general to be preferred to those […] being in a further future […]." (Ibid. p. 216)[29]

I think that the most plausible way of understanding these principles is to take them to provide more specific versions of, or shortcuts towards satisfying, the principle of maximising expected utility. The idea is that the general adherence to these principles, when deliberating over which action to perform, helps the agent realize more intrinsic value than any alternative approach. These rules are all motivated by their assumed tendency, if observed, to increase the probability of attaining the goods of which they are means for securing:

> We have, however, not only to consider the relative goodness of different effects, but also the relative probability of their being attained. A less good, that is more likely to be attained, is to be preferred to a greater, that is less probable, if the difference in probability is great enough to outweigh the difference in goodness." (Ibid. p. 215)

Is it true that these practical principles provide the best means to maximise expected value? Their truth or falsehood is contingent on several different empirical assumptions, regarding the agent's psychology (motivational set-up and dispositions) and the agent's abilities to affect states of affairs at temporally and spatially more or less remote distances. Are these assumptions plausible?

Of course, a judgement about this must be extremely speculative. Short of an actual empirical investigation, it seems that we have to resort to more or less qualified guesswork. I will only present some possible counter-arguments to these principles, which, at least in a number of situations, seem to make the agent adhering to these principles fail in maximising expected utility. This does not establish the utility of their being generally adhered to, though. But

---

concern himself with the happiness of others as much as with his own." (Sidgwick, 1874, p. 431) If this restriction could be justified, several things would be won. It can explain and justify our bias for ourselves and our near and dear. It would also simplify the deliberation in a way that would increase the methods practicability.

[29] Cf. also the discussion of epistemic discounting in section 5.4.

without a positive defence of these principles, it is hard to see why we should endorse them in the first place.

Principle 1 is not entirely clear. The agent is supposed to try to attain (or aim at) goods, in our case utility, for which she has a strong preference or 'natural inclination'. As it turns out, principle 1 seems to constitute the reason for accepting principle 2, rather than being an independent practical guideline[30]. Moore assumes that "*almost every one* has a much stronger preference for things which closely concern himself[…]and those in whom be (sic! [he]) has a strong personal interest[…]" (Moore, 1903, p. 216. My italics.). This, in turn, is why such agents ought not to 'attempt a more extended beneficence'.

In our day and age, and for many agents in the rich and affluent countries, principle 2 is implausible. It would seem to constitute an illegitimate limitation of our moral outlook. It could be argued that though we are in a favoured position to do good to ourselves and our near and dear, yet our ability to prevent great evils befalling distant strangers is even more favourable. For example, a handful of dollars can save lives and help persons suffering from absolute poverty, but have little impact on our own lives. In many situations, it seems, the direct opposite of the principle is likely to be a more valid approach.

Another problem with principles 1 and 2 is that of potential bias. It seems that this kind of 'egocentric' method of decision runs a particularly high risk of including elements of egoistic bias and wishful thinking. Moore stresses this problem when discussing the dangers of agents breaking generally useful moral rules, making an exception in their own case. When Moore argues for the present guidelines, however, he does not mention this problem at all.

If Principle 3 is understood in a weak sense, as dealing only with goods of equal quantity, then it seems to be trivially true. AU is neutral to time in the sense that *when* some quantity of wellbeing is realized is irrelevant. Principle 3 on this interpretation only states that an agent ought to prefer a *certain* good to a probable good of equal magnitude. A more interesting interpretation, and also the one Moore endorses is, of course, discounting a future good in proportion to the improbability of its being attained. And this is implied by the principle of maximising expected utility. Principle 3 can also be given an even stronger interpretation, in line with the 'ripples in the pond'-postulate that is the topic of the next section.

5.3 The 'ripples in the pond'-postulate

There are different views on how far-reaching the outcomes of actions are. Do the outcomes of actions reach far into the distant future, perhaps to the end of

[30] Agents with a strong preference for the wellbeing of strangers or animals, ought of course accordingly to strive for these goods according to principle 1. In this sense principle 1 is more fundamental than principle 2, which seems to be only a more specific instantiation of principle 1, for 'almost every one'.

time, or do they gradually disappear like the ripples in a pond after a stone has been dropped into the water? Some philosophers, e.g. Smart and Moore, thought that it is necessary to assume that the latter view is true in order to make utilitarianism 'workable':

> […W]e do not normally in practice need to consider very remote consequences, as these in the end approximate rapidly to zero like the furthermost ripples on a pond after a stone has been dropped into it. (Smart, 1973, p. 33)

And further:

> [...]we need some sort of 'ripples in the pond' postulate to make utilitarianism workable in practice.(Smart, 1973, p. 34)

A similar idea is presented already in Moore (1903), where he claims that:

> It is difficult to see how we can establish even a probability that by doing one thing we shall obtain a better total result than by doing another. (Moore, 1903, p. 201)

The difficulty lies in the fact that our actions seem to have consequences which we, for different reasons, know very little, if anything, about. More specifically, it is our inability to know temporally far-reaching (Moore even speaks of an 'infinite future') consequences of our actions. To solve this problem, Moore proposes an 'assumption' or 'large postulate':

> [...I]f a choice guided by such considerations is to be rational, we must certainly have some reason to believe that no consequences of our action in a further future will generally be such as to reverse the balance of good that is probable in the future which we can foresee. This large postulate must be made, if we are ever to assert that the results of one action will be even probably better than those of another. (Ibid. p.202)

The reason for accepting the 'ripples in the pond'-thesis is primarily practical. *If* this postulate could be defended, *if* this is a characteristic of the outcomes of our actions, then it could be argued that the deliberations would be substantially simplified. Although both Moore and Smart seems to make an ontological claim (this is, anyhow, the way I will interpret them), I will also discuss an epistemological 'version' of this postulate in section 6.3 below. Consider the following claim made by Smart:

> The necessity for the 'ripples in the pond' postulate comes from the fact that usually we do not know whether remote consequences will be good or bad. Therefore we cannot know what to do unless we can assume that

> remote consequences can be left out of account. *This can often be done*. (Smart, 1973, pp. 33-34. My italics)

Moore presents some reasons for accepting this:

> As we proceed further and further from the time at which alternative actions are open to us, the events of which either action would be part cause become increasingly dependent on those other circumstances, which are the same, whichever action we adopt. (Moore, 1903, p. 202-203)

And further:

> The effects of any individual action seem, after a sufficient space of time, to be found only in trifling modifications spread over a very wide area, whereas its immediate effects consist in some prominent modification of a comparatively narrow area. Since, however, most of the things which have any great importance for good or evil are things of this prominent kind, there may be a probability that after a certain time all the effects of any particular action become so nearly indifferent, that any difference between their value and that of the effects of another action, is very unlikely to outweigh an obvious difference in the value of the immediate effects. (Ibid. p. 203)

Notice that Moore is talking of an ontological assumption rather than an epistemic one. It is an assumption about how the world is constituted, not about what we can know about it. There are good reasons for believing this assumption to be false, at least for some actions.

What is wrong with Moore's assumption? A general objection to this kind of assumption is that it seems strange to assume this kind of symmetry between actions. Independently of how far-reaching, 'prominent' or 'important' the consequences may be, what is the reason for assuming that the consequences of different actions share these characteristics to the same degree? A more plausible assumption, at least *prima facie*, is that different actions have effects that differ with respect to these factors. An action can have its most 'prominent' effects just after its performance, e.g. the lighting of a firecracker or after a long time, e.g. the burying of nuclear waste (when the earthquake eventually occurs). It should also be noted that although it is possible that according to Moore's theory of intrinsic value, 'trifling modifications' would not affect the realization of intrinsic value in any significant way, this can not equally plausibly be maintained on other possible axiological theories in combination with AU, i.e. preferentialism or hedonism. Many small differences in wellbeing spread over a wide area can, for example, be *very* important.

In so far as the assumption rests on the metaphor of 'ripples in the pond', it is unhappy. When a stone is dropped into a pond, the subsequent ripples will all eventually disappear. Not of course into thin air, but the release of energy produced by the plunge will take different forms and the effects which is detectable by vision, i.e. the ripples, will eventually disappear altogether. This is not the view that Moore defended with regard to the effects of different actions. He explicitly defended the view that the effects diminish, eventually to consist only of 'trifling modifications'.

Although Moore did not use the metaphor of 'ripples in the pond', his assumption bears enough resemblance to this example to make a few comments appropriate. As Shaw points out, by modifying and questioning the character of this metaphor, the problem with this example will surface.

> Instead of a pebble, toss a live trout into the pond, and just as the initial ripples die out, the trout will jump out of the water yards away. (Shaw, 1995, p. 112)

Shaw also gives an example that is intended to show the falsity of Moore's assumption, the story of Son of Carl and Daughter of Deborah. Deborah Debtor owes Carl Creditor some money. She can either pay him the money, or shoot him. If she shoots him, Carl's would be future son, Son of Carl, is prevented from coming into existence. If he had come to exist, he would have become a deranged mass murderer. But if she shoots him, she will get caught, and due to her subsequent imprisonment she thereby misses her one opportunity of bringing a daughter, Daughter of Deborah, into the world which would become the "secular savior[31] of the twenty-first century, bringing far more good to the world than Son of Carl takes from it." (Shaw, 1995, p. 111) Whatever alternative she chooses, her action will have further effects more prominent than the immediate ones. This and other examples, e.g. burying nuclear waste, show the implausibility of Moore's assumption. The general point shown in these kinds of examples is that actions (or omissions), which appear to be relatively insignificant, are often necessary conditions for future actions, which can have more 'prominent' outcomes.

Moore's particular way of discounting future effects of actions is a mistake. It assumes that the world is constituted in a way that it probably is not. We must of course find a way of dealing with our epistemic limitations, but the question of how far-reaching the effects of our actions are is an empirical one and cannot be plausibly dealt with by way of a mere *a priori* assumption. Moreover, available evidence seems rather to support the opposite claim to the 'ripples in the pond' assumption: many actions do indeed seem to have quite far-reaching effects. The rival view on outcomes is 'the thesis of

[31] The argument does not depend on the Daughter of Deborah being the 'secular savior of the twenty-first century'. Her coming to existence must only outweigh the bad effects of Son of Carl's existence.

endlessness'[32]. According to this view, the outcome of (at least some) actions may continue until the end of time (or at least until the end of the world). The question of which of these two views–if anyone– is true depends on, among other things, what sort of events that should be included in the outcome of an action. For example if my thinking of Socrates' suicide (right now) is a (very remote) outcome of Socrates' action of drinking the hemlock (in 399 BC), then it seems plausible to believe that 'the thesis of endlessness' is true for some actions. Which view on the outcomes of actions should we believe in? It seems impossible to give any conclusive reason for believing any one of these views to be the true account. Despite all of this, I think that there are good arguments for rejecting the ripples in the pond view, at least for some actions. It seems plausible to assume that at least some actions can have outcomes that extend very *far* into the future and that their 'most prominent' effects arise after a very long time. The following examples speak in favour of this last point: The government of some country decides to bury nuclear waste deep down in the primary rock. Everything looks good, but a hundred years later, there is an earthquake and the waste spreads in the eruption. Another example would be to have children. After a couple of hundred years, there would probably exist hundreds of people that would never have existed, had it not been for my decision to procreate. We cannot show either of these views to be true, but postulating that one is true in order to be able to use AU is not a plausible "solution". The important question is what kind of beliefs we can have about the outcomes of particular actions? Can we have justified beliefs about the *total* outcome of any action[33]? This seems highly unlikely.

5.4 Epistemic discounting

Perhaps we are justified in discounting effects in the further future on epistemic grounds. Both Smart and Moore justify their proposed discounting of far-reaching outcomes on the ground that there are no far-reaching outcomes of actions (or that they are insignificant or that they cancel each other out). These are all empirical claims. But we might perhaps justify the discounting on epistemic grounds instead. This way of discounting is suggested by defenders of versions of AU. Thus Kagan writes:

> (Of course it remains true that there will be a very small chance of some totally unforeseen disaster resulting from your act. But it seems equally true that there will be a correspondingly very small chance of your act resulting in something fantastically wonderful, although totally unforeseen. If there is indeed no reason to expect either, then *the two possibilities will cancel each other out as we try to decide how to act*.) (Kagan, 1998, p. 65. My italics.)

[32] Bergström (1966), p. 110.
[33] Cf. Lenman (2000) and Miller (2003).

Should we conceive of the deliberative approach as involving this idea? Is it a plausible idea? Lenman (2000) launches the following objection. Suppose that an agent can choose between two different courses of action X and Y. She does not know the total outcomes of either X nor Y. But she knows the following. One of the actions will have a disastrous outcome, while the other action will have a heavenly outcome. She also knows that if she performs X, on top of whatever total outcome X might have (disastrous or heavenly), part of X's outcome is that someone, for a brief moment will experience a pleasurable sensation.

Now, according to the idea of epistemic discounting, the agent ought to disregard the outcomes that she does not know anything about. This seems to leave the agent only with the knowledge of part of X's outcome. And she ought accordingly to perform X (a brief pleasurable experience is better than nothing).

Does the agent have a reason for performing X rather than Y? Does her belief in AU provide her with a reason for performing X? Lenman thinks that this would only be a *very* weak reason indeed, but he does not show that it is not a reason[34]. I would agree that the reason is in a sense weak but, and this is important, it is the strongest reason the AU-agent has in her predicament. A weak reason might well be one's strongest reason. The practical relevance of this point is, however, doubtful. On the one hand, it seems reasonable to discount the parts of an outcome that we do not know, or cannot have justified beliefs about. But in order to justify this discounting, with reference to AU, we would need reasons for believing that making this kind of discounting in our deliberations increases the validity of the deliberative approach. We are not in a position to justify this belief. Furthermore, if the agent cannot justifiably assign probabilities or estimate utility amounts, then the deliberative approach is not applicable. It cannot guide the AU-agent's actions.

6.0 Assigning probabilities

Let us turn to the fourth step: To assign probabilities to the relevant outcomes. According to Smart, the relevant probabilities to be used in the deliberation process are of the objective kind: "But until we have an adequate theory of *objective* probability utilitarianism is not on a secure theoretical basis." (Smart, 1973, p. 41) The notion of objective probabilities can be given different interpretations. There are logical views, relative frequency views and propensity views. I will not take a stand on the question of the plausibility of these views. The important thing for the Ideal subjectivist approach to maximising expected utility is that whatever truth or plausibility these different notions of probability may have, this approach demands of the agent

[34] Cf. Lenman (2000), p. 351-366.

that she knows the answers to these questions. On the Ideal subjectivist interpretation, the relevant probabilities are the subjective probabilities of a cognitively ideal agent[35]. If objective probabilities exist, this ideal agent would know them. A cognitively ideal agent will have true beliefs about the objective probabilities *if* there are any and whatever their particular nature is (propensities, frequencies or whatever). For this ideal agent there will be no discrepancies between objective probabilities (if there are any) and her subjective probabilities. Whether there exist objective probabilities or not, is not something I have to take a stand on here. However, on the Ideal subjectivist approach a human AU-agent, would have to assign only the 'true' probabilities to the different outcomes. She would have to *know* which view on probability that is true.

According to the Pure subjectivist approach the relevant probabilities are the agents actual probability assessments. There exists, in the literature on decision theory, accounts of how to construe subjective probabilities in terms of hypothetical lotteries. The idea is that our degree of faith in the truth of a proposition is reflected in our willingness to bet money on the proposition being true or false. Even if one should acknowledge the *theoretical* soundness of this approach, the *practical* relevance of this idea, however, is very doubtful. These methods are very time-consuming and it is doubtful if they are applicable in practice. The 'hypothetical betting approach' has serious drawbacks if considered as an approach whereby to determine subjective probability in practical situations of choice, whatever merits it possesses as a theoretical device. The account

> […] deals with an agent in a situation in which […the agent…] must place a series of bets on a […] set of initial statements as well as all negations, conjunctions, disjunctions and conditional bets that can be formed using these statements. […][The agent] will be expected to take conditional bets […] for all of the infinitely many bets that are constructible from the initial set. (Resnik, p. 69-70)

Consider an example. "If I send £100 to Oxfam, more utility would be realized than it would be were I to send the money to MSF." What is my degree of confidence in this statement being true? Determining this in accordance with the above mentioned method seems practically hopeless. Of course, the practical difficulties associated with this technique will, if employed by human AU-agents, affect the Ideal subjectivist- as well as the Reasonable subjectivist approach as well.

[35] Once again, it is only if determinism is false that an ideal agent would need probability assignments. Otherwise, she will know the outcome of every action with certainty.

Another problem for the Pure subjectivist approach is that the agent might not hold beliefs of probability that are specific enough. She might be unable to make up her mind as to the probability of a certain outcome. For example, probability assessments may have to be very exact. Suppose an action, *a*, is believed by an agent to have two possible outcomes. If things turn out for the best it will result in a net utility of 10 units, if not of 6 units. Suppose further that the agent believes that refraining to perform *a*, the performance of not-*a*, will lead to a net utility of 8 units. The agent is now trying to make up her mind as to the probabilities of success if she performs *a*. It seems as if the accuracy in this ascription is important here. If she sets the probability of success at .51 and that of failure at .49, then the expected utility of performing *a* is 8,04, and *a* ought to be performed. If on the other hand set the probability of success at .49 and that of failure at .51, then the expected utility of *a* is 7.96, and *a* ought not to be performed. Are there situations, in real life, in which the possibility of this kind of accuracy is attainable for moral agents?

Suppose that *a* is the giving of a present on a certain occasion. The agent is not sure how the gift will be received. Due to a previous quarrel between the agent and the potential receiver of the gift, the agent suspects that the recipient might come to believe that the agent is trying to suck up to her, not giving the present out of goodwill[36]. The agent believes that there is roughly a fifty-fifty chance of success, but this is only a rough approximation. The example shows that even the slightest vacillation in the determination of probabilities determines whether she ought or ought not give her friend the present. For an agent trying to perform subjectively right actions, this choice is not decided, if it's rationale is supposed to be the principle of maximising expected utility, as easily as it might be for an agent with a more main-stream conception of morality. Refusing to go into meticulous detail here would involve abandoning trying to maximise expected utility, leaving the action without the rationale supposedly provided by that principle. The fact that small alterations in the probability assignments, which the agent is hard pressed to give reasons for preferring one assignment rather than the other, can determine whether the action ought to be performed or not, tell against the practicability of the method. If she cannot make up her mind on this issue, she sometimes cannot perform the required calculations. This problem will arise also for the Reasonable subjectivist approach.

According to the Reasonable subjectivist approach the AU-agent must provide good reasons for her probability assignments. In order to assign a particular probability to a possible outcome of an action, she would have to have good reasons for believing that this outcome is just as probable as the assignment reflects. An AU-agent with normal cognitive capacities will be

[36] We assume that receiving the gift and believing it to be a way of sucking up, correspond to 6 utility units (the gift is nice so it is still a positive value). Receiving the gift and believing it to be given out of goodwill correspond to 10 utility units.

hard pressed to provide such reasons. Such agents are seldom in a position to have justified beliefs about the probabilities of all of the possible outcomes of the different alternative actions in a relevant alternative set. For one thing, the sheer number of possible outcomes seems to be too big to make the assignments of probability practically possible. The probabilities of the possible outcomes of one action must sum to 1 to obey the probability calculus. An outcome is a state of affairs. That an outcome is 'possible' means that the probability of this state of affairs occurring due to the performance of the action is somewhere between 1 and 0. It is certain to occur if the probability is 1 and it is certain not to occur if the probability is 0. For example, consider the action of sending £1000 to Oxfam. There are surely, for all I know, many *possible* outcomes of this action. One *possible* outcome of this action is that 10 lives are saved. Another is that 11 other lives are saved. Another is that 20 other lives are saved, that no life is saved and so on. Every empirically possible outcome of this action, taken to be relevant by the agent, would have to be assigned a numerical value, consistent with the numerical value assigned to the other possibilities. The greater the ability of the agent to discern different relevant outcomes of an action, the greater the number of (subjectively acknowledged) possible outcomes. The practical problems of actually working out a plausible assignment of probabilities to an action's possible outcomes can become very hard indeed.

Let us return to the example with the gift above. Suppose I have decided to assign the outcome that the gift is appreciated the probability of .51, and decide to give it to my friend. But before I give her the present, I ask myself if I have good reasons for assigning this probability to this outcome. What is my reason for assigning .51 instead of .49? I do not know the specific (psychological) mechanisms of my friend in such details the would enable me to say "My reasons for assigning the probability of .51 are stronger than my reasons for assigning the probability of .49!"

7.0 Calculating the expected utilities of each alternative

Finally, let us turn to the fifth step of the deliberative approach. An ideal agent will make no mistakes when assigning expected utilities to her different alternatives. On the Pure subjectivist approach human AU-agents can, however, make mistakes when trying to assign expected utility to different alternatives. She can make mathematical miscalculations. She may also fail to have the requisite beliefs about the utility and probability that is needed. (More about this in 8.1 below.) On the Reasonable subjectivist approach, the problem is serious. The problems that arose when trying to assign probabilities and utilities to outcomes that are justified, or well grounded, relative to the available evidence is often very hard. Hare suggests that the relevant probability and utility assignments ought to survive an exposure to logic and the facts:

> […A] full account would need to contain a method of weighing combinations of probabilities and utilities against one another, by asking which combination one would prefer, after exposure to logic and the facts. (Hare, 1981, p. 133)

Because the 'facts' of the matter are very complicated and often very hard to determine, i.e. that human AU-agents are hard pressed to gain justified beliefs about these facts, it will often or always be very hard to arrive at a point where the human AU-agent is justified in her assignments.

8.0 The validity and practicability of the deliberative approach

How does *the deliberative approach* score in terms of practicability? How likely is it that an agent who wants and sets her self the goal to maximise expected utility will succeed? Obviously, the answer to this question depends on what constraints are imposed on the beliefs and probability assignments of the deliberating agent, i.e. whether one defends *Pure subjectivism*, *Reasonable subjectivism* or *Ideal subjectivism*. In the remaining of this chapter I will try to say something about the validity and practicability of the different versions of the deliberative approach. Let us begin with pure subjectivism.

8.1 Assessing the Pure subjectivist approach

As the foregoing discussion seems to show, there are considerations that strongly suggest that the Pure subjectivist approach is relatively practicable. If *any* subjective belief held by the agent, no matter how ignorant or foolish it might be, together with *any* subjective probability assessment, however off the mark it might be (provided that it does not violate the probability calculus) is acceptable, then little can go wrong. The only source of error is, it might seem, that the agent fails to perform the required calculations correctly. Still, compared to more demanding versions, the Pure subjectivist approach is likely to be more practicable.

There is a further problem. The agent might not be able to make up her mind about probabilities, outcomes and utilities in a situation. If she does not have a clue about the probabilities, the Pure subjectivist approach might not be very practicable for the agent. This is also pointed out by Smith:

> For example, act utilitarians often advocate supplementing their principle just with the rule of maximizing expected utility. But to know what this rule prescribes typically requires an agent to perform a complex arithmetical calculation involving the probabilities and values of the possible consequences of each alternative action. Obviously not every decision-maker will feel certain what values or probabilities are to be assigned to the various possible consequences. Even those who can make such assignments may not feel certain that the result of the computation

> is mathematically accurate. […] All of these decision-makers are still subject to the Problem of Doubt, since they remain uncertain which act their principle (indirectly) prescribes. (Smith, 1988, p. 98)

So, the practicability of the Pure subjectivist approach might not be as high as one might first expect.

On the other hand, we have little reason to think that the Pure subjectivist approach is particularly valid. The decisions of an agent can represent successful attempts to follow the approach, even if they are based on beliefs and evaluations that are the result of wishful thinking, bias and lack of time. We have no reason to think that someone who makes decisions on such a basis will approximate the overall goal of AU.

8.2 Assessing the Ideal subjectivist approach

As for the Ideal subjectivist approach the degree of practicability and validity for normal human beings depends upon how successful normal human beings are in obtaining justified true beliefs. According to the Ideal subjectivist approach only justified true beliefs and assignments of probability and utility are allowed in deliberation. They are the relevant inputs. Ordinary human agents will inevitably entertain some beliefs that are false and/or unjustified. But perhaps some of their beliefs are true and justified. Only these beliefs are allowed in deliberation. The degree to which the beliefs of ordinary agents are true and justified will obviously be different for different agents but, in general, normal human agents are not likely to entertain such beliefs. Entertaining justified true beliefs about the possible outcomes of every alternative action in a situation, including their probability and the amount of utility associated with each and every one of them is often practically impossible for human agents. And even if it were *possible*, finding them out would probably consume so much time as to make the agent occupied with this particular problem for a lifetime. The criteria of relevance for the input of the deliberation, supplied by the Ideal subjectivist approach, clearly set too high standards for human agents. It is highly unlikely that an actual human agent can ever meet these demands. As a suggestion of an interpretation of how to construe the deliberative approach as a sufficiently practicable method of decision for AU, the Ideal subjectivist approach is a straw-man. The Ideal subjectivist approach has obvious and serious drawbacks in terms of practicability.

What should we say about the validity of the Ideal subjectivist approach, then? Is it likely that *if* an AU-agent could adhere to this approach that she would approximate the overall goal of AU?

Arguing that this approach makes an AU-agent approximate the overall goal of AU better than any other proposed method of decision-making has a certain intuitive appeal. Accordingly, it seems that the validity of this

approach would be relatively high. But it is not certain that this is the most valid approach. Suppose that the AU-agent can perform either *a* or not-*a*. Suppose further that *a* maximises *expected* utility. The AU-agent accordingly perform *a*. But it is still possible that *if* the agent had performed not-*a* instead, which does *not* maximise *expected* utility according to the ideal agent, then utility would *in fact* have been maximised. Even improbable outcomes sometimes occur. Accordingly, we cannot conclude that the method of maximising expected utility, even if adhered to, is the most valid, given the circumstances. Because this leaves open the possibility that things sometimes will not turn out actually to maximise utility, it is possible that another method, i.e. a method that does not consist in maximising expected utility, will actually, however unlikely this is, turn out to realize a greater amount of utility overall.

The Ideal subjectivist approach is perhaps a relatively valid method of decision-making for AU. But in order to establish this thesis it needs to be argued that by maximising expected utility over time, we actually achieve better results, than we would if we adhered to another method of decision-making. And this has not yet been showed. I will return to this question in the next section.

In chapter IV, Hare's method of decision-making is discussed. As we will see, the Ideal subjectivist approach has a role to play in this method. So in this sense, the approach is not a straw-man. But as an interpretation of the deliberative approach, for human agents, it is. Our human imperfections are clearly too great to make the method sufficiently practicable for us. Attempting to adhere to the Ideal subjectivist approach is not likely to succeed for normally endowed AU-agents.

8.3 Assessing the Reasonable subjectivist approach

The purpose of introducing the reasonable subjectivist approach is to provide an example of what standards a method of decision-making for AU should plausibly satisfy in order to get support from AU. Because human moral agents are wanting in certain respects, we need to make tradeoffs. We need to buy increases of practicability at the expense of validity. Could some version of the reasonable subjectivist approach be taken to constitute a plausible method of decision-making AU? How should we evaluate the reasonable subjectivist approach in terms of practicability and validity? Well, because the input is less ideal on this approach than on the Ideal subjectivist approach, it seems plausible to assume that the validity of this method would be less than the validity of the Ideal subjectivist approach. But, for the same reason, the practicability of this method seems to be higher. Perhaps we are more likely to entertain beliefs of the quality required by this approach. We can also compare this method with the Pure subjectivist approach. It seems plausible to assume that when the standard set on the input is raised, this would make the

method less practicable. This raising of the standards might also be taken to increase the validity of the method. Of course, it is still likely that different agents will succeed to different degrees, and that an agent might succeed to a greater or lesser extent in different situations.

Tännsjö seems to think that the deliberative approach is a (relatively) practicable method of decision-making:

> However, there is nothing in this goal that should be in principle impossible to achieve […]. Therefore, to try to maximise expected happiness seems to be a viable goal. We have reached it when we have *done our best* in forming a certain kind of opinion and when we have achieved a certain kind of consistency between our various different assessments and our actions. We have reached it even if these judgements are wide of the mark. If we succeed in reaching it we perform an action that is subjectively right. (Tännsjö, 1998, p. 36. My italics.)

That we have 'done our best' in forming our opinions seem to suggest that Tännsjö has some version of the Reasonable subjectivist approach in mind. Not just any opinion would suffice. Tännsjö claims, furthermore, that we can have justified beliefs about whether or not an action maximises expected happiness:

> Now, in a particular case, whether an action maximises expected happiness or not does seem open to inspection. We may hold a justified belief about this. (Tännsjö, 1998, p. 24)

This means that we would be able to determine the practicability of this approach. I believe that this is extremely difficult, if not impossible, for human agents. To the extent that human AU-agents are able to live up to the standards of the reasonable subjectivist approach, i.e. *if* they were able to gain justified beliefs regarding the possible outcomes of their alternative actions in terms of the realised utility and *if* they were able to justify their assignments of probabilities and *if* they were able to provide a way of construing of the relevant alternatives in situations of choice, this approach constitutes a viable method of decision-making. But there are vast problems with this. Justifying a belief that doing one thing rather than another would probably make the universe contain a greater balance of positive over negative intrinsic value is not an easy thing. This makes the method highly impractical for human AU-agents. Furthermore, justifying a belief to the effect that this method is highly valid for a human AU-agent is probably beyond such an agent's abilities.

The deliberative approach is sometimes thought to constitute a valid approach to decision-making given AU:

> Doing this [trying to maximise expected utility] may result in our *doing the wrong thing*, but it is safe to assume that if our actions maximise expected utility, then we will *succeed in producing the most happiness over the long run*. (Shaw, 1999, p. 30. My italics.)

This is a bold conjecture about general strategies of decision-making. The problem is, of course, to give reasons to think that it is true:

> It might be thought that if we try consistently to maximise expected happiness, then, in the long run, probably, we end up with better results than by consistently applying any conceivable alternative strategy. This may be so, and if 'probably' is taken in a subjective sense, it is probably true (of many utilitarians, at any rate). However, there exists no conclusive argument to the effect that, by consistently trying to maximise expected happiness, we obtain good results. (Tännsjö, 1998, p. 24)

In order to test this conjecture we would have to be justified in our beliefs regarding very complicated empirical matters of facts. We would also have to provide an account of the relevant alternative-sets. Justifying these beliefs and providing an account of the relevant alternative-set is something that we haven't done and something that we are extremely unlikely in ever accomplishing.

Tännsjö presents an argument for trying to adhere to the deliberative approach. This is not, strictly speaking a reason for taking this approach to be valid, but I will comment on it here anyway:

> […][E]ven if, for all we know, by applying the method [trying to perform subjectively right actions] we may not produce better results than we would do if when taking hard decisions we were to flip a coin, at least we know that we concentrate on aspects that, according to utilitarianism, are of moral importance. And we know that we concentrate on no aspects other than these. This means that, by adopting the maximising method, we are at least doing our best. (Tännsjö, 1998, p. 25)

But this is not very plausible. First, we are not concentrating *only* on aspects that are relevant according to AU[37]. Probability is not a 'relevant aspect' according to AU, neither are our beliefs about different *possible* outcomes of our actions. It is only the *actual* outcome of our actions, as compared to what would have been the outcome, had we done something else, that is of relevance to the rightness or wrongness of our actions. The agent also needs to be justified in her belief that she is considering the relevant alternative-set.

[37] Cf. Bergström (1996) p. 91-92.

I argued in chapter II that an AU-agent, in order to be justified in trying to adhere to a particular method of decision-making, should have good reasons for believing that i) she would succeed in adhering to the method if she tries and ii) that adhering to the method would make her approximate the overall goal of AU. It seems that agents with normal cognitive capacities cannot provide good reasons for this belief. There is yet another complication:

> […I]t takes a rule-utilitarian argument to carry us from the premise that it pays in the long run to try to maximise expected happiness, if this premise could be established, to the conclusion that, in a particular case, we ought to do so. (Tännsjö, 1998, p. 24)

Even if the premise *could* be established, which it cannot, we would face this problem. For an AU-agent this is of course a serious problem. She faces a problem of demarcation. If trying to maximise expected utility is not the best approach in every situation, how can an AU-agent justify a belief to the effect that she either *is* or *is not* in a situation where trying to maximise expected utility *is* or *is not* the best thing to do? It appears that she often, if not always, is not justified in her beliefs to this effect.

9.0 Conclusions

Maximising expected utility has played an important role as a part of different methods of decision-making for utilitarian thinkers. In view of this, the lack of arguments for the principle of maximising expected utility, as a part of a method of decision for AU, is remarkable. As will become apparent in the next chapter, utilitarian thinkers have presented several arguments for *restricting* the scope of the deliberative approach. But arguments in *favour* of trying to adhere to the method itself are few. Tännsjö's argument is the exception to the rule. But the argument is not convincing. Tännsjö's appeal to 'important aspects' is untenable and seems irrelevant. And Shaw's "assumption" that the long-term consequences of maximising expected utility makes an agent approximate the overall goal of AU stand unsupported.

I have assumed that the deliberative approach is a mode of decision-making that an AU-agent can try to adhere to. This method was given different interpretations by applying different criteria of relevance to the input of the deliberation-process. I argued that, given a plausible criterion of relevance on the deliberation-input, which demands justification for the beliefs and assignments, it is hard to justify a belief that an agent has satisfied the methods criterion of successful deliberation, i.e. that trying to adhere to this method would make the agent succeed. We are not justified in believing that the reasonable deliberative approach would score high in terms of practicability for human AU-agents. This makes it hard for agents to justify trying to adhere to this method.

In this chapter, the object of investigation was the *unrestricted deliberative strategy*. This method was a 'straw-man'. At most, the deliberative approach is but one part of a method of decision-making for AU. No defender of utilitarianism has, to my knowledge, defended the deliberative approach to decision-making in this unrestricted form. In certain situations, trying to act subjectively right is not believed by these thinkers to be the best strategy. Situations of temporal confusion, wishful thinking, threat and so on, are perhaps examples of situations where trying to maximise expected utility will make agents fail to do just that. This approach seems particularly unsuitable in situations characterised by lack of time and wishful thinking. This "calculating" approach also seems inappropriate in situations concerning relations to our near and dear[38]. *If* restrictions of this method are needed, i.e. if there are situations in which the AU-agent ought not try to act subjectively right, because this would make them fail to do so, then this shows that this approach is wanting in terms of validity. The need to propose restrictions on when this approach ought to be used reveals, if justified, a flaw in the methods validity.

According to the *restricted deliberative strategy*, the agent ought in certain kinds of situations to follow secondary rules of conduct instead. The "retreat" from the unrestricted deliberative strategy raises the question of the status of these secondary principles. Can these principles be given a ground in AU? The distinction between criteria of subjective rightness and methods of decision-making enables a proponent of AU to say that in situations where the agent does not have the time to deliberate, her actions (e.g. the following of the rules of common sense morality) might still satisfy the criterion. Thereby even these actions could be given an indirect rationale grounded ultimately in AU. Of course *arguing* along this line is one thing, establishing the connection is quite another. These questions will be addressed in the next chapter.

[38] See for example Tännsjö (1998), chapter 3. This question is also discussed in Lif (2003). For a (partial) defence of consequentialism against this argument see Jackson (1991).

Chapter IV
Secondary Rules

1.0 Introduction

In this chapter, we shall turn to the role of secondary rules of conduct within different possible methods of decision-making for AU. Different utilitarian thinkers have proposed different suggestions as to how and to what extent a utilitarian agent should make use of secondary rules of conduct in her decision-making. However, the differences are marginal enough as to warrant speaking of a difference of degree rather than of kind. Common to the vast majority of utilitarian thinkers is that they take secondary rules to be an important part of their preferred methods of decision-making. They adduce similar reasons such as cost-effectiveness, timesaving, avoidance of bias and the like, for preferring these rules, at least in some situations, to the deliberative approach, i.e. trying to maximise expected utility in the way examined in the previous chapter. The methods of decision-making I discuss in this chapter are all versions of the restricted deliberative strategy. But while some, (e.g. Moore), give some of these rules a quite *absolutist* place in their methods, allowing for no exceptions, others (e.g. Tännsjö, Sidgwick and Hare), give these rules a less intrusive role.

I shall especially focus on the role of secondary rules in the methods presented by Tännsjö, Sidgwick, Moore and Hare, all of whom are representative writers in the utilitarian tradition. I think that they cover a representative spectrum of approaches on this subject. They also have more or less elaborate ideas on the role of secondary rules. More specifically, I shall examine their methods of decision-making in relation to AU, as it is spelled out in chapter II, regardless of the fact that, e.g. Moore's and Hare's ethical theories in particular, can in certain respects be taken to diverge from AU. I examine their different views on the role of secondary rules and relate their ideas to AU.

The relation between secondary rules, whether they are called, 'rules of thumb' (Smart), 'prima facie principles' (Hare), 'subordinate principles' (Mill) or 'middle axioms' (Sidgwick), or anything else, on the one hand and the rules of common sense on the other, is only contingent and coincidental. But utilitarians who emphasise the role of secondary rules tend to identify these rules with the rules of 'common sense'. There are several reasons for this. One is that this will help fending off attacks, launched by their critics, that AU sanctions repugnant, evil, cruel, and in other ways (what is conceived of as) morally abhorrent actions, in certain situations. Another reason might

be that the very fact that the common sense rules have a certain adherence- and acceptance-utility[1] is partly due to the fact that they are accepted. These reasons are not very convincing though. The first is question-begging relative to AU. The second does, even if valid, not prove much. A certain amount of adherence- and acceptance-utility is not enough. To be justified, relative to AU, in adhering to them, reasons must be adduced for believing them to have the *highest* adherence-utility possible. AU is a maximising theory. In order to establish that the common sense rules correspond to the secondary rules that AU would recommend, we need arguments based on AU. Below we will see what some of the utilitarian thinkers themselves have to say about this. Because the notion 'the rules of common sense' is notoriously vague, I will not try to decide on the relation between AU and specific interpretations of common sense rules of conduct though. I will limit my investigation to examine the general relation between secondary rules of conduct and AU. Can reasons, based on AU, be given that justifies specific secondary rules?

After having recapitulated what Tännsjö, Hare, Sidgwick and Moore have to say about the role of second order principles in relation to practical action-guidance, presenting my reconstruction of their respective method of decision-making, a common structure transpires. Although the *role* of the secondary rules is disputed and despite minor differences between their respective methods, there is a general agreement that secondary moral rules are important when an AU-agent is making her decisions.

I will focus on elements that are common to the different approaches. It should be kept in mind that the utilitarians I examine in this chapter have not really developed a determinate account of just what role secondary rules are to play, nor which specific secondary rules that ought to be acknowledged. What they do say is often indeterminate and sketchy. They do not present any systematic method of decision-making. This makes an evaluation of their suggestions difficult. Trying to interpret these ideas by giving them a precision that they do not have, runs the risk of attributing to the writers views that they do not endorse. While some ideas are quite straightforward, making this problem rather small, other ideas are more problematic. I will do my best to avoid, as far as possible, making any serious mistakes with regard to this issue. The reader should keep in mind that I am trying to make a reconstruction of what I take their favoured methods of decision-making to be. My purpose here is not to criticise their actual views on this subject, but to examine different possible candidates for methods of decision-making for AU. However, the examination should be made, if not for other reasons, since it illustrates the lack of concrete action-guidance which these suggestions provide to moral agents.

[1] I will not try to give any exact definition of adherence-utility and acceptance-utility. It suffices to say that the adherence-utility of a set of rules is the utility realized through successful adherence to the set. The acceptance-utility of a set of rules is the utility realized through agents accepting the set.

2.0 The problem of demarcation

In order to evaluate a method of moral decision-making that constitutes a blend or combination of different "pure" approaches, e.g. the deliberative approach and the secondary rule-following approach, we need to know the role and relative importance of each of the "pure" approaches within the method. Should the AU-agent use the deliberative approach often or more seldom? Does the method prescribe following secondary rules most of the time or only on certain rare occasions? Drawing the line, specifying when one approach is to be used and not the other, I call *the problem of demarcation.* It is an important problem and a problem that has not yet received an appropriate solution by the utilitarian thinkers that I will be discussing in this chapter.

In order to illustrate this problem, I will consider what J. J. C. Smart has said in this context, as well as the criticism that his suggestions have evoked. The deliberative approach plays a prominent role in Smart's method of decision-making. Although no utilitarian has defended the idea as the sole method of moral decision-making, i.e. as the method to be used in *every*[2] decision an agent makes (what I have called the unrestricted deliberative strategy), Smart is, to my knowledge, the one who comes closest to such a position. As I have already pointed out, Smart defends maximising expected utility as a *criterion* of 'rational' action. But he also defends the deliberative approach of trying to maximise expected utility through a rather "direct" calculation:

> [The act-utilitarian] criterion is […] *applied* in cases in which he does not act habitually but in which he deliberates and chooses what to do. (Smart, 1973, p. 45)

And further:

> When […the agent…] has to think what to do, then there is a question of *deliberation* or choice, and it is precisely for such situations that the utilitarian *criterion* is intended[3]. (Ibid. p. 43. My italics.)

---

[2] Strictly speaking not every decision can be reached through deliberation. This would give rise to the regress mentioned in chapter II. If this approach is used in determining which situations it should be applied to, if it is not applied to a specific decision-matrix, but is also used in setting up the matrix , then it seems that we are back in the regress.

[3] One thing should be noted in this context. Smart does not conceive of adhering to rules of thumb as a *part* of a method of decision: "[…A]ct-utilitarianism is meant to give a method of deciding what to do in those cases in which we do indeed decide what to do. On these occasions when we do not act as a result of deliberation and choice, that is, when we act spontaneously, no method of decision, whether utilitarian or non-utilitarian, comes into the matter." (Ibid. p. 44) This passage suggests that adhering to rules of thumb, i.e. secondary rules of conduct, is not part of a method of decision at all. They seem instead to govern "spontaneous" actions. Every time the AU-agent actively *decides* to do something, when she makes a *choice*, she should use the deliberative strategy of trying to maximise expected utility. We can choose to habituate ourselves to abide by rules of thumb but we cannot in a particular situation of choice, choose to adhere to such a rule. 'Choices' and 'decisions' ought always to be taken by using the deliberative approach of

Trying to meet this criterion consists in trying to adhere to the deliberative approach. How do the secondary rules enter the picture? This is the topic of the chapter in Smart (1973), entitled "The place of rules in act-utilitarianism".

According to Smart, agents ought sometimes not to try to maximise expected utility. They ought instead to adhere to 'rules of thumb'. What is a rule of thumb? Smart does not go deeply into that question. He says that they are "stereotype" rules that AU-agents are to follow "habitually"[4]. One example is the rule of keeping promises. Another is to try to save drowning persons (in situations where this is possible).

When ought the agent deliberate and when ought she to follow 'rules of thumb'? A plausible answer would be to claim that the AU-agent should guide her actions by using rules of thumb when this makes her *approximate the overall goal of AU*. The *motive* of the agent is to approximate the overall goal of AU, and the agent believes that the best means to accomplish this end is to adhere to a particular 'rule of thumb' in the particular situation. The 'rule of thumb' is a mere means for the agent trying to realize her goal. But this does not help us solve the problem of demarcation. This is because we cannot identify which situations that satisfy this characterisation. Smart suggests another point of demarcation:

> Normally, [the agent] will act in accordance with [rules of thumb] when he has no time for considering probable consequences or when the advantages of such a consideration of consequences are *likely* to be outweighed by the disadvantages of the waste of time involved. (Ibid. p. 42. My italics.)

This passage raises many questions. To be justified in believing that the advantages of considering probable consequences are *likely* to outweigh the disadvantages of this consideration, it seems that the AU-agent must already involve herself in the kind of deliberation that the 'rules of thumb' were designed to spare her from. The 'has no time'-clause is, therefore, dubious. Richard T. Garner and Bernard Rosen (1967) have argued that this leads to serious problems:

> Smart suggests that one ought not to consider the probable consequences when the advantages of doing so are likely to be outweighed by the "disadvantages of the waste of time involved," but this suggestion is totally unacceptable, for it implies the impossible–that sometimes one can decide on act utilitarian grounds whether or not to calculate the consequences of a particular action. For example, let A be the action of deciding whether or not to take a trip after considering the probable

---

maximising expected utility. My use of the notion of a 'method of decision' is, however, more inclusive. Adhering to secondary rules can be a part of a method of decision-making.

[4] Cf. Smart (1973), p.43.

> consequences of doing so, and let B be the action of deciding without considering the probable consequences. Smart seems to be suggesting that one can decide whether to do A or B on act utilitarian grounds. But no one could do this without knowing the relative value of the consequences likely to follow such actions. And, in order to know this, he would need to know what he would do if he had calculated and what he would do if he had not. In order to know what he would do if he had calculated, he obviously needs to know the consequences of taking the trip–something he can learn only by engaging in the calculation he is trying to decide whether or not to avoid. Of course it might be possible to suggest that there are good act utilitarian reasons for not calculating when to calculate, but once again one could only provide an act utilitarian justification for not calculating when to calculate whether to do A or B by calculating, and again, one thing he has to know in order to perform this calculation is the probable consequences of doing both. Thus, it seems that Smart has utterly failed to show how the act utilitarian can justify, on act utilitarian grounds, the use of rules of thumb. (Garner, 1967, p. 68-69)

Of course, taking the distinction between criterion of rightness and method of decision seriously, we can escape the 'implication of the impossible'. But, and this is important, it only takes us so far. It solves the theoretical problem, but not the practical.

Smart's view seems to be that the AU-agent should, through the deliberative approach, determine which rules of thumb she ought to acquire the disposition to act on and also in which situations this disposition should "kick in". We would need secondary rules that are not questioned on every particular occasion. The AU-agent ought not to try to determine, case by case, if there is a lack of time or not. The deliberative approach cannot be taken to be the default-mode of moral decision-making because the secondary rules could not then play the role they are supposed to play, e.g. saving time etc. It is rather that there are situations in which the agent should simply act in certain ways, without any deliberation. An example is where someone is about to drown and the AU-agent could save that person. In this situation she ought, without further ado jump into (or throw a life-buoy into) the water[5]. According to this way of looking at rules of thumb their justification comes from the deliberation leading to their formulation and this deliberation should of course also be justified. Both of these justifications are problematic, however.

Smart writes that the criterion is intended for situations where the agent can deliberate. But the *criterion* is more plausibly regarded as intended to cover *all* kinds of situations. In situations where the agent deliberates, where 'he has to think what to do', Smart says, that the agent ought to use the deliberative

[5] Given, of course, that this kind of disposition is part of the best method of decision-making.

approach to try to *satisfy* the criterion. In these situations the agent ought to try to maximise expected utility as a method of decision.

Accordingly, the AU-agent should adhere to rules of thumb in situations where she has reason for believing that this is the kind of situation where acting on them would *maximise expected utility*. Suppose an agent reaches a decision by simply adhering to a secondary rule. Suppose further that she decides to do A. Of course, it is possible that A maximises expected utility. That is, it is possible that *if* the agent had deliberated, A is the practical conclusion she would have reached. But, as a matter of fact, she used another method. In order to retain the rationale behind maximising expected utility as a way of approximating the criterion of right action of AU, the restricted deliberative strategy, i.e. the method which involves sometimes following rules without deliberating, one would have to argue that excluding some situations from deliberation would help the agent to meet the criterion of maximising expected utility, without adhering to the method, i.e. without deliberating. Because maximising expected utility can be conceived of as a criterion of rationality perhaps the rationale for adhering to secondary rules could be cast in terms of satisfying the criterion without using the deliberative approach. If this could be accomplished, the *restricted deliberative strategy* would be much more respectable. But justifying that belief is also very hard, as the forgoing discussion has indicated.

The problem of demarcation will surface on many occasions throughout this chapter. It is time to turn to an overview of different proposals for methods of decision-making for AU.

3.0 Tännsjö's 'List'

Tännsjö defends several exceptions to the deliberative approach, which in Tännsjö's method means trying to act subjectively right, i.e. trying to satisfy the SUR criterion. He presents a list of situations where the AU-agent ought not to try to act subjectively right. I will comment on each exception in turn. Tännsjö's proposal is that we adopt a list of different methods, which each seems to fit in different situations of choice. This list is only vaguely specified[6]. It contains elements such as 'sometimes following common-sense moral principles', 'sometimes acting in a spontaneous, un-reflected, manner', 'sometimes ignoring threats', 'sometimes acting subjectively right'. This approach can plausibly be regarded as a version of the restricted deliberative strategy, where the deliberative approach is, on some occasions, restricted by an appeal to secondary rules. In these situations, instead of deliberating, the agent ought to try to adhere to certain secondary rules of conduct.

[6] This suggested method of decision is not a supposed to be complete in any way. It is a rather sketchy suggestion. I will, however, examine the method and comment on it anyway, perhaps sometimes making more substantial interpretations of the ideas than could be supported by the text.

It should be kept in mind that the very point of this kind of list is to guide actions. This imposes certain demands on it. In particular, the list should be specific enough to give definite[7] answers to what we ought to do. It must also be 'manageable', in the sense that it must not be too complicated. It should be noted that the list is not in any way regarded as exhaustive, or complete. Despite this, the approach raises many interesting questions. Let us consider Tännsjö's proposed method of moral decision-making in some detail.

We have already discussed one part of this list, i.e. trying to perform subjectively right actions. This is the most important part of the method. I think that the other items on the list are best seen as amendments to this main part[8]:

> […W]e want to say that, in general, it is responsible for a person to perform a subjectively right action. (Tännsjö, 1998, p.34)

When should AU-agents refrain from trying to perform subjectively right actions and use some other 'method' on Tännsjö's list? Tännsjö gives examples of situations where the deliberative approach is not to be pursued, I will recapitulate them one by one.

3.1 Lack of time

One kind of situation where trying to act subjectively right might seem to be a bad idea is when the agent does not seem to have the time (relative to a specific alternative-set) required to carry through the calculations needed to determine which action maximises expected utility. This kind of situations includes instances where an action needs to be performed immediately (or at least very quickly). They also include situations in which the agent exercises the performance of specific skills or habits, e.g. car driving:

> In order to be a good driver, it is essential in many decisions not to deliberate. Our reactions should rather be conditioned responses to certain typical stimuli. If we attempt consciously in these situations to do our best, we will perform poorly as drivers. As a matter of fact, what distinguishes a good driver from a bad one is that the good driver is capable of not deliberating in these situations. This is not to say that, when we decide whether we should drive at all, or whether we should drive today, we should not try to maximise expected happiness. It may very well be the case that we should. (Tännsjö, 1998, p. 37)

What does this tell us about acting morally reasonably or responsible? Of course, the argument is convincing given the goal set, i.e. performing good *as*

[7] How definite should we demand the prescriptions of the method to be? At present moment, I do not have any suggestion.

[8] Ideally, the amendments should only take effect in situations where *trying* to act subjectively rightly would, in fact, be counterproductive.

a driver. But what about performing good *as a moral agent*, i.e. approximating the goal set by AU? There is always the alternative of stopping the car and doing something else in stead. Furthermore, when driving one has the ability to reason in many ways at the same time. One ponders over questions about what to do when one arrives at one's destination or one chats with the passenger in the seat next to oneself. At the same time one is driving the car. But decisions reached (on issues that do not concern one's car-driving) whilst driving the car should perhaps be reached through deliberation.

If Tännsjö's argument is accepted in this kind of situation, similar arguments ought to be accepted regarding other circumstances. Consider the case of a physician deciding whether to provide euthanasia to a patient with considerable suffering and bleak, but not negligible, chances of recovering. Suppose that trying to maximise expected utility in this situation suggests that the physician should to provide euthanasia for the patient. But here one *could* argue, in a parallel manner, that performing well *as a physician* requires the physician to abide by the laws and norms of the society he works in and that giving the patient euthanasia is therefore not what she ought to do. It is hard to see why the argument is valid in the one case, but not in the other.

The point is that establishing that when driving one ought to let simple stimuli-response mechanisms govern ones behaviour is not of much help in deciding what to do. When driving, the alternatives of stopping or not stopping the ride are open to the agent. This decision, it could be argued, is not suitable to make relying on only stimuli-response mechanisms. And as long as these options are viable, perhaps the agent is not justified in sticking to her stimuli-response pattern of behaviour[9].

When is an agent in a situation with lack of time? We could not answer "In situations where the agent would fail to maximise expected utility if she tried to do so!" This would beg the question relative to SUR. These situations need to be characterised independently, without reference to the SUR criterion, to be useful here. The rationale for abiding by these amendments is, presumably, that doing so is the best possible way for an agent to approximate the overall goal of AU. But we need a method of decision for deciding when to try to satisfy the SUR criterion by trying to maximise expected utility and when to go for some of the amendment-clauses (determining whether this is a situation in which the amendments are applicable). We seem to need independent and substantial characteristics of the situation indicating that there is a lack of time. But what would they be?

One suggestion would be some kind of subjective criterion, perhaps cast in terms of subjectively felt, or experienced, time-shortage. There are also possible criteria of a more objective kind. These could be cast in terms of

[9] Of course, if the agent is justified in believing that AU supports her driving the car, *then* driving the car in the usual, unreflective way, is presumably what the agent ought to do.

absolute time "shortage" relative to some specific alternative-set. There are problems with both of these approaches. The first suggestion would suffer from problems akin to those of the Pure subjectivist approach, discussed in the previous chapter. The felt time-shortage need not be a case of 'actual' time-shortage, i.e. the agent might be mistaken. The second suggestion would have to give an account of how to determine the relevant alternative set, a problem that is still unsolved. I mention these possibilities just to point to these problems. I will not take a stand on this issue.

3.2 Likely bias or wishful thinking

There are situations where we, through bias or wishful thinking, are likely to deliberate poorly, i.e. where the agent will make faulty assessments of probability or of the facts of the situation[10]:

> I am thinking of situations where bias or wishful thinking is likely to lead us astray, such as the one where we have drunk (moderately) and contemplate whether we should drive or leave the car. If we try to maximise expected happiness it is very probable, due to wishful thinking, that we will end up with the (unwarranted) conclusion that we ought to drive. Therefore, in the situation, we should not think in terms of expected value at all. Rather we should stick to established rules. Drinking and driving do not go together, period. (Tännsjö, 1998, p. 37)

This is of course speculative. It seems as though the categorical prescription that one ought always, when slightly intoxicated, to *believe* that one is never to drink and drive, is inferred as a conclusion from AU. Drinking and driving might, sometimes, be the right thing to do, but *thinking* that it is, is always wrong in these circumstances, according to Tännsjö. Is there evidence for such a claim that together with AU supports the conclusion? How intoxicated should the agent be for the rule to kick in?

The example might be questioned. Perhaps even a sober deliberation would suggest that one ought to drive drunk. The probability of hurting oneself or other people might be low enough as to make the enterprise worthwhile. What are the possibilities of identifying these situations by the agent? Even if we should accept the conclusion that we always should think that drinking and driving do not go together, period, in situations where we have been drinking moderately, what does this tell us? It does not help much, in terms of action-guidance, to exclude drinking and driving as an option. Ought the intoxicated agent to walk home or pick up a cab or go by the tram or subway or what? And how is the agent to determine this? Through trying to adhere to the deliberative approach? By following established secondary rules of conduct? What do they say about cases like this one?

[10] Cf. Hare (1981), p. 38.

The bias amendment raises another question of a different kind. The agents for which the problem of bias and wishful thinking arises are, precisely because they are wanting in certain respects, particularly unlikely to be able to identify situations where the bias is present. Because of the fact that normal agents occupy this position, it is difficult to give principled answers to questions about wishful thinking, i.e. when does the agent have good reasons for believing that she is in a situation where she is thinking wishfully or exhibit other biases? The more biased and the more wishful an agent is, the less likely it is that she will acknowledge that she is.

3.3 Intimacy

Another kind of situation in which trying to act subjectively right has been taken to be a bad strategy is when personal relations and commitments are threatened[11]:

> […T]here are situations of intimacy, where a calculating approach would ruin important values. When in love, and ones a firm relationship is established, it is not appropriate to contemplate whether a change of the object of one's love and affection would be a change for the better. A prudent person who believes in AU, then, and who is prepared in many situations to aim at subjectively right actions, will not do so when conducting his or her most intimate life. Instead, such a person will have conditioned himself or herself to be a person who acts in his or her most intimate life out of spontaneous affection. However, there is an appropriate limit to this. There may come a time when divorce is a viable option. A prudent person does not condition himself or herself, then, never to give up this thoughtless approach. When life has become hell, it may be time to aim once again for subjective rightness. (Tännsjö, 1998, p. 37)

It is one thing to claim that calculating whether it would be better to replace, e.g. ones wife, with another person is often a bad idea. Most people have firm intuitions supporting such a claim. But what does AU say? The biggest problem seems to be the problem of demarcation. Which situations concern one's most intimate life? The problem is, of course, to give a definition of 'situations that does not concern personal commitments' that does not beg any questions relative to AU. Let me give an example. Suppose that I am wondering whether to donate half of my salary to some foreign-aid organisation. Is this a situation that concerns my 'most intimate life'? People near and dear to me will obviously be affected (I will have less disposable money to spend on them). Perhaps they will wonder (with or without good reason) whether I care more about distant people than I care about them etc. But if this is a situation that concerns my most intimate life, where I ought not

[11] Cf. Parfit's discussion in (1984) of 'blameless wrongdoing'.

to try to maximise expected utility, which situations are left? When should I try to act subjectively right? There are difficulties of demarcation here that need to be solved in order to make the method determinate.

3.4 Threat or blackmail

Tännsjö also wants to make an exception to the deliberative approach in situations of threat:

> […I]n situations of threat or blackmail it might be a good idea to be a person who is prepared irrationally to refuse to give in, since being such a person means that one is not easily exposed to threats. (Tännsjö, 1998, p. 37-38)

No doubt, refusal to give in to threats will *sometimes* be the best thing to do. But how do we ascertain that we deal with a rational robber and not a psychotic one? Determining this will often be very hard. We would also have to make sure that this disposition is known to would be blackmailers, but perhaps without anybody else knowing about it. (If other people knew of this disposition, this could have unfavourable effects. Persons close to the agent may feel uncomfortable if they knew that she would not, under any circumstances, yield to threats.) Furthermore, it seems to me that this refusal is primarily suitable for states or other big organisations, rather than to individual agents. States can "afford" to make the sacrifices necessary to make their strategy credible. For individual agents, this is harder.

3.5 Justifying the list

Is it possible to give reasons, in terms of AU, for an AU-agent to try to adhere to Tännsjö's list, i.e. his favoured version of the restricted deliberative approach?

> When those who believe in the truth of AU[12] and SUR consistently adhere to the methods on the list, they do so in the hope that, by doing so, they are, in the long run, producing better outcomes than they would have done had they consistently held on to any alternative method that they can think of. This is what they hope. But can they give any reasons why their favoured list is superior to any putative competing one? I think that such reasons can be given, but they are far from conclusive. In the first place, and negatively, we can try to ascertain that our method of decision making (our list) does not violate Gilbert Harman's 'rational equilibrium test', i.e., the list is such that following the procedures on it would not lead one to modify it. Secondly, and more positively, we can try to show, in a piecemeal manner, that what is on the list deserves to be there. We can show in relation to each item that, when we have followed the kind of decision procedure described, the results, as far as we have

[12] Remember that Tännsjö uses 'AU' to refer to his hedonistic version of act-utilitarianism.

> been able to assess them, have been better than when we have followed a different procedure. The reasons we put forward, then, are of a general inductive nature. (Tännsjö, 1998, p. 38)

How do we 'ascertain' that a particular list does not violate the 'rational equilibrium test'? Perhaps inductive support can be taken to speak in favour of the method satisfying Harman's test. But how do we determine this? Perhaps trying to adhere to the method will lead us to revise it in the future. The empirical evidence needed to justify this hypothesis is not easily acquired.

The second reason appeals in a straightforward manner to possible empirical evidence. How are we supposed to show that what is on our list deserves to be there? What would the comparisons referred to in the citation be like? I believe that the comparisons needed in this context are as difficult to make as the comparisons needed to apply the criterion of rightness directly in a situation of choice. The reasons for believing that an item on the list deserves to be there are weak in proportion to the uncertainty of the empirical evidence. As a matter of fact, I think that they would be *very* weak.

This list of situations, where trying to perform subjectively right actions is not recommended, is not thought to be complete by Tännsjö. He acknowledges that there might be other situations not accounted for by him where this approach should be avoided. This means that Tännsjö does not present a complete method of decision. Furthermore this means that the problem of demarcation is not solved by this method.

4.0 Sidgwick: Restricted Empirical Hedonism

While emphasising the importance of general adherence to common sense rules by human agents, Sidgwick gave secondary rules a rather secluded role within the method of decision that he thought was appropriate. At first glance it might seem as if Sidgwick gave secondary rules a very prominent place. Although trying not to violate the rules of common sense is something of a default mode in his method of decision-making, a closer examination reveals that the exceptions to this general prescription are many. An agent abiding by Sidgwick's suggested way of making decisions is likely to have to perform a great deal of deliberation, i.e. by using the method that Sidgwick referred to as 'Empirical hedonism'. Empirical hedonism is Sidgwick's version of the deliberative approach.

How ought a utilitarian agent to decide what to do, according to Sidgwick? His answer to this question is rather complex. He writes:

> We must conclude, then, that we cannot take the moral rules of Common Sense as expressing the *consensus* of competent judges, up to the present time, as to the kind of conduct which is likely to produce the greatest

> amount of happiness on the whole. It would rather seem that it is the unavoidable duty of a systematic Utilitarianism to make a thorough revision of these rules, in order to ascertain how far the causes previously enumerated (and perhaps others) have actually operated to produce a divergence between Common Sense and a perfect Utilitarian code of morality. (Sidgwick, 1874, p. 467)

Sidgwick argues at great length for the 'general utilitarian basis of the morality of common sense'[13]. What role should common sense rules play in a utilitarian method of decision-making? I will try, as far as possible, to state what I take to be Sidgwick's proposal in as systematic a form as I can.

Sidgwick's method of decision-making can be stated thus:

1. Adhere to the prescriptions of the set of rules endorsed by common sense[14], unless
a) Common sense gives conflicting prescriptions in a particular situation of choice[15].
b) Common sense does not give any determinate prescription, or any prescription at all, in the particular situation of choice[16].
c) Common sense prescribes the following of a rule which it would, according to the belief of the agent, be sub-optimal to follow in the particular situation of choice[17].

2. If a, b or c applies, use empirical hedonism[18].

---

[13] Cf. Sidgwick (1874), primarily p. 423-495. Sidgwick's arguments do not pretend to show that the common sense rules of conduct are effective means to *maximising* utility. He only claims that they can be given a 'general utilitarian basis,' i.e. that they have a certain amount of positive acceptance- and adherence utility.

[14] Speaking about how a 'scientific Utilitarian' ought to act, Sidgwick claims that "Generally speaking, he will clearly conform to [the Positive Morality of his age and country], and endeavour to promote its development in others." (Ibid. p. 475)

[15] Cf. Ibid. p. 425-426. Also: "[…]that the apparent first principles of Common Sense may be accepted as the 'middle axioms' of Utilitarian method; direct reference being made to utilitarian considerations, in order to settle points upon which the verdict of Common Sense is found to be obscure and conflicting." (Ibid. p. 461)

[16] Ibid.

[17] "For a Utilitarian must hold that it is always wrong for a man knowingly to do anything other than what he believes to be most conducive to Universal Happiness." (Ibid. p. 492)

[18] Two different possibilities are open here. The agent might want to determine what she ought to do in the situation. This is what I discuss in this essay. Sidgwick poses the question of "what method of determining right conduct the acceptance of Utilitarianism will practically lead" (Ibid. p. 460) "The most obvious method, of course, is that of Empirical Hedonism, […]; according to which we have in each case to compare all the pleasures and pains that can be foreseen as probable results of the different alternatives of conduct presented to us, and adopt the alternative which seems likely to lead to the greatest happiness on the whole." (Ibid. p. 460) But the agent might also be considering the general status of the common sense rules, i.e. she might be trying to change the set of rules that is commonly adhered to in her society. The method by which a utilitarian should try to reform Common Sense morality is according to Sidgwick that of 'Empirical Hedonism': "Here our investigation seems, after all, to leave Empirical Hedonism as the only method ordinarily applicable for the ultimate decision of such problems–at least until the science of Sociology shall have been really constructed." (Ibid. p. 476)

According to Sidgwick, then, the default mode of utilitarian decision-making is to adhere to the prescription of common sense rules of behaviour.

> I hold that the utilitarian, in the existing state of our knowledge, cannot possibly construct a morality *de novo* either for man as he is (abstracting his morality), or for man as he ought to be and will be. He must start, speaking broadly, with the existing social order, and the existing morality as a part of that order: and in deciding the question whether any divergence from this code is to be recommended, must consider chiefly the immediate consequences of such divergence, upon a society in which such a code is conceived generally to subsist. (Sidgwick, 1874, p. 473-474)

But there are areas where common sense is silent as to what an agent should do. Take, for example, the question of how much time, energy and money that an agent should spend on charitable actions. Common sense does not provide an answer to this question. Is it better to work as a doctor for the MSF, than to work as a librarian?

After having questioned the possibility of drawing a line between private and public spheres of individual responsibility, Sidgwick made the following claim:

> But further: even supposing that we could mark off the "sphere of individual option and self-guidance" by some simple and sweeping formula, still within this sphere the individual, if he wishes to guide himself reasonably on utilitarian principles, must take some account of all important effects of his actions on the happiness of others; and if he does this methodically, he must, I conceive, use the empirical method which we have examined in Book ii [i.e. empirical hedonism]. (Ibid. p. 478)

It seems as though an agent trying to adhere to Sidgwick's method of decision-making will have to guide a substantial part of her decisions by using 'empirical hedonism', i.e. the deliberative approach.

> A hedonistic method, indeed, that would dispense altogether with direct estimates of the pleasurable and painful consequences of actions is almost as inconceivable as a method of astronomy that would dispense with observations of the stars. (Ibid. p. 178)

Does Sidgwick's method solve the problem of demarcation? Well, there seem to be situations where the agent will be in doubt whether adhering to a certain common sense rule of conduct would, or would not, be 'most conducive to Universal Happiness'. In these cases, which can easily be imagine to be rather numerous, the method will not give any definite answers.

5.0 Hare: Levels of moral thinking

In this section the method of decision-making defended by R. M. Hare, his two-level approach to moral thinking, is examined. One important preliminary note is warranted. Hare does not speak explicitly in terms of a substantial criterion of rightness of actions and a corresponding method of decision. Rather, Hare's version of act-utilitarianism is a consequence of the alleged logical properties of the moral concepts (such as 'ought' and 'must'). He says: "[...] the formal, logical properties of the moral words, [...] yield a system of moral reasoning whose conclusions have a content identical with that of a certain kind of utilitarianism"[19]. However, it is not difficult to recast the discussion in terms of a substantial criterion of rightness. I will disregard this part of Hare's theory, and examine rather the use an AU-agent could make of the two-level approach.

Hare, in *Moral Thinking*, advocates a two level ethical theory[20]. The two levels are called the *critical* and *intuitive* level of moral reasoning. Theoretically, the critical level is the more important one of the two. But according to Hare we pursue, and indeed should pursue, most of our moral thinking at the intuitive level. From a practical perspective the intuitive level becomes at least as important as the critical level because of this fact. One of the purposes of making this distinction between the two levels is, according to Hare, that it can be used "[…]in order to defend a version of utilitarianism against an extremely common type of objection which would not be made by anybody who understood the distinction."[21] I am interested in examining if the two level approach can play a role in making utilitarianism applicable in practice.

Before I turn to a discussion of the two levels, it is appropriate to introduce two *personae* that Hare makes use of in his discussion. The first is the *archangel*. The archangel is characterised in the following way:

> […A] being with super human powers of thought, super human knowledge and no human weaknesses. […] When presented with a novel situation, he will be able at once to scan all its properties, including the consequences of alternative actions, and frame a universal principle (perhaps a highly specific one) which he can accept for action in that situation, no matter what role he himself were to occupy in it. Lacking,

---

[19] Hare, R. M. (1981), p. 4. At the *critical* level, we are forbidden to take moral principles of substance into consideration. AU's criterion of rightness is a substantial moral principle.

[20] Hare is actually including three levels in his approach; the third being the meta-ethical level, where the meaning of evaluative words are discussed. My interest in this thesis is restricted to the other two levels exclusively.

[21] Hare (1981) p. 25. Exactly which objection Hare has in mind here is not quite clear. I think that it is the objection that utilitarianism has counterintuitive implications in some cases. But Hare also seems to think that the two-level approach can be used in defending utilitarianism from arguments of impracticability. (Cf. Ibid. p. 35-36, 38) The two-level approach seems, at least partly, to play a very similar role as the distinction between criterion of rightness and method of decision.

> among other human weaknesses, that of partiality to self, he will act on that principle, if it bids him act. (Hare, 1981, p. 44)

This being is an ideal AU-agent. At the other extreme, we have the *prole*:

> […A] person who has these human weaknesses to an extreme degree. Not only does he, like most of us, have to rely on intuitions and sound prima facie principles and good dispositions for most of the time; he is totally incapable of critical thinking (let alone safe or sound critical thinking) even when there is leisure for it. Such a person, if he is to have the prima facie principles he needs, will have to get them from other people by education or imitation. (Ibid. p. 45)

Much of what Hare says concerning the question of practical action-guidance turn on the question of how much human agents resemble *archangels* or *proles*. This question raises the problem of demarcation, to which I will return below. Having characterised these personae, it is time to turn to Hare's method.

5.1 Hare on 'Intuitive moral thinking'

What is the purpose of thinking at the intuitive level of moral thinking? According to Hare:

> Intuitive thinking has the function of yielding a working approximation to this [what the archangels would prescribe] for those of us who cannot think like archangels on a particular occasion. If we wish to ensure the greatest possible conformity to what an archangel would pronounce, we have to try to implant in ourselves and in others whom we influence a set of dispositions, motivations, intuitions, prima facie principles (call them what we will) which will have this effect. (Ibid. p. 46-47)

At this level the agent is operating with relatively simple *prima facie* principles. Due to this fact they will be able to give us only rough guidelines as to how we ought to act. But this is a good, or even necessary, feature of these principles. This is because they would need to be unspecific enough to be applicable in different situations and at the same time, help agents guide their actions in practical situations of choice:

> […]there is also a practical reason related to the circumstances of their use. Situations in which we find ourselves are not going to be minutely similar to one another. A principle which is going to be useful as a practical guide will have to be unspecific enough to cover a variety of situations all of which have certain salient features in common. (Ibid. p. 35-36)

Which set of principles should we try to adhere to?

> The best set is that whose acceptance yields actions, dispositions, etc. most nearly approximating to those which would be chosen if we were able to use critical thinking all the time. This answer can be given in terms of acceptance utility, if one is a utilitarian […]. (Ibid. p. 50)

What is a *prima facie* principle? Hare:

> Such principles express 'prima facie duties' […], and, although formally speaking they are just universal prescriptions, are associated, owing to our upbringing, with very firm and deep dispositions and feelings. Any attempt to drive a wedge between the principles and the feelings will falsify the facts about our intuitive thinking. *Having* the principles, in the usual sense of the word, is having the disposition to experience the feelings, though it is not, […] incompatible with submitting the principles to critical thought when that is appropriate and safe. (Ibid. p. 38-39)

Hare refers to Ross[22] when characterising the principles. The principles express *prima facie* duties. It is far from clear what kind of duties these are. Ross also speaks of them as 'conditional' duties. They are said to arise due to the fact that certain actions are of a certain kind. For example, the fact that an action is of the promise-keeping kind creates a *prima facie* duty to perform it. They are conditional in the sense that if an action is of a *prima facie* duty-creating kind and if it does not violate any other *prima facie*-duty, then this fact would make the action a duty proper. In the actual world, however, actions seldom are of just one *prima facie* duty creating kind. An action may be an instance of both a promise-breaking and a beneficent kind. Here two *prima facie* duties "pulls in opposite directions". At this point Ross and Hare diverges from each other. In this kind of conflict Ross thinks that we should act so as to fulfil the *prima facie* duty which is "more of a duty"[23]; to know which of the alternatives that is 'more of a duty' we have to make a practical 'judgement' of a rather intuitive kind. As a matter of fact, Ross explicitly rejects the idea that what would in this case determine our duty proper would be the fulfilment of the duty which maximises utility. And this is precisely what rationalises our choice of one rather than the other according to AU[24].

What is the function of *prima facie* principles? This is not an easy question. They are to "guide" our actions and we are to "follow" them. My interpretation is that they are to be taken as determining a class of actions which, on the intuitive level, it is obligatory or permissible for us to perform. To put it in a rather crude, and perhaps misleading, way: If you do not violate any of your *prima facie* principles (or duties) then what you are doing is

[22] Cf. Ross (1930).
[23] Ibid. p. 18.
[24] If Hare is right, then the archangelic prescription would amount to the equivalent choice.

morally permissible (and sometimes obligatory)[25]. An agent's set of *prima facie* duties are, in a manner of speaking, a *filter*. If the proposed action passes the grid of *prima facie*-duties, either in the sense that the action is *prescribed as a duty* or that it is *not forbidden*, then it is permissible for the agent to perform it. It would perhaps be illustrative to consider some examples of prima facie principles. Hare presents a situation where two prima facie principles conflict:

> I have promised to take my children for a picnic on the river at Oxford, and then a lifelong friend turns up from Australia and is in Oxford for the afternoon, and wants to be shown around the colleges with his wife. (Hare, 1981, p. 26-27)

According to Hare, this situation puts him in a position where two prima facie principles conflict. He ought, thinking now at the intuitive level of moral thinking, to keep his promise to his kids. All the same, still on the intuitive level, he ought to show his lifelong friend around the colleges[26]. Other examples of prima facie principles would be of the type "Always tell the truth", "Always return your debts" or "Never commit murder"[27]. According to Hare these kinds of principles are the ones that should guide us in our 'every day' moral thinking.

I think that this fits in quite well with the idea that (non-optimific) actions that do not violate any *prima facie* duties are to be considered permissible, and sometimes obligatory, at the intuitive level. If these actions where not morally permissible, then the very purpose of having these *prima facie* principles would collapse.

5.2 The Critical level

Hare, when discussing the prospects of putting his theory into practice, emphasises the fact that intuitive thinking alone is not enough. We need critical moral thinking as well:

> One thing, however, is certain: that we cannot all of us, all the time, behave like proles (as the intuitionists would have us do) if there is to be a system of prima facie principles at all. For the selection of prima facie principles, and for the resolution of conflicts between them, critical thinking is necessary. If we do not think that men can do it, we shall have

[25] This is of course another sense of 'permissible' and 'obligatory' than the one used on the critical level, or, in the language of AU, different from the sense in which only optimific actions are permissible according to the criterion of right action.

[26] As a curious remark one might wonder what the prima facie *duty* is that would have him show his friend the colleges would look like. Do we have a prima facie duty to satisfy the wants of lifelong friends (and their wives)? Is this putative duty so strong that it would even begin to compete with the keeping of promises (to one's children)?

[27] "Always wear seat belt when driving." is another example given by Hare. Some of the principles will no doubt be considerably more detailed and specific. But for now I will disregard this complication.

> to invoke a Butlerian God to do it for us, and reveal the results through our conscience. (Ibid. p. 45-46)

There are two major purposes of thinking at the critical level. The first is to find out which *prima facie* moral principles we ought, generally, to follow at the intuitive level, i.e. the search for what I have been calling *secondary rules of conduct*. The second is to resolve conflicts between *prima facie* principles at the intuitive level. According to Hare, it is the *prima facie* principles which, in the actual world, have the highest acceptance value that will be prescribed by the archangel, i.e. the ideal AU-agent. This ideal critical thinker possesses characteristics far from those which actual human beings possess. And this discrepancy has important consequences when it comes to the problem of guiding actions.

An archangel will, because of his superior characteristics, not need the intuitive level of moral thinking. This archangel can do the direct calculation needed. Because we are not archangels, we will not be able to perform perfect critical thinking. This is an *unattainable* ideal for human agents. When we, *qua* human agents, reason at Hare's critical level our reasoning can take different forms depending upon how we choose to handle our epistemic limitations. As we will see, for human agents, thinking at the critical level ought most plausibly be interpreted as adopting what I have been calling the deliberative approach, under the reasonable subjectivist interpretation.

5.2.1 Designing Principles

Human agents ought to design their methods of decision-making at the critical level of moral thinking. How should they decide which *prima facie* principles they ought to adhere to at the intuitive level? This question is supposed to be answered by thinking at the critical level of moral thinking using *the deliberative approach*. To be sure, human agents are not archangels. Human critical thinking is bound to be imperfect:

> To archangels, who can do it perfectly, I have ascribed superhuman powers and superhuman knowledge[…]. The most that human beings can ask for, when they are trying to do the best critical thinking they can, is some way of approximating, perhaps not at all fully, to the thought-process of an archangel. (Ibid. p. 122)

According to Hare, this approximation will take the form of trying to design the, would be, prima facie principles, so that their acceptance-utility is maximised:

> […]intuitive or prima facie principles have to be selected for their acceptance-utility in the actual world[…]. (Ibid. p. 113)

What we are aiming at is to implement the set of *prima facie* principles that has the greatest acceptance-utility[28]. For human agents, however, even this is to set our aims too high. We would probably not come to a decision if we set our aims this high. Human agent, thinking on the critical level, would have to rest content with aiming to design their set of prima facie principles so that they maximise *expected* acceptance utility.

> It is quite consistent even with act-utilitarianism […] to admit the utility of inculcating such principles, so that people do, without thinking, what is *likely* to be optimific. To do what they do may turn out, in extraordinary cases, not to be *right* by the act-utilitarian criterion; but even by that criterion what they do is *justified*, because the criterion, regarded as a guide to actual choices at the time that they are made, can only require people to do what maximizes *expectable* utility; and this, in nearly all cases, given the weakness of our human nature and the limitations of our knowledge, is likely to be what the sound general principles prescribe. (Hare, 1973, p. 16)

We are now back in the deliberative approach, except for the fact that instead of *actions* we are now dealing with *sets of prima facie principles*. The problems with this approach, however, are still there.

5.2.2 Resolving Conflicts

The resolving of conflicts at the intuitive level by critical thinking takes at least two different forms. First, we might qualify our *prima facie* principles, i.e. make them more specific or qualify them with a clause which tells us which other principles the principle should be overridden by and which principles should be overridden by it. The second way is to disregard the *prima facie* principles altogether and try, through using the deliberative approach, to maximise expected utility.

> 'The principles, since they are in conflict, cannot be altogether relied on; I am compelled to depart from one or the other, and do not know which. So let me put the principles aside for the time being and examine carefully the particular case to see what critical thinking would say about it.' This is possible for critical thinking, in so far as humans can do it, and it is what the […] crude act-utilitarians might recommend in all cases. It is, as we have seen, a dangerous procedure; but sometimes we may be driven to it. (Hare, 1981, p. 51)

This reasoning is the same as the method of trying to maximise expected utility, i.e. the deliberative approach. At this level the human agent is

[28] "Critical thinking aims to select the best set of prima facie principles for use in intuitive thinking. […] The best set is that whose acceptance yields actions, dispositions, etc. most nearly approximating to those which would be chosen if we were able to use critical thinking all the time." (Hare, 1981, p. 49-50.) Cf. also Ibid. p. 156 and 159.

supposed to estimate probable outcomes of different alternative actions and choose the one with the highest expected utility associated with its outcome.

According to Hare, this is not a method that should be used very often:

> Some critics of utilitarianism have supposed that the utilitarian is committed, whenever faced with a particular moral problem, to doing an elaborate calculation of utilities, involving interpersonal comparisons, in order to arrive at the optimum choice. I have adopted the view that such calculations are in practice usually impossible and that to undertake them would often be dangerous. We do better to stick to well tried and fairly general principles. (Hare, 1981, p. 121)

Hare thinks that we should do most of our thinking on the intuitive level, taking this mode of thinking to be the default mode of the method of decision-making. All the same, in order to design the prima facie principles, and resolving conflicts between them, we sometimes have to think at the critical level.

Another way of 'resolving' a conflict between *prima facie* principles does not make use of critical moral thinking on the spot. Instead, one *prima facie* principle may be adhered to, even if it is suspected that the principle ought to be qualified or altered. When there is not enough time to work out exactly *how* the principle ought to be altered, it may sometimes be adhered to in that particular situation and then be qualified, using critical thinking, at a later time, in a 'cool hour'. (Cf. Ibid. p. 51)

In other situations Hare seems to think that the 'conflict' can be resolved without recourse to critical thinking:

> In simpler cases we may 'feel sure' that some principle or some feature of a situation is *in that situation* more important than the others […]. We shall then be able to sort the matter out intuitively, letting one principle override the other in this case, without recourse to critical thinking. (Ibid. p. 50)

To 'feel sure' is, presumably, to make an immediate 'judgement'. According to Hare this is something quite distinct from weighing different outcomes in terms of value. I do not know why Hare makes this claim. If one 'lets one principle override the other' without recourse to critical thinking, the overriding principle must in a sense already have been qualified to override the other one in a situation of conflict. Or else, what would 'more important' mean in this context?

5.3 Hare on the justification of adhering to 'Prima facie principles'

Hare adduces *practical*, as well as *psychological*, reasons for adhering to these principles. One reason for following prima facie principles is that they are relatively practicable. The prima facie principles provides

> what is in fact an indispensable help in coping with the world […], namely the formation in ourselves of relatively simple reaction-patterns […] which prepare us to meet new contingencies resembling in their important features contingencies in which we have found ourselves in the past. (Ibid. p. 36)
>
> […I]n morals, the principles which we have to follow if we are to give ourselves the best chance of acting rightly are not definitive of 'the right act'; but if we wish to act rightly we shall do well, all the same, to follow them. A wise act-utilitarian, unlike his caricature mentioned earlier, will agree with this […]. (Ibid. p. 38)
>
> A further reason for relying in much of our moral conduct on relatively general principles is that, if we do not, we expose ourselves to constant temptation to special pleading […]. (Ibid. p. 38)

Although this would not be relevant for an archangel, for an AU-agent as defined in this essay, these points are, presumably, of relevance.

The general approach to decision-making that Hare seems to advocate could be schematically represented in the following way[29]:

1. Adhere to the *prima facie* rules of the intuitive level of moral thinking[30] except where
   a) They give contradictory prescriptions. (Cf. Ibid. p. 50-51)
   b) They do not apply to the situation. (Cf. Ibid. p. 40)
   c) Sufficient information is available that shows that the intuitively right act would not be for the best. (Cf. Ibid. p. 138)[31]
2. If a, b or c applies, then launch into critical moral thinking, i.e. use the deliberative approach. (Cf. Ibid. 133)

---

[29] William Frankena, represented the structure of Hare's method of decision thus: "In other words, the rational human universal prescriptivist has three ways of dealing with any particular situation on Hare's view: (1) to apply a [prima facie principle] if no conflict is involved and an adequate critically selected one is at hand, (2) to use [critical moral thinking] to find a satisfactory new [prima facie principle] to apply, or (3) to apply [critical moral thinking] directly to the particular case without bringing in any [prima facie principles]; and he or she may and even should use the third way on occasion. We now have a general picture of Hare's conception of human [moral thinking] as ideally including (a) no pure [intuitive moral thinking], (b) some pure one-level [critical moral thinking], (c) some two level [moral thinking] in which the [critical moral thinking] involved is impure in the sense of issuing in [prima facie principles], and (d) much impure [intuitive moral thinking], i.e. codal thinking that is critically based." (Frankena, 1988, p. 46)

[30] "*In so far* as the intuitions are desirable ones, they can be defended on utilitarian grounds by critical thinking […] if they can be so defended, the best bet, even for an act-utilitarian, will be to cultivate them and follow them in all normal cases […]." (Hare, 1981, p. 137)

[31] Hare is very sceptical regarding the possibility of ever entertaining this information. (Cf. Ibid. p. 138-140) It is not quite clear if Hare explicitly defends the c) clause. It is suggested in relation to a fictive debate between a defender of utilitarianism and a critic of the doctrine. However, I think that Hare *ought* to accept the clause. Rejecting it would go against the spirit of utilitarianism.

5.4 Alternating between the levels

How do we justify thinking at one, rather than the other, level of moral thinking in our decision-making? Again, this is the problem of demarcation. This is a very important practical question, which is not given enough room in Hare's writing. Before this question is answered, it is not clear what adhering to Hare's method of decision amounts to.

It is clear that Hare does not think that this question should not be decided anew on every occasion that the agent thinks he is facing a situation of moral choice. This would make the intuitive level collapse into the critical. This is because the agent would have to estimate the probable gains and losses of operating at the different levels and this would, *ex hypothesi*, be to think at the critical level. What is needed is some kind of principled approach, where the limits of each level are drawn. What Hare says about this is:

> Our question then is, 'When ought we to think like archangels and when like proles?' Once we have posed the question in this way, the answer is obvious: it depends on how much each one of us, on some particular occasion or in general, *resembles* one or the other of these two characters. There is no philosophical answer to the question; it depends on what powers of thought and character each one of us, for the time being, *thinks* he possesses. (Ibid. p. 45. My italics.)

In proportion to the degree to which we resemble the prole we ought, according to Hare, to refrain from attempting to think at the critical level and instead think on the intuitive level of moral thinking. Notice the shift from objective criterion to subjective criterion. Actual 'resemblance' versus 'believed possession of powers'. These two criteria can have very different implications. If actual resemblance is the criterion, then (probably) human agents should very seldom think at the critical level. If, on the other hand, it is the agents beliefs about which cognitive powers she possesses that should determine whether she should think at the critical level, then 'proles' with a high self-esteem should often think at the critical level.

To what degree could a human being be said to resemble an archangel? In terms of my terminology, to what degree can human beings live up to the strict demands of the Ideal deliberative approach? Even if we take her to be the most intelligent, wisest and good person on earth, I think that we would have to say that she resembles the prole more than the archangel. Human cognitive limitations are substantial. However, Hare has something more plausible to add:

> A person with any deep experience of such situations will have acquired some *methodological* prima facie principles which tell him when to launch into critical thinking and when not; they too would be justified by critical thinking in a cool hour. (Ibid. p. 52)

Sadly, Hare does not give any examples of such principles. Presumably, the justification for these principles would have to refer to some assumption that these methodological principles are part of the most valid set of *prima facie* principles. This suggestion is more satisfying from a theoretical perspective, but it is still not of much help from a practical perspective. We cannot justify a belief to the effect that we have found this set.

6.0 Moore: Some absolute prohibitions regarding calculation

In this section I shall examine G. E. Moore's proposal for how an agent ought to decide what to do in practical situations of choice, i.e. his method of moral decision-making. This problem is the topic of chapter V of Moore (1903), entitled "Ethics in Relation to Conduct". Moore defends a position on this topic that is conservative, with respect to common sense. Contrary to many other thinkers of a consequentialist bent, Moore stresses categorical adherence to some common sense rules of conduct. However, because there are many situations of choice in which common sense rules of conduct is silent, e.g. situations where several actions are permissible, the agent will have to guide her choice by "direct considerations", i.e. by deliberating. It is necessary to say something about my reasons for including Moore (an explicit non-hedonist, or an "ideal utilitarian") in this thesis. There are several reasons for this. Despite his rejection of hedonism as an answer to the question of what is intrinsically valuable, Moore defended a version of act-consequentialism. According to Moore:

> Our 'duty', therefore, can only be defined as that action, which will cause more good to exist in the Universe than any possible alternative. And what is 'right' or 'morally permissible' only differs from this, as what will *not* cause *less* good than any possible alternative. (Moore, 1903, p. 198)

The structural similarity of this consequentialism with AU is obvious. The difference between AU and Moore's ethics concerns the axiological component. Their teleological structure is common ground[32].

Most of what Moore says in "Ethics in Relation to Conduct" bears relevance to different consequentialist theories in general and AU is no exception. It could even be claimed that, due to Moore's more "pluralistic" theory of value, e.g. that all pleasure need not be intrinsically valuable (according to Moore some pleasures can even be positively bad) and his idea of organic wholes, the problem of action guidance could be taken to apply

[32] There are further differences, e.g. that Moore kept the possibility of actions having intrinsic value open, but I will ignore them. Because my arguments is directed at a modified version of Moore's method, I believe no injustice is done by adopting this strategy.

with an even greater strength than it does to AU[33]. However, my aim is not to examine Moore's version of AU, but to examine the possibility of an AU-agent using, *mutatis mutandis*, Moore's method of moral decision-making when deciding what to do in practical situations of choice.

Another reason why an investigation of Moore's method in this thesis is both natural and relevant is what Moore himself says in Moore (1903) in relation to an important part of his method, namely rules which are generally recognised, generally practised and useful:

> And (2) these rules […] can be defended *independently of correct views upon the primary ethical question of what is good in itself.* (Ibid. p. 207. My italics.)

This is, strictly speaking, only a reason for including the rule-based part of Moore's method, but it is easy to modify the rest of the method to fit whatever axiology AU is supplemented with.

All through chapter V in Moore (1903), Moore is upholding a sharp distinction between a criterion of rightness and a method of decision-making. My objective is to examine the prospect of an AU agent using a modified version of Moore's method of decision-making when trying to act morally. Moore defends a version of the restricted deliberative strategy.

What makes Moore's theory special is that he argues that agents are never *justified* in violating a generally accepted and useful common sense rule of conduct. Despite Moore's scepticism regarding our ability to determine whether a particular action is right or wrong, he defends the possibility of having practical normative knowledge of a probabilistic general kind. Moore claims that we can know that the general adherence to certain rules of conduct would, in general, be useful[34]. The rules Moore is trying to vindicate are, again, the rules of common sense[35]. Moore argues using this and an additional argument from epistemic limitation, that an agent should *always* adhere to these rules in particular situations of choice.

6.1 Moore's method of moral decision-making

"Ethics in Relation to Conduct" begins with stating the general question to which the chapter is devoted, and makes certain important qualifications of its scope. The question he will try to answer is: What ought we to do? This question

---

[33] I have in mind here the fact that it seems to complicate matters if we have to determine whether a particular pleasure is intrinsically valuable or not. Moore thought that not all pleasures are good ones (e.g. pleasure resulting from sadism). Another complicating factor has to do with the fact that Moore suggested that deserved suffering might make the world a better place. Determining desert is often a tricky affair.

[34] It is not clear how Moore uses 'useful' in this context. I will return to this issue below.

[35] Again, the 'rules of common-sense morality' is not a particularly determinate conception, and specifying these rules is not an easy matter.

> […] introduces into Ethics, […] an entirely new question-the question what things are related as causes to that which is good in itself; and this question can only be answered by an entirely new method- the method of empirical investigation; by means of which causes are discovered in the other sciences. To ask what kind of actions we ought to perform, or what kind of conduct is right, is to ask what kind of effects such actions and conduct will produce. Not a single question in practical Ethics can be answered except by a causal generalisation. (Ibid. p. 196)

Moore writes in terms of *kinds* of actions and *kinds* of effects, i.e. of types rather than of tokens[36]. General rules of conduct must be formulated as either pre- or proscribing rather general kinds of actions. Moore's categorical prohibition on murder is an example of this. The opposite of 'general', in this sense, is 'specificity', and could be exemplified by the following rule: "Never commit murder unless you are a cricket-player shorter than five feet, own a Jack-Russell puppet younger than two month of age, …" When discussing general rules of conduct, Moore seems to have in mind *very* general rules. *Always* tell the truth, *Always* keep your promises, *Never* commit murder etc. But the rules Moore wants to defend are the rules 'recognised by common sense', and it is not quite clear how general these rules are. Consequently, it is not quite clear *which* rules they are. This is a source of great complication in Moore's approach, as well as for the other approaches discussed in this chapter. As the quotation indicates, Moore thought that actions of a certain kind have certain *kinds* of effects. Is it plausible to claim that every instance of a certain kind of action, say murder, has effects of a certain kind? Well, one kind of effect that is common to all instances of murder is that at least one human being is killed. But this is true by definition, and therefore of only marginal interest in the present inquiry. Are there any other kinds of effects that every 'murder' have in common? The important question, from the point of view of AU, is whether there is any symmetry between the effects, common to all actions of murder, with regard to their value. Are the effects of every instance of murder bad? This does not seem very plausible. That *some* murders can have good or at least indifferent effects, according to AU, is a more plausible hypothesis. And Moore seems to concur in this[37]. Moore defended a more modest thesis:

> An ethical law has the nature not of a scientific law but of a scientific *prediction*: and the latter is always merely probable, although the

[36] It seems, though, that he makes an exception to this when addressing the question of directly estimating the effects of an action in individual decision-making. Be that as it may, the difference is one of degree, rather than of kind. A sufficiently qualified and precise specification of a type, will only have one token, or at least so one could argue.

[37] At least he says "We can secure no title to assert that obedience to such commands as 'Thou shalt not lie', or even 'Thou shalt do no murder', is *universally* better than the alternatives of lying and murder." (Ibid. p. 204)

> probability may be very great. An engineer is entitled to assert that, if a bridge be built in a certain way, it will probably bear certain loads for a certain time; but he can never be absolutely certain that it has been built in the way required, nor that, even if it has, some accident will not intervene to falsify his prediction. With any ethical law, the same must be the case; it can be no more than a generalisation: and here, owing to the comparative absence of accurate hypothetical knowledge, on which the prediction should be based, the probability is comparatively small. (Ibid. p. 204-205)

Are the effects of most instances of murder bad? This claim is, intuitively, more plausible. But as we will see below, Moore uses this generalisation to justify the *universal* adherence to this kind of moral rules or practical laws. This, I will argue, is not a legitimate conclusion.

Moore's method could be schematically represented as a step-by-step procedure. The method can be represented in the following scheme:

A moral agent, confronted with a situation of choice, should go through the following steps when trying to decide what to do.

1. Determine if there is a *generally recognised and generally practised useful rule* that applies in the situation[38]. If there is: Always adhere to the rule[39].
2. If a rule applies that is generally recognised and generally practiced but not useful: Either adhere to the rule anyway or engage in a "direct consideration"[40].
3. Determine if there is any *alternative* general moral rule, that *would* be better if generally recognised, that applies in the situation. If there is, and in so far that adhering to this alternative rule tends to break down the existing custom: Adhere to the alternative rule. (Cf. Ibid. p. 213-214)[41]
4. If no generally recognised and generally practiced rule applies and no useful alternative general rule can be established: Engage in a "direct consideration". (Cf. Ibid. p. 214-215)

[38] "The individual can therefore be confidently recommended *always* to conform to rules which are both generally useful and generally practiced." (Moore, 1903, p. 213)
[39] Ibid.
[40] "There is, therefore, a strong probability in favour of adherence to an existing custom, even if it is a bad one. But we cannot, in this case, assert with any confidence that this probability is always greater than that of the individual's power to judge that an exception will be useful […]. (Moore, p. 213)
[41] Whether or not the agent should adhere to the new rule depends on how her choice will affect the conduct of other persons. If she can make others follow her example, thus exchanging the old rule with a better one, this will be what she ought to do. Determining this 'in cases of doubt' seems to require some deliberation in the form of direct estimates. "It seems, therefore, that, in cases of doubt, instead of following rules, of which he is unable to see the good effects in his particular case, the individual should rather guide his choice by a direct consideration of the intrinsic value or vileness of the effects which his action may produce." (Ibid. p. 214)

Moore's way of determining the relevant alternatives has already been discussed in chapter III, so I will not comment on this. Furthermore, when Moore speaks of making a "direct consideration" he has in mind roughly what I have been calling the deliberative approach.

6.2 General rules of conduct

This thesis deals with individual decision-making and not, at least in any direct way, with the question of which general rules ought to be observed in a society. This much being said however, we must, to understand Moore's method, make a digression to discuss the notion of general rules of conduct. The idea of 'generally recognised and generally practised useful rules' plays an important role in Moore's proposed method of individual decision-making. He discusses the relation between general rules of conduct and individual methods of decision-making and presents his version of the restricted deliberative strategy.

Some initial remarks are necessary. The general usefulness of a rule should be understood as the general usefulness of the actions that, as a result of agents adhering and/or trying to adhere to the rules, are performed. What Moore tries to do is to provide a vindication or *rationale* for the rules most universally recognised by common sense, i.e. rules that are generally accepted and useful. These generally observed common sense rules come in different orders according to Moore. Let us call them the stronger and the weaker order of rules. The strong order of rules are those of which it can be shown that their general observance is necessary for upholding a civilised social order, i.e. preventing something like a state of nature. The weaker variety of rules includes those of which it can be shown that the observance of them is useful in a particular state of society. Moore:

> If, now, we confine ourselves to a search for actions which are *generally* better as means than any probable alternative, it seems possible to establish as much as this in defence of most of the rules most universally recognised by Common Sense. (Ibid. p. 205)

Remember that Moore is talking about act-*types*, rather than act-*tokens*. He tries to show that some act-types are generally more useful than others and that the actions governed by the 'rules most universally recognised by Common Sense' are among these.

Moore mentions a few rules that can be shown to be useful in the strongest sense, i.e. in the sense that they are necessary in every state of society. Examples of these kinds of rules are the prohibition of murder, theft, lying and the breaking of promises[42].

[42] Moore also mentions the dispositions of industry and temperance in this context. (Cf. Ibid. p. 206)

These most fundamental or important rules, share two characteristics according to Moore. Firstly:

> (1) They seem all to be such that, in any known state of society, a general observance of them *would* be good as a means. The conditions upon which their utility depends, namely the tendency to preserve and propagate life and the desire of property, seem to be so universal and so strong, that it would be impossible to remove them; and, this being so, we can say that, under any conditions which could actually be given, the general observance of these rules would be good as a means. For, while there seems no reason to think that their observance ever makes a society worse than one in which they are not observed, it is certainly necessary as a means for any state of things in which the greatest possible goods can be attained. (Ibid. p. 207)

Notice the 'can' in the last row. Moore takes general observance to these rules to be a necessary, though not a sufficient, condition for 'attaining the greatest possible goods'. Secondly:

> And (2) These rules, since they can be recommended as a means to that which is itself only a necessary condition for the existence of any great good, can be defended independently of correct views upon the primary ethical question of what is good in itself. On any view commonly taken, it seems certain that the preservation of civilised society, which these rules are necessary to effect, is necessary for the existence, in any great degree, of anything which may be held to be good in itself. (Ibid. p. 207)

But Moore claims that "[a] similar defence seems possible for most of the rules, most universally enforced by legal sanctions[43] [...] and for some of those most commonly recognised by Common Sense […]" (Ibid, p. 206). Other rules can only be shown useful in certain conditions of society. Moore mentions only "[…] rules comprehended under the name of Chastity […]" (Ibid. p. 208) in this category:

> But it is not difficult to imagine a civilised society existing without them; and, in such a case, if chastity were still to be defended, it would be necessary to establish that its violation produced evil effects, other than those due to the assumed tendency of such violation to disintegrate society. Such a defence may, no doubt, be made; but it would require an

[43] If an action is illegal, then our reasons for not performing it is even greater than it is if the action only violates a generally accepted and useful common-sense rule of conduct: "One of the chief reasons why an action should not be done in any particular state of society is that it will be punished; since the punishment is in general itself a greater evil than would have been caused by the omission of the action punished. Thus the existence of a punishment may be an adequate reason for regarding an action as generally wrong, even though it has no other bad effects but even slightly good ones. The fact that an action will be punished is a condition of exactly the same kind as others of more or less permanence, which must be taken into account in discussing the general utility or disutility of an action in a particular state of society." (Moore, 1903, p. 208-209)

> examination into the primary ethical question of what is good and bad in itself, far more thorough than any ethical writer has ever offered to us. Whether this be so in this particular case or not, it is certain that a distinction, not commonly recognised, should be made between those rules, of which the social utility depends upon the existence of circumstances, more or less likely to alter, and those of which the utility seems certain under all possible conditions. (Ibid. p. 208)

The difference between the two orders of rules is, at least in theory, important. If Moore is right, if some rules can be defended independently of which theory of intrinsic value that one wants to defend, then this fact would make some practical questions easier to answer. I will not pursue this issue any further. The reader is asked to keep in mind that my present object of investigation is a modified version of Moore's method. Presumably, a hedonistic theory of value would result in different weaker order rules than would a Moorean theory, but I will not try to account for the details of this difference.

Moore is careful to point out our limitations in how much we can show as to the usefulness of these rules.

> We can secure no title to assert that obedience to such commands […] is *universally* better than the alternatives […] no more than a *general* knowledge is possible […]. (Ibid. p. 204)

But, contrary to this, he argues that these rules ought to be universally obeyed. This is due to the cognitive limitations shared by every individual agent. Moore poses the question: "Can the individual ever be justified in assuming that his [rule violation] is one of these exceptional cases?" (Ibid. p. 211) Moore adduces three reasons for 'definitely answering the question in the negative'.

The first reason is that

> […] if it is certain that in a large majority of cases the observance of a certain rule is useful, it follows that there is a large probability that it would be wrong to break the rule in any particular case[…]. (Ibid. p. 211)

Remember that we presuppose that the agent has a will to approximate the overall goal of AU. As Shaw points out, this reason's plausibility hinges on how one characterises the situation in terms of the degree of knowledge the agent possesses.

> In the pertinent case, however, the moral agent is claiming to know something further about the situation that makes it reasonable to believe that better results would come from ignoring rule R than following it. Ex

> hypothesi, the agent believes that the present circumstances are such that what is usually the case is not now the case. (Shaw, 1995, p. 150)

How can the agent be more confident in her belief about general utility, than with the particular in this specific case? Does not general utility consist of the aggregation of particular utilities? It seems highly unlikely that these cases would not sometimes arise. And this is enough, it seems, to refute Moore's claim that the agent ought always to abide by the rule.
Secondly,

> […A]nd the uncertainty of our knowledge both of effects and of their value, in particular cases, is so great, that it seems doubtful whether the individual's judgement that the effect will probably be good in his case can ever be set against the general probability that that kind of action is wrong. (Moore, 1903, p. 211-212)

Thirdly,

> Added to this general ignorance is the fact that, if the question arises at all, our judgement will generally be biassed by the fact that we strongly desire one of the results which we hope to obtain by breaking the rule. (Ibid. p. 212)

Moore claims that we ought always to obey rules which are generally observed and are seen as useful:

> It seems, then, that with regard to any rule which is *generally* useful, we may assert that it ought *always* to be observed, not on the ground that in *every* particular case it will be useful, but on the ground that in *any* particular case the probability of its being so is greater than that of our being likely to decide rightly that we have before us an instance of its disutility. (Ibid. p. 212)

This does not amount to a form of rule-utilitarianism, but the prescriptions of the methods are similar. Moore's criterion of right action is act-utilitarian, yet, due to our cognitive limitations, we ought always to follow these rules.

Is this a plausible contention? I think not. There are several problems with it. Telling the truth, it may be admitted, is *generally* useful. But if some lunatic threatens to kill a person if I do not lie, maybe I ought to tell this one lie. *This* one lie is not likely to undercut the general institution of truth-telling. And, it might be argued, I have at least as good a reason for believing that this exception to the rule will have a more favourable outcome, as I have in believing that the following of the rule will have. (Other examples lend themselves to be plausibly used as further counterexamples, i.e. the desert island death-bed promise.)

Moreover, generally useful rules very often give conflicting prescriptions. And the very assumption to regard a rule as useful may be questioned. How do we know that a rule of conduct is generally useful? Moore maintains that it is only actually followed rules that can be shown to be generally useful[44]. Though, one may agree that an actually followed rule could be shown to be *useful*, it is still quite a step from this to show that the rule is part of the best set of rules that we can imagine, i.e. the best approximation to an optimal set of rules.

Furthermore, these rules seem possible to adhere to in different degrees. Take for instance the rule of helping fellow humans in need. We can help persons from time to time, or we could make it our sole occupation in life.

Moore argues that a rule issuing a respect for property is generally useful. But there are different ways of construing property rights. These different systems of property probably will have different outcomes in terms of utility. According to James Griffin:

> How far can we assess an institution of property? We can assess egregiously bad forms of it. We can assess this or that part of decent forms of it. But institutions of property in advanced societies are so complex that there will be a wide range of acceptable forms that no doubt differ in quality among themselves, but that we have no hope of ranking. (Griffin, 1992, p. 128-129)

AU would have us implement the *best* institution of property, not one that is just 'good enough'. And even if we can identify particularly bad institutions, this does not take us very far in determining how we ought to set up the best institution.

Moore seems to think that there is a presumption in favour of rules recognised and observed by common sense. One problem with other rules is, according to Moore, that "[…] the actions which they advocate are very commonly such as it is impossible for most individuals to perform by any volition." (Moore, 1903, p. 209) This is because of the 'peculiar dispositions' needed to will such actions. Moore does not give any examples of such actions, but Shaw argues[45] that examples of these dispositions would be to love our enemies and not coveting our neighbour's possessions. It is certainly possible that Moore is right that with regard to several alternative moral rules, they would be too demanding for ordinary people. But this could hardly be true of every alternative rule[46]. For example, the rules: 'Do not eat meat from factory farmed animals' and 'Rich people to donate their surplus of wealth to famine relief' are hardly of this kind.

---

[44] Cf. Ibid. p. 209-210.
[45] Cf. Shaw (1995), p.141.
[46] Furthermore, if 'impossible', is taken in its usual sense, then the rules prescribing actions impossible to perform by the agent, simply are not normative for the agent. This is because "ought" implies "can".

Another problem with rules that are not prescribed by common sense is that although the actions they recommend "[…] themselves are possible, […] the proposed good effects are not possible, because the conditions necessary for their existence are not sufficiently general."(Moore, 1903, p. 210) This is because human nature is what it happens to be, i.e. wanting in certain respects. Perhaps absolute pacifism would be an example, but my two examples mentioned above are not obviously, if at all, of this kind.

A third alleged defect of alternative rules is the following: "(3) There also occurs the case in which the usefulness of a rule depends upon conditions likely to change, or of which the change would be as easy and more desirable than the observance of the proposed rule." (Ibid. p. 210) Again, the two examples mentioned above seem to escape even this kind of defect.

If, as I have argued, there are rules which are not generally observed in society, but yet do not suffer from the defects pointed out by Moore, these rules would be relevant too. The usefulness of these rules, if they are possible to implement in society, should be compared with the common sense rules.

How do we determine if a rule is useful? Moore's use of the concepts of 'usefulness' and 'useful' is ambiguous. On the one hand he maintains that "…'right' does and can mean nothing but 'cause of a *good* result,' and is thus identical with 'useful'…" (Ibid. p. 196. My italics.) But on the other hand he maintains that "what is 'right'[…] only differs from this [i.e. 'our duty'], as what will *not* cause *less* good than any possible alternative." (Ibid. p. 198) Moore exploits this ambiguity in the transition from his act-consequentialist criteria of right, to the establishment of the general usefulness of generally observed rules of conduct. 'Useful', presumably, can be used to refer to many different things. A rule or an action could be called useful if the general recognition of and the general practice of adhering to the rule a) makes a positive utility contribution to the universe as a whole (as compared with some alternative rule or action, or b) is valid, i.e. adhering to the rule makes (us) approximate the overall goal of AU. In which of these senses should a rule be 'useful', in order for it to be justified?

Well, it seems obvious that if it is useful in the first sense this gives us *some* reason for adhering to the rule. But this is not sufficient to satisfy AU's demands. The rule has to be a good means to approximate the overall goal of AU standard. Trivially, the rules *necessary* for a 'civilised social order' are observed to a sufficiently high degree if our social order *is* civilised. But of course, this shows, at best, that these rules are useful in one sense, e.g. that the observance of these rules are better than the observance of certain other rules (and or better than having no rules at all). Perhaps some *qualified versions* of these rules would make the social order even more civilised. There could also be *additional* rules which, if followed, would make the outcome even better. For a set of rules to be supported by AU the set ought to be the most valid set; and for us to have any reason for believing that a certain set of rules *is* the

most valid set, we would need some reason to believe that adherence to these rules makes agents approximate the overall goal of AU. As I have already argued, it is hard to come by good reasons for believing, of any specific set, that it is.

It is also important to keep in mind how little is won, even if we grant that Moore has established that a generally observed rule of conduct is useful. From the point of view of an individual agent deliberating over what to do in a practical situation of choice, this can hardly be of much help. The most that it could be maintained that Moore has shown is that the *general* observance is useful. If Moore is right, when considering rules which are *not* generally practised, that "[…]there is a large probability that he will not, by any means, be able to bring about its general observance[…]" (Ibid. p. 210), then this should apply also in the parallel case of generally observed rules. It could, per analogy, be maintained that one agent could seldom, if ever, affect the general observance in either way; neither enforce the general observance of the rule nor make the general observance deteriorate. The argument is a double-edged sword, that cuts in both directions. Even if this is true of many agents (with little or no influence in questions of social policy), those of which it is not true hardly enforce or diminish general observance mainly through observing these rules themselves, rather they have the power of political, legislative and/or social influence to affect the behaviour of others in this regard.

Moore does not provide a solution to the problem of demarcation. Although Moore claims that certain rules ought always to be adhered to, there are still many situations in which the agent will be in doubt whether to engage in a "direct consideration", i.e. to use the deliberative approach, or whether she ought to adhere to a secondary rule of conduct. There will be cases in which the agent is in doubt regarding the degree of usefulness of a particular rule (practiced as well as non practiced ones) as well as cases where the agent is in doubt regarding the effects of her violation or adherence to a certain rule on the future adherence frequency of herself as well as other agents.

7.0 Assessing the restricted deliberative strategy

What should we say about the validity and practicability of the restricted deliberative strategies defended by Tännsjö, Sidgwick, Moore and Hare, then? I have interpreted their respective methods of decision-making as involving two main parts, the deliberative approach and the secondary rule-following approach. They are all different versions of the *restricted deliberative strategy*. There are three major points that affect the validity and practicability of these methods of decision.

The first is the deliberative approach. Obviously, the evaluation of this approach will depend on how we conceive of the "input" of this approach, i.e. what criteria of adequacy we accept for this "input". The validity and practicability of the deliberative approach was discussed in chapter III. The

conclusion was that an AU-agent with normal cognitive abilities is not in a position to justify that the deliberative approach, interpreted in the reasonable subjectivist way, is a sufficiently valid method of decision-making. The method is probably also relatively impracticable for human AU-agents.

The second point concerns the use of secondary rules. The third concerns the drawing of the demarcation line between the different approaches, i.e. between deliberating and adhering to secondary rules. The validity and practicability of these methods of decision will depend on all three of these elements. A tentative overall judgement, as to the validity and practicability of the methods, will be made in the end of this chapter.

In trying to evaluate these methods of decision, I will proceed in the following manner. I will take up general criticism that seems to apply to all these methods. For example, in different degrees the problem of demarcation affects all there proposed methods. The writers are hard pressed to specify under which circumstances the AU-agent ought to use the deliberative approach or when to adhere to secondary rules instead. A human AU-agent is very unlikely to be able to provide good reasons for preferring one particular point of demarcation to every other possible point.

7.1 Assessing trying to adhere to secondary rules

I will not comment on any specific systems of secondary rules. The writers I have examined in this essay have not given precise formulations of their respective system. And there are enough general problems with the approach as to make the evaluation interesting.

In so far as a method of decision prescribes the following of a specific set of secondary rules of conduct, the method seems relatively practicable. If an agent tries to abide by the rules, she is likely to succeed in doing so. These rules are supposed to be relatively general. They are designed to be relatively easy to use, that is part of their purpose. But how much support from AU do actions following from adhering to these secondary rules have? How well does trying to adhere to secondary rules meet the validity desideratum? This is, of course, a very difficult question to answer. But there are reasons for being sceptical regarding the validity of this approach. First, whichever set one considers, it seems highly unlikely that there is not another set that has a higher validity than this one. Suppose that the rules of common sense could be given a sufficiently determinate understanding. It is not plausible to believe that the rules of common sense, under this particular interpretation, are the most effective means to maximise utility. Adhering to these rules might well have a favourable outcome, but they could hardly be thought to be the best possible set[47].

[47] Cf. Parfit (1984) p. 40.

Another reason is that adhering to secondary rules of conduct is relatively unspecific. Because of this it allows for a number of different actions in any particular situation of choice. Thus, it does not determine the course of action in a way that lets us determine whether the method's validity is relatively low or relatively high. Or so it might be argued. But in order to justify adherence to a particular set of secondary rules, an AU-agent would need to establish that her adherence to that particular set is the most effective means for her in approximating the overall goal of AU. Because of this indeterminate connection that these rules have to AU, justifying the rule-adherence with reference to AU's overall goal is *very* hard. That the agent is hard pressed to provide such a justification means that she does not possess sufficient reasons for believing the set to be the most valid set. Even if the set were actually the most valid set of all, relative to AU, this is not something that the agent is in a position of establishing. The cognitive shortcomings of human AU-agents, which was the most important reason for introducing the secondary rule approach, seems to push us back to square one:

> What has happened is that we have eliminated the Problem of Doubt at the level of applying moral principles to acts, only to have it re-appear at a higher level, the level of choosing auxiliary rules. Uncertainty as to which act satisfies the principle [AU] has to be replaced by uncertainty as to which auxiliary rule would be most appropriate. Either form of uncertainty prevents the agent from constructing a bridge between her principle and an act to be done. Since this difficulty is just a reappearance of the Problem of Doubt, we might suppose it could be solved at this new level in the same way we solved it at the lower level. Suppose the agent's uncertainty about which auxiliary rule to apply has the following source: she possesses a standard for ranking the different auxiliary rules, but lacks sufficient information about the rules to determine for certain which one satisfies her standard. (Smith, 1988, p. 100-101)

We might take the 'standard' to be the validity desideratum. Thus, according to this standard, an auxiliary rule, R, is to be preferred to another, R', if and only if, R makes the agent adhering to R approximate the overall goal of AU to a greater extent than trying to adhere to R' would. That a human AU-agent is ever in a position to justify a belief to the effect that this standard is satisfied for a specific system of secondary rules seems highly unlikely.

As far as we can tell, there are many possible systems of secondary rules that could be claimed to obtain *some* support from AU. Because we can hardly be justified in assuming that one of these systems is more valid than another, and that these sets might well give incompatible prescriptions, it seems plausible to conclude that, for each of these systems of secondary rules, an AU-agent cannot justify her belief that adhering to one such system of secondary rules, rather than another, is the most valid method of decision-making for her.

It seems rather uncontroversial to conclude that the problems with determining the degree of validity of trying to adhere to specific sets of secondary rules, for a human AU-agent, are significant. A human AU-agent is very unlikely ever to be justified in believing that a particular set of secondary rules is the most valid set, and this is what she has to establish in order to be justified in trying to adhere to this set from the point of view of AU.

7.2 The problem of demarcation revisited

The problem of demarcation, i.e. the problem of determining when to try to adhere to the deliberative approach and when to try to adhere to secondary rules of conduct, is a source of impracticability for every method of decision-making that does not give a solution to it. If the method does not specify, in a way that is practically action-guiding for the AU-agent, when the deliberative approach is to be used and when secondary rules ought to be followed, then the method will not be successful in guiding actions in practice. Because settling this question, with reference to AU's overall goal, is unattainable for human agents, due to our inability to justify our beliefs in complicated empirical facts and interpersonal utility-comparisons, the restricted deliberative strategy remains a highly impracticable method of decision-making for AU.

8.0 Conclusions

That secondary rules are a necessary part of any plausible method of decision for AU can hardly be doubted. The problems with the unrestricted deliberative strategy made this clear. The problem is giving reasons, from the point of view of AU, for preferring one set of secondary rules to another. We know *that* a set of secondary rules is needed, but not *which* set.

Another problem concerns how to specify the situations in which an AU-agent is justified in deliberating and the situations in which she ought rather to use some secondary rule. Giving reasons, in terms of AU, for drawing the line at a certain place, rather than another, is not an easy task. None of the thinkers managed to present a plausible point of demarcation, i.e. giving a determinate account of when to adopt the deliberative strategy and when to follow secondary rules.

Trying to adhere to secondary rules is a relatively practicable way of making decisions in many situations, at least if the rules are given determinate formulations. If one tries to adhere to them one will often succeed. Remember that our point of departure was the question of what an agent who wants to let AU guide her actions ought to do. Incorporating secondary rules in one's method of decision, then, tends to make the method more practicable. Of course, these rules must be relatively well specified, the situations in which they are to be followed also needs to be specified and the rules must be

coherent (at least they must be complemented with some device for preferring one over another in cases of potential conflict).

None of the methods of decision-making that I have considered in this essay have succeeded in establishing a set of secondary rules of conduct and justified this set by showing it to be preferable, according to the AU standard, to every other possible set. The justification would, at least to a certain degree, be accomplished if the preferred set could be shown, through using the deliberative strategy of maximising expected utility, to be the best set. However there are innumerably many possible sets that would need to be compared in order to reach a decision as to which set would be the most valid. And we cannot have justified beliefs about how close an agent adhering to one set, rather than another, will come in approximating the overall goal of AU.

Chapter V
On the justification of AU

1.0 Introduction

I have argued that there are overwhelming problems with trying to apply AU (construed as a criterion of rightness) in practical situations of choice. No proponent of AU has presented a method of decision-making such that we are justified in believing that it meets the desiderata of practicability and validity to a sufficient degree. In this chapter, I turn to the implications of these results. What are the consequences for AU? What are the consequences for us?

One set of questions concern the truth of AU. Does the problem of applying AU entail that it cannot be true? Given a realist view on the truth of moral principles, it seems obvious that the answer is No. However, some theorists are constructivists. They believe that there are moral facts but that these are human constructions. Given some versions of such a view, if we cannot know which actions are right, according to AU's standard, then AU cannot plausibly be taken to be true. Accordingly, I shall argue that AU is much more easily reconcilable with realism than with these versions of constructivism. Accepting some versions of the latter view would cast considerable doubt on the putative truth of AU.

Another set of questions concerns the justification of AU. It might be thought that the fact that AU entails that we cannot determine whether a particular action is right or wrong leaves us with no reason for accepting AU. After all, presumably, to help us answer such questions is part of the whole purpose of moral theories. For example, it has been argued that the problem of applying AU entails that AU cannot be justified through the method of reflective equilibrium. However, I shall argue that such a sceptical conclusion is premature. Although the sceptical conclusion about the possibility of applying AU implies that there is no straightforward way of testing AU against our intuitions concerning the rightness of actions in actual situations, the method of reflective equilibrium leaves room for other arguments and strategies. I shall sketch one such strategy; a strategy that allows us to see that a belief in AU may still be justified, despite its implications for the impossibility to determine how we ought to act in practice.

The plan of this chapter is as follows. AU entails that we cannot determine whether a particular action is right or wrong in a particular situation of choice. I shall call this implication 'practical scepticism'. This position will be characterized in more detail in section 2.0. In sections 3.0-3.3. I start by

exploring the consequences of the fact that a theory has this implication for its truth. I argue that, given some versions of moral constructivism, if a theory entails practical scepticism, it cannot be true. In section 4.0-4.1, I turn to the question of justification. In particular, I examine an argument to the effect that, if a moral theory entails practical scepticism, it cannot be tested and justified through the method of reflective equilibrium. The main conclusion is that this is true, at best, if the judgements against which theories are tested is exhausted by normative judgements, whereas it is not true if we also include value judgements. Since there is no reason to exclude such judgements, the fact that AU entails practical scepticism does not exclude it being a justified theory. In section 5.0, I outline an alternative way of justifying a belief that AU is true based on judgements of value.

2.0 Practical scepticism

Scepticism comes in many varieties[1]. Some are *global*, others *partial,* some are *extreme* some are *moderate*. The form of scepticism I will defend in this chapter is partial (rather than global) and moderate (rather than extreme). The position I defend qualifies as a partial variety of scepticism, because it regards only a limited area of (possible) knowledge. It is a sub-species of normative scepticism, which in turn is a sub-species of moral scepticism. It is very important for my purposes that this is so. Particularly important is the fact that my favoured form of scepticism does not presuppose a global and extreme variety of scepticism. If this were the case, the implications for normative ethics, i.e. the relevance of this inquiry, would loose its bite. If the claim that we cannot, in practical situations of choice, know (or have justified beliefs regarding) what action satisfies the AU criterion of right action, were based on the idea that we cannot know anything (or nearly anything), the claim would be trivial[2].

Neither the natural sciences, nor common sense, should have any objection to accepting the crucial premise of practical scepticism, namely that the extremely complicated empirical questions that would have to be answered in order to determine if an action is optimific cannot be answered in a reliable way by normally endowed human agents.

By 'practical scepticism' I mean the doctrine that no actual (human) agent is justified in believing, in any particular situation of choice, that a particular action is morally right. I will not, within the scope of this essay, state

[1] My list of different forms of scepticism does not pretend to be exhaustive.

[2] Cf. Frazier (1994) p. 50 and Shaw (1995), p. 106. According to one particular variety of scepticism the future is out of our cognitive reach. This is a principled doctrine. According to one version of this doctrine, beliefs about the future cannot ever be true. There could be different reasons for this. One is that propositions about the future cannot be true because the future is not determinate. Practical scepticism does not imply scepticism about the future. But scepticism about the future implies practical scepticism. This means that practical scepticism is the 'weaker' theory of the two. If practical scepticism gained its force from scepticism about the future, then it would be a rather trivial matter.

necessary and sufficient conditions that a belief must satisfy in order to be justified for an agent. But I do not think there is any need. It suffices to say the following. We are here engaging in complicated predictions of future events and the problem of determining the relevant alternative set in the situation. It is commonly held that predictions of future events, even if they are restricted to a rather narrow area, are very speculative. Because of the fact that the optimificness of an action depends on what actually will happen, as compared to what would happen if another action were performed, over vast areas of space and time and that even small changes could affect the relative amount of realized positive and negative value, it seems plausible to assume that human AU-agents are not in a position to determine with adequate reasons whether an action is optimific or not. And this would be necessary in order to be justified in a belief that the action satisfies AU's criterion of rightness.

This view seems to have been accepted by Moore:

> In order to shew that any action is a duty, it is necessary to know both what are the other conditions, which will, conjointly with it, determine its effects; to know exactly what will be the effects of these conditions; and to know all the events which will be in any way affected by our action throughout an infinite future. We must have all this causal knowledge, and further we must know accurately the degree of value […] of all these effects; and must be able to determine how, in conjunction with the other things in the Universe, they will affect its value […]. And not only this: we must also possess all this knowledge with regard to the effects of every possible alternative; and must then be able to see by comparison that the total value due to the existence of the action in question will be greater than that which would be produced by any of these alternatives. But it is obvious that our causal knowledge alone is far too incomplete for us ever to assure ourselves of this result. Accordingly *it follows that we never have any reason to suppose that an action is our duty*[3]: we can never be sure that any action will produce the greatest value possible. (Ibid. p. 198-199. My italics.)

William H. Shaw claims that Moore sets unreasonably high epistemic standards when addressing this problem. Moore writes in terms of "to know", "be sure" and "to shew" that an action is right. Shaw argues that in setting the standards as high as Moore does, his claims are open to accusations of irrelevance:

> But scepticism of a global epistemological character is irrelevant here, because the denial that we have knowledge of, or justified beliefs about, our duty (as defined by the consequentialist principle) gets its force only insofar as our ignorance here is supposed to contrast unfavorably with the

[3] Notice the italicised part. *This* inference is surely a *non sequitur*. But we can replace 'any reason' with 'good reason' in order to make the claim more plausible. I will ignore this in what follows.

> security of our knowledge about other matters. The consequentialist need not worry about the sceptic who believes we lack knowledge of duty simply because he believes we lack knowledge of anything. (Shaw, 1995, p. 106)

Due to the complexity of the empirical questions which would have to be decided in order to determine the outcome of any action, the epistemic standards applied must be adapted to the nature of the questions. To demand certainty in predictions about future events and their causal relations, would of course be vain, perhaps even meaningless. As was argued in chapter II, the problem for AU is not that we cannot "know" or "be sure" that an action satisfies AU's criterion of right action. It is that even given weaker epistemic demands, satisfying demands like 'having good reasons to believe', they are still extraordinarily hard to satisfy. Moore's argument is thus applicable to weaker epistemic standards as well.

The conclusion from the previous chapters is that the acceptance of AU's criterion of rightness of actions (probably) leads to practical scepticism (for human agents). This is because it is extremely unlikely that a human agent, in a practical situation of choice, would be justified in believing that she has satisfied AU's criterion of rightness of actions. The doctrine does not claim that the agent cannot have strong reasons to judge an action to be superior, as far as the agent can tell, to another action. This scepticism need not be principled[4], but for the time being it should qualify as a principled version. The implications of practical scepticism can be divided into two categories:

(a) What are the implications of the fact that AU entails practical scepticism for the *truth* of AU?
(b) What are the implications of the fact that AU entails practical scepticism for the possibility of *justifying the belief* that AU is true?

The answer to (a) depends on how we conceive of the foundation of morality, whether realism or constructivism is the most plausible account of the foundations of morality. Thus, I will start with an examination of the implications of practical scepticism for the truth of AU, given a constructivist idea of the foundation of morality. I conclude that practical scepticism and some versions of constructivism do not go well together. A competing view is that morality has a realist foundation. When (b) is discussed I will assume this view. Here I ask whether it is possible to combine a belief in practical scepticism with a belief that AU is justified.

[4] That is, at least if we could give a plausible account of the relevant alternatives in situations of choice. This, however, it seems highly unlikely that we would ever be able to do.

3.0 AU and Moral constructivism

The reason why the previous results do not affect the combination of moral realism and AU is that, on a purely realist view, the truth of a moral judgement, is independent of our ability to achieve knowledge of it. Thus, the fact that AU implies that we cannot achieve such knowledge does not exclude AU from being true[5].

However, suppose that we instead accept a constructivist view on moral discourse. What implications does such an approach have for the truth of AU? I will concentrate my investigation on issues were the implications of practical scepticism are relevant. More specifically, the question is whether there are versions of constructivism that does not square well with AU and practical scepticism.

3.1 Definition of Moral Constructivism

According to *The Cambridge Dictionary of Philosophy*, 'ethical constructivism' is "[…] a form of anti-realism about ethics which holds that there are moral facts and truths, but insists that these facts and truths are in some way constituted by or dependant on our moral beliefs, reactions, or attitudes."[6] On this account norms, as well as values, are brought into the world by moral agents. There are no hard moral facts 'out there' in the world for us to find. It is we who, in one way or another, are the source of the validity of our moral judgements. However, there are many different versions of constructivism.

3.2 Mackie: Inventing morality

J. L. Mackie is, perhaps, an example of a proponent of the kind of view I have in mind here[7]. In his *Ethics, Inventing Right and Wrong* he defends the idea that morality is not to be discovered but invented. According to this view:

> Morality is not to be discovered but to be made: we have to decide what moral views to adopt, what moral stands to take. […T]he object is […] *to decide what to do, what to support and what to condemn, what principles of conduct to accept and foster as guiding or controlling our own choices and perhaps those of other people as well.* (Mackie, 1977, p.106. My italics.)

According to Mackie, we need moral norms and ideals to promote co-operation and to further our interests. Hence, we 'invent' the notions of 'moral rightness' and 'wrongness' and 'desert', 'good' and 'bad'. We need

---

[5] If moral discourse is given a realistic interpretation, if moral questions are a matter of objective fact, a part of the fabric of the universe, then it could be argued that the problem of practical scepticism is of only marginal concern (Cf. Parfit, 1984, p. 29). It is a problem with *us*, not with our moral theories (the moral facts as it were) that we cannot know what we ought to do in practical situations of choice. This is of course somewhat disappointing, but it is a defect of *ours*, according to realism, not a defect of AU.

[6] Audi (1995) p. 243.

[7] Mackie does not call his position constructivism, but I think that his view can plausibly be taken to be a version of this idea.

precepts like these to check on our behaviour, to make possible the fulfilment of our non-moral or pre-moral goals.

Mackie thinks of morality as "a device for counteracting limited sympathies" (Mackie, 1977, p. 107) More generally he argues, with reference to G. J. Warnock, that morality is needed because

> of certain general and persistent features of the human predicament, which is 'inherently such that things are liable to go very badly' - badly in the natural, non-moral sense that human wants, needs and interests are likely to be frustrated in large measure. Among the factors which contribute to make things go badly in the natural course of events are various limitations - limited resources, limited information, limited intelligence, limited rationality, but above all limited sympathies. Men sometimes display active malevolence to one another, but even apart from that they are almost always concerned more with their selfish ends than with helping one another. The function of morality is primarily to counteract this limitation of men's sympathies. We can decide what the content of morality must be by inquiring how this can best be done. (Ibid. p. 107-108)

He also points to problems of co-ordination, such as prisoner dilemmas. I think that it is quite obvious that the function of morality according to this idea must be to provide us with moral principles that can help us guide our actions in practical situations of choice. Morality is invented or constructed, at least partly, to solve *practical* problems. Moral principles are to provide methods of moral decision-making, helping us to cope with practical problems like "How do we come to terms with situations of a prisoner's dilemma kind?" What we need is norms that make people act in a way that does not lead to counter-productive results. To take the original example of the two prisoners, they need to have incorporated the norm that you do not ever squeal on your brother in crime. This norm would "solve" the practical dilemma in the sense that the outcome would be better for each of the agents compared to if they both confess. Due to different facts of the human condition, people need to restrict their actions in ways that let them, as a group as well as (often) as individuals, pursue their goals in more effective and secure ways. Some principles are better than others in this regard.

Mackie himself talks of act-utilitarianism as the 'ethics of fantasy' and stresses the 'impracticability' of AU. But Mackie's reason for taking act-utilitarianism to be impracticable is different from mine:

> But why, it may be asked, are such moralities of universal concern impracticable? Primarily *because a large element of selfishness* […] *is a quite ineradicable part of human nature.* (Mackie, 1977, p. 132. My italics.)

It is because moral agents are too selfish that adhering to the prescriptions of AU is highly impracticable. But if we accept the distinction between motives of actions and criteria of right action, and the dictum that ought implies can, this is of course no problem for AU[8]. If an agent's selfishness makes it *impossible* to perform some actions, e.g. those involving great sacrifices, then AU does not demand that these actions be performed. And although an agent is driven by selfish motives her actions can still be right. Motives are of no direct relevance to the rightness of actions, according to AU.

AU is unrealistically demanding in at least two senses. One is that we cannot satisfy AU because of our self-love. The other is that we cannot intentionally satisfy AU because we are not justified in believing that a particular action is the action prescribed by AU. Mackie argued along the first line that human agents have strong selfish dispositions. These dispositions would prevent them from acting morally right in many situations, according to Mackie. This is one kind of reason. This could be taken to mean that even when an agent has good reasons for believing that a particular action would be morally right, due to self-love she chooses another action. My argument is different. I assume that the agent is *motivated* to do the right thing. She would, I assume, willingly be prepared to sacrifice her own well-being or the like, for the total good. However, due to the truth of practical scepticism, she does not know which of her alternative actions that satisfy AU's criterion of rightness. Hence, my claim is stronger than Mackie's. The major problem for AU, given the meta-ethical setting provided by Mackie, is its impracticability in the sense argued for in this thesis. By the rather unrealistic assumption of ideal motivation on the part of the AU-agent, we can disregard the *further* problem regarding AU and practical action-guidance. Even an ideally motivated agent would not be able to satisfy AU's criterion of rightness of action. On Mackie's view we *do* have use for plausible and highly practicable methods of moral decision-making. Presumably, we have *no* use for a criterion of rightness of actions, that does not help us decide what to do in practical situations of choice though. AU's criterion of rightness of actions does not help us decide what to do in practical situations of choice. Because this is the end which an "invented" morality is designed to promote and because AU's criterion of rightness is useless relative to this end, AU cannot

[8] At least, that is, if realism is true. This is also pointed out by Parfit: "I also believe that, even if we became convinced that Consequentialism was the best ethical theory, most of us would not *in fact* become pure do-gooders. Because he makes a similar assumption, Mackie calls Act Utilitarianism 'the ethics of fantasy'. Like several other writers, he assumes that we should reject a moral theory if it is in this sense *unrealistically demanding*: if it is true that, even if we all accepted this theory, most of us would in fact seldom do what this theory claims that we ought to do. Mackie believes that a moral theory is something that we *invent*. If this is so, it is plausible to claim that an acceptable theory cannot be unrealistically demanding. But, on several other views about the nature of morality, this claim is not plausible. We may *hope* that the best theory is not unrealistically demanding. But, on these views, this can only be a hope. We cannot assume that this must be true." (Parfit, 1984, p. 29)

be plausibly regarded to be true. If we invent morality to co-ordinate our actions and make our actions predictable to one another and to enable us to further our own interests as well as the interests of others, AU is not a good tool for accomplishing our end. This is due to AU's lack of practical action-guidance.

One way, and indeed the most natural way, of practically justifying a moral principle, given this account of morality, is to show how the acceptance of, and our abiding by, the norm actually achieves the goals which morality is invented to promote. But the acceptance of AU's criterion of right action does not have any determinate practical consequences that we can have justified beliefs about. It does not help us solve our practical problems. Thus, this way of justifying moral principles does not seem open in the case of AU. This is a strong reason for believing that this kind of constructivism and AU does not constitute a viable theoretical position[9].

3.3 Rawls: 'Kantian constructivism'

Another example that would qualify as a moral constructivism is John Rawls' position in his Dewey-lectures. One idea that seems to exclude AU from being true, given a constructivist account, is provided by Rawls' 'publicity condition'; a condition that is intended to constrain the choice of the parties of the original position. This condition is in part intended to capture the role of impartiality in Rawls' account of moral reasoning. However, it also involves more substantive requirements.

> Thus a conception of justice is framed to meet the practical requirements of social life and to yield a public basis in the light of which citizens can justify to one another their common institutions. Such a conception need be only precise enough to achieve this result. […] The moral conception is to have a wide social role as a part of public culture and is to enable citizens to appreciate and accept the conception of the person as free and equal. Now if it is to play this wide role, a conception's first principles cannot be so complex that they cannot be generally understood and followed in the more important cases. Thus, it is desirable that knowing whether these principles are satisfied, at least with reference to fundamental liberties and basic institutions, should not depend on information difficult to obtain or hard to evaluate. To incorporate these desiderata in a constructivist view, the parties are assumed to take these considerations into account and prefer (other things equal) principles that are easy to understand and simple to apply. The gain in compliance and willing acceptance by citizens more than makes up for the rough and

[9] For a similar conclusion see Miller (2003) p. 61, where he writes "[…W]hatever types of beings the universe may contain, moral standards are only meant to apply to humans – a proposition that might follow from a conception of morality as a human construct or institution. From this meta-ethical proposition, …[together with additional premises] …, it would follow that if no possible humans can satisfy a moral standard then it cannot be a valid one." Of course, it is intentionally satisfying the moral standard that is interesting in this context.

> ready nature of the guiding framework that results and its neglect of certain distinctions and differences. (Rawls, 1980, p. 561)

Rawls here seem to apply a practicability desideratum to a moral conception's first principles. By setting up these requirements that a moral principle must meet in order to be true, Rawls' account becomes harder to combine with AU. Knowing whether AU is satisfied surely 'depends on information difficult to obtain or hard to evaluate'. Neither is it 'simple to apply'. If desiderata such as these are given the status of *requirements*[10] on moral truths, then this squares badly with AU.

To sum up, some versions of moral constructivism set up requirements on moral truth that are incompatible with practical scepticism. Mackie would have us invent moral principles that help us promote our pre-moral goals in practice. Rawls takes the fact that a moral principle is 'simple to apply' to be a desideratum. If desiderata such as these are taken to be *necessary* conditions for the truth of moral principles, then AU is not a moral truth.

4.0 Does practical scepticism yield epistemic scepticism?

In this section, I will criticise previous attempts at justifying AU's criterion of rightness from the point of view of coherentism, or reflective equilibrium. The attempts I will criticise deals with normative properties such as 'rightness' and 'wrongness'.

One way of justifying beliefs about rightness and wrongness of actions, given AU, would be if we could observe the optimificness (or non-optimificness) of actions. In his famous example[11] about the young hoodlums igniting a neighbourhood cat, Gilbert Harman argues that although we can 'observe' that the hoodlums are doing something morally wrong; we are not justified in inferring from this observation that the act has the property of 'moral wrongness'. This is because the putative property of wrongness of the action is not a part of the best explanation of why we make the immediate judgement, without any conscious reasoning going on, that the action is morally wrong. The best explanation of our judgement is instead that we have received a certain moral education in our upbringing, conditioning us to condemn such acts. More generally, Harman's thesis is that moral facts or properties do not play any part in our best explanations of any observations and that this makes a strong case against moral realism.

---

[10] 'Requirements of practicability' on moral principles are also proposed by Gruzalski (1981) and Smith (1986). Gruzalski argues that a probabilistic interpretation of utilitarianism is to be preferred to actualistic versions. One of his reasons is that "[…] any adequate account of rightness must allow us to assess actions in the real world […]" (Gruzalski, 1981, p. 166) My discussion in chapter III seem to cast considerable doubt on the possibility, even for a reasonable probabilistic interpretation of AU, to meet this requirement. Also cf. Smith (1986) where she claims that "A principle's capacity to serve as a decision-making guide affects its acceptability as a theoretical account of rightness." (Smith, 1986, p. 342) She believes that if a principle cannot be used for making decisions, it cannot be used as a criterion of rightness either. (Cf. Ibid)

[11] Harman (1977) p. 4.

Proponents of AU are wise to reject Harman's contention that we can observe moral rightness. There is a perfectly valid explanation for why rightness and wrongness cannot be observed, according to AU. That we cannot observe the wrongness of the hoodlum's action is due to the fact that we cannot observe the total outcome of their action[12]. This is why "normative" observations cannot justify a belief in AU. According to AU, because the rightness and wrongness of particular actions depends on their outcomes, and because these outcomes are, presumably, far reaching, together with the fact that rightness and wrongness depends on what were the alternatives to the particular action, it is not very plausible to take rightness and wrongness to be observational properties. Rightness is better construed of as a theoretically derivative property. It is based on a generalisation from our judgements of value in particular cases, actual or imaginary. Or so I will argue below.

There are more indirect considerations that could play a part in justifying a belief that a particular action is morally right. There are perhaps other ways to gain justified beliefs about the optimificness of a particular action than direct observation?

Folke Tersman has argued that we cannot have justified beliefs about the optimificness of a particular action and that this means that AU cannot be justified given the idea of reflective equilibrium[13]. His argument has received criticism from Tännsjö[14]. I do not think Tännsjö's criticism is altogether successful. A stronger case for AU can be made. However, this question seems worthy of closer scrutiny. In the following I will state Tersman's case against the possibility of justifying AU within a coherentistic setting. I will then state Tännsjö's attempt to salvage AU from Tersman's attack. My own view on this issue is that the solution to this problem lies in shifting focus, from norms to values. I outline a possible way of justifying AU in section 5.0.

In Tersman (1991) the thesis that it is highly unlikely that we are ever justified in our belief that a particular action is right according to utilitarianism (AU) is defended. One premise of this argument is the claim that, in practice, we are unable to justify a belief to the effect that a particular action is optimific. The question is not whether we can *observe* that an action is optimific or not, but rather whether we are ever *justified in believing* that a particular action is optimific. In other words, to test AU against the assumption that an action is optimific we have to be justified in our beliefs regarding, for example, relevant alternatives, the outcomes of actions and the value of these outcomes. But we are not justified in these beliefs[15].

---

[12] Cf. Tännsjö (1990) p. 80-81.
[13] Cf. Tersman (1991).
[14] Cf. Tännsjö (1998).
[15] "[…] a person P, in order to be justified in accepting utilitarianism, must be *justified* in believing that the actions he judges to be right have at least as good consequences as any of their alternatives. This is not a casual matter. By "consequences" is meant the consequences of the actions with respect to the total sum of pleasure and pain of all sentient beings in the entire universe, from the time of their performance until the end

The underlying idea seems to be the following. We test the plausibility of our moral theories against our judgements in concrete cases. A theory can only gain inductive support from this procedure if it implies judgements about the concrete cases. These judgements are implied by the theory only relative to certain auxiliary hypotheses concerning complicated empirical comparisons between outcomes of different alternative actions. The problem, relative to AU, in this context is that we are never justified in believing that these particular auxiliary hypotheses needed for the inference hold true in concrete cases.

A direct form of justification[16] would be possible if we could observe optimificness in actions. This is not possible when it comes to actual particular actions in actual situations of choice. This way of confirming AU is a dead end. This direct way of confirming AU is rejected by Tännsjö, as well as by Tersman. The reason is that we can probably never "be justified in our observation that a particular action is optimific." (Tännsjö, 1998, p. 23)

However, Tännsjö has argued that we could still be justified in a belief in utilitarianism (AU). It is just that our justification is of a more indirect form:

> [I]f utilitarianism is correct, then, at least for a person who wants to be moral, it would be rational in many cases to try to maximise expected happiness. If utilitarianism is correct, to try, in the circumstances, to maximise expected happiness would be to act 'subjectively rightly'. Now, in a particular case, *whether an action maximises expected happiness or not does seem to be open to inspection. We may hold a justified belief about this*. If, in a particular case, we realise that a particular action does maximise expected happiness, and if, in the same situation, we form the immediate judgement, without any conscious reasoning having taken place, that, from a moral point of view, this action is responsible, and if we conclude that it is responsible because it maximises expected happiness, this may constitute evidence, directly, for the method; and, *indirectly for the utilitarian formula*. That formula explains why it is responsible to maximise expected happiness (in these circumstances, for a person who wants to behave morally), i.e., why doing so would be subjectively right. (Ibid. p. 24. My italics.)

---

of time. It seems obvious that in most, if not all cases, it is extremely unlikely that anyone has justified beliefs about such matters." (Tersman, 1991, p. 400)

[16] "It is tempting to argue, then, that when in a particular case we observe that an action is right and optimific (Oa&Ra), this may constitute evidence for utilitarianism. Indeed, Oa&Ra is evidence for utilitarianism if we find that the utilitarian formula is not only consistent with our observation in the particular case, but provides also the best (moral) explanation of it. This case constitutes evidence for utilitarianism if it is because the action in question is optimific that it is right. If, on the other hand, we face a situation where an action that is found to be optimific is not right (Oa&-Ra), then this disconfirms utilitarianism. If we find an action that is optimific and also right (Oa&Ra), but where we find a better explanation of its rightness (than its being optimific), then this case does not confirm utilitarianism either. If this method is to the point, utilitarianism does allow for a successful application of the method of coherentism." (Tännsjö, 1998, p. 22)

Is this a plausible way of justifying a belief in AU? There are major problems with this justification. Tännsjö thinks that we, at least sometimes, can have justified beliefs about whether an action maximises expected happiness or not. That this is 'open to inspection'. As I argued in chapter III the validity as well as the practicability of the reasonable subjectivist version of the deliberative approach is doubtful. Normal human agents are hard pressed to justify a belief to the effect that by adhering to this method of decision, the agent approximates the overall goal of AU. That it is 'open to inspection' whether or not an action maximises expected utility is highly doubtful given the criteria of relevance on the input that this method sets up.

Tännsjö suggests another indirect applications of the method of coherentism to AU:

> [I]t seems plausible to assume that we can sometimes be justified in our belief that a certain action is not optimific (to boil a certain child in oil for no particular reason, say). We can also be justified in our belief that, from a hedonistic point of view, the world would be better if we avoided this action (no matter how exactly we were to do this). We have good reasons to believe that, from a hedonistic point of view, this action would make the world (in one respect) worse and no reason to believe that it would make the world (in any respect) better. We can also have a firm normative belief that even if the action we were to perform, if we avoided boiling this child in oil, would be wrong, it would be 'less' wrong than the action we would perform if we did boil this child in oil. Classical hedonistic utilitarianism explains why this is so. (Ibid. p. 23-24)

This is supposed to be an example of an actual, not imaginary, case. So we have to consider this kind of example in practice. Otherwise our judgement cannot be caused by its being true[17]. We must set ourselves in that kind of situation in practice in order to justify a belief in AU. Perhaps Tännsjö is right about the possibility of having these kinds of justified beliefs. His case is not as strong as one could hope, though. This defence rests on the assumption that we are justified in believing, of a particular action, that it is not optimific. But even believing that an action is *not* optimific is quite hard to justify. As the arguments of chapter II showed, we must have justified beliefs about complicated empirical comparisons as well as justified beliefs about the relevant alternatives in the situation of choice. The child-boiling example also involves (implicitly) the notion of probabilities. This is, I think, unfortunate.

Of course, other normative theories can also be taken to explain why this is wrong; AU needs to be the best explanation if it is to get any support from its ability to explain the wrongfulness of this particular action. Let us grant that AU is the best explanation, however. Does this line of reasoning help justify a belief in AU? Tännsjö is, I think, on to something useful here. The clue to a

[17] Cf. the next section (4.1).

more solid defence lies in the idea of degrees of wrongness. The case for AU can be made even stronger. There are cases which do not involve probability assessments and in which we, instead of relying on somewhat complicated empirical beliefs, can base our defence of AU directly on judgements of intrinsic value instead. This defence rests on a *comparative* idea, i.e. we rank different states of affairs on a scale of 'better than' and exploit the transitivity of this relation in order to justify a belief in AU. (Cf. 5.0 below.)

4.1 Imaginary Cases

What about testing normative theories in thought experiments? Perhaps it is sufficient if the implications of a particular criterion of rightness could be tested in hypothetical thought experiments, and that the answers to our fictive moral questions could be compared to each other and teach us something about which of the theories is the most plausible? This is surely regarded by many to be a plausible procedure when it comes to evaluating theories. This procedure has been the dominating way of ranking how well theories fare relative to one another in the discussion amongst professional ethicists.

This way of investigating the reasonableness of normative theories goes roughly as follows. You construct a hypothetical situation of choice which has certain features. You then ask what the theory would recommend as the best solution to this particular problem of choice, and what the competing theory would recommend. You can then compare these competing answers and try to figure out in what relation they stand to your own 'intuitions', or considered moral judgements regarding what would be the right answer to this problem of choice. The theory that provides the most plausible answer is then to be preferred.

As I mentioned above, imaginary cases seem to play an important role in our assessments of moral theories. Here, we just hypothesise that a certain action is optimific. Just like laboratory experiments, thought experiments give us the advantage of being able to disregard irrelevant factors as well as manipulating different variables one at the time. Many crucial observations, the ones we rely on when we refute or inductively support our theories, are made in laboratory settings. Here the influence of irrelevant factors can be reduced to a minimum and sometimes be dispensed with all together. This last point is crucial. In ethics we can assume that we know what action is optimific, and ask ourselves if it is also right according to our moral intuitions.

Tersman argues, however, that resting a justification of AU solely on imagined cases is not sufficient. Why is it not sufficient for a vindication of AU, to rest purely on imaginary cases? Tersman's answer is that

> [O]ur evidence for moral principles cannot reasonably consist solely of judgements on imagined cases in order for us to be justified in accepting

> the principles. There is a good reason why more than thought experiments are required in other possible areas of knowledge, e.g., the natural sciences. This is that judgements about only hypothetical cases–unlike actual observations–never are caused by their being true. (Tersman, 1991, p. 404)

In thought experiments we do not 'observe' the optimificness of actions, we *assume* that an action is optimific and ask ourselves if it is also right. Thus I concur in Tersman's demand that observations need to be caused by them being true. Tersman has argued that utilitarianism cannot be justified via a reflective equilibrium and that this separates utilitarianism from other propositions, e.g. some empirical propositions in the natural sciences. This argument raises the question of what we should require from a justification, and how questions of justification should be settled. I believe, nevertheless, that thought experiments, just like laboratory experiments, have a very important role to play. We can view imaginary, hypothetical cases in the same way as we view carefully controlled experiments in laboratories. There are a lot of *as ifs* in the natural sciences too. In different experiments you make assumptions that contains *as if* clauses. You test the hypothesis as if the laboratory environment were the same as the world *out there*.

Imaginary cases can be part of an indirect reasoning concerning the truth of AU's criterion of rightness. Tersman is right in his claim that we cannot rely solely on imaginary cases, but this does not preclude them from playing an important role. They provide us with *some* support for, or against, ethical theories.

5.0 Scalar morality

Both the direct and the indirect ways of justifying AU, discussed in sections 4.0 and 4.1, deal with the notions of 'optimificness', 'rightness' and 'wrongness'. Because of this it is not surprising that they fail. I think that there is an alternative way of proceeding that is more successful. In describing this alternative way, I will exploit the difference between norms and values. I will be outlining a way in which 'evaluative cognitivism' might be justified. 'Evaluative cognitivism' is the idea that it is possible to justify a belief to the effect that AU's way of ranking situations in terms of the amount of intrinsic value realized in them is true. The idea is that AU might emerge as part of the best explanation of judgements to the effect that one state of affairs is better than another, judgements for which we may have independent grounds (in that these judgements are issued in direct response to a situation without any conscious reasoning going on). This idea underlies a possible (speculative) defence of AU, a defence that is rather indirect. It should be emphasised that I take this to be a *possible outline* for justifying a belief in AU. I am not claiming that I have justified a belief in AU. I will only argue that practical scepticism *need* not imply evaluative scepticism.

AU can be conceived of as a normative theory, providing a criterion of rightness of actions. This is how I have conceived of AU until now. But AU can also be conceived of as an evaluative theory, a theory that ranks items or states of affairs as better or worse. Although I will not try to give an account of a complete moral theory, I will now extend my formulation of AU to include also its evaluative part. I will take AU to consist of the following five principles:

1. An account of positive intrinsic[18] value.
2. An account of negative intrinsic value.
3. An account of how 1 and 2 determines a judgement, all things considered, of the overall intrinsic value of a state of affairs. (The maximisation thesis)
4. An account of instrumental moral value of actions.
5. A translation rule that allows us to translate judgements of intrinsic overall value into judgements of instrumental moral value of actions.

I will not elaborate 1-3. Because these parts of AU are determined by the favoured specific theory of intrinsic value that one accepts, e.g. hedonism or preferentialism, I will assume that they can be spelled out in one way or another. My interest lies with 4 and 5. In what follows I will, however, use a hedonistic understanding of intrinsic value, in order to illustrate my line of argument.[19] Notice again that my aim is merely to show how it might be *possible* to justify a belief in AU.

I will be arguing two main points:

1) That AU, conceived of as an evaluative theory, is compatible with AU's criterion of rightness. (This will be done by formulating an evaluative counterpart to AU's criterion of rightness that is extensionally equivalent with it.)
2) That this way of conceiving of AU makes it possible to provide evidence in support of AU.

Let us start with 1. Consider point 4, in my more extensive statement of AU. We have to conceive of AU, not only as a criterion of rightness of action, but also as a theory of *instrumental moral value* of actions. In *Utilitarianism* Mill states the utilitarian ideal in the following way:

[18] Throughout, I will talk about 'intrinsic' value. This is because the writers I discuss uses this term. It should always be kept in mind however that the value I have in mind is also 'final'.

[19] Tännsjö has defended the hedonistic account of intrinsic value, using particular observations of intrinsic value to infer, by means of the best explanation, that we have reason to believe in hedonism. Cf. Tännsjö (1990).

> [T]he Greatest Happiness Principle, holds that actions are right *in proportion as they tend* to promote happiness, wrong as they *tend to promote* the reverse of happiness. (Mill, 1987, p.278. My italics.)

If we take Mill by his words, the most natural interpretation of this passage is that he is putting forward an (exhaustive) account of right- and wrong-making *characteristics* of actions, not a *criterion* of rightness of actions[20]. As already pointed out by Roger Crisp[21], this reading makes 'rightness' a matter of degree[22].

Usually, AU is formulated as a non-gradual criterion of right action. What, then, should we say about Mill's formulation? My impression is that the difference between the "standard" view and the one put forward by Mill is often ignored[23]. Instead of ignoring it, we should exploit it. I think that we can learn something by taking the difference seriously. I think Mill's formulation hints at the fundamental or basic idea of AU, i.e. that the important thing is not to perform right actions, e.g. as many as possible, but to realize as much intrinsic value as possible.

Michael Slote has discussed this topic when he presented his 'satisficing act-consequentialism'. According to this theory an action need not be optimific in order for it to be right. It suffices that the action is *good enough*, where 'good enough' is not given any precise definition, but is said to be a function of the degree in which it's consequences approximates the goodness of the consequences of the optimific action in the situation[24].

> The choice between satisficing and optimizing act-consequentialism would then be seen as involving a somewhat arbitrary decision about when the consequences of acts were good enough to qualify them as morally right, but by the same token any choice between differently stringent forms of satisficing act-consequentialism would also have to be seen as arbitrary. No dividing line between right and wrong action would be thought of as corresponding to anything objectively valid, and the desire to present a non-arbitrary consequentialist theory of the morality

[20] One could of course conceive of this passage as stating a criterion albeit a graded one. What we *call* it is of no importance however.

[21] Crisp, R, *Mill on Utilitarianism*, 1997, p. 96.

[22] Under "standard" formulations of utilitarianism, rightness is an 'all or nothing' affair. Of course, Mill's formulation can be supplemented by a principle to the effect that the action that promotes happiness to a maximum degree is *the* right action. This can be taken to constitute the criterion of right action. Tännsjö also exploits this idea: "We can also have a firm normative belief that even if the action we were to perform, if we avoided boiling this child in oil, would be wrong, it would be *'less' wrong* than the action we would perform if we did boil this child in oil. Classical hedonistic utilitarianism explains why this is so. At least it does if, as we should do, we add to the criterion of rightness an idea that what is stated in the criterion is an ideal that could be reached *more or less perfectly*. " (Tännsjö, 1998, p. 24. My italics.) The last sentence bears a striking resemblance to Mill's formulation.

[23] One exception is Slote (1985).

[24] "[…]it may well be possible formally to elaborate the notion of enough-ness as some sort of percentage or other mathematical function of the best results attainable by the agent. I shall not attempt to spell out here the details of any particular plausible way in which this might be attempted." Slote (1985), p. 52.

> of actions would presumably then lead one to espouse a scalar form of act-consequentialism according to which an act *a* counts as morally better than an alternative *b* just in case *a* has better consequences than *b*. (Slote, 1985, p. 80)

I will not discuss the plausibility of Slote's 'saticficing act-consequentialism'. My focus is AU. However, the idea of a *scalar* moral theory is a useful one relative to my purposes in this chapter. I give the following account of the instrumental moral value of actions, thus meeting point 4 in the above list of the five principles of AU:

> IMB: An action, A, is (instrumentally) morally better than an alternative action, A', if and only if, the universe contains a greater balance of positive over negative intrinsic value if A is performed than if A' is performed.

This definition is stated in comparative terms. It allows us to rank all actions, in theory, on a (transitive) 'better than' scale. AU, formulated as a scalar moral theory, says that the right action is the (instrumentally) morally optimal action. We are now in a position to give a criterion of 'the instrumentally morally optimal action'.

Instead of stating a criterion of rightness of action, AU can be taken to state an extensionally equivalent statement in terms of instrumentally morally optimal actions:

> IMO: An action is (instrumentally) morally optimal, if and only if, there was nothing the agent could have done instead in the situation, such that had the agent done this, the universe (on the whole) would have been better[25].

This formulation is the evaluative counterpart of the normative criterion of rightness. This criterion should be taken as implying that the more intrinsic value an action realizes, the stronger our reason to perform it. As Kagan said in his *The Limits of Morality*, "the reason to promote the good […] is a *pro tanto* reason."[26] Thus, we have the strongest reason to perform the morally optimal action.

The translation-rule, in point 5 of my list, says that we can translate talk of 'rightness' into talk of 'instrumental moral optimality'. If we assume that an action ought to be performed if, and only if, it is instrumentally morally optimal, we have arrived at the equivalent of AU's criterion of rightness. Accordingly, the right thing to do, is also the 'instrumentally morally optimal'

[25] I have accepted a hedonistic understanding of intrinsic value here, so 'better' is to be understood as containing a greater sum balance of wellbeing (pleasure) over ill-being (suffering).
[26] Kagan (1989), p. 17.

thing to do[27]. 'Optimal'[28] should here be understood in terms of instrumental moral value. In fact, I think Tännsjö was right when he claimed that normative problems are only evaluative problems in disguise[29].

Let us turn to 2. As it stands IMO evaluate actions. But AU can also evaluate other things. It can evaluate states of mind in terms of intrinsic value that they contain:

> SOM: A state of mind, S, is intrinsically better than a state of mind, S', if and only if, S contains a greater net sum of positive intrinsic value over negative intrinsic value than does S'.

This principle follows from the evaluative part of AU. Just as we have to deduce testable hypotheses from scientific theories in the natural sciences, we can do the same thing with AU. It seems that it is *possible* to provide evidence for believing this statement to be true. And it seems that this evidence is much more credible than the evidence presented in sections 4.0 and 4.1. SOM can be supported by independent judgements of intrinsic value in a cool hour. Let me illustrate how this might be done.

Right now I am sitting by my desk, writing on this essay. I reach out and take a sip of my newly brewed coffee. But, unfortunately, I forgot that the coffee is extremely hot. As a result my mouth is sour. I have a rather unpleasant experience of pain. Let us call the state of mind I was in before sipping on the coffee S1 and the state of mind I was in just after my sip S2. I now make, without any conscious reasoning going on, the following judgement of intrinsic value: "S1 is (was) intrinsically better than S2".

Can AU be the best explanation of this evaluative judgement? We should distinguish between (i) the best explanation of my *passing* the judgement and (ii) the best explanation of the *content* of my judgement. In order for AU to get support from this the best explanation of (i) must be provided by the content of my judgement[30]. And the best explanation of the content of my judgement must be AU. Let us assume that the best explanation of my passing the judgement is that S1 *does* contain a greater sum balance of intrinsic value than does S2. Can AU be the best explanation of this, i.e. of the content of my judgement? Can AU explain why S1 is intrinsically better than S2?

---

[27] Cf. Crisp (1997), p. 96.

[28] This translation requires, to be plausible, that 'best' be understood so that it does not preclude two alternatives or states to be 'best'. This, however, does not seem to be a problem. Under the value-interpretation, a 'best' action available to the agent will be a 'right' action for her. In other words, 'the best action' will be extensionally equivalent with 'the right action'.

[29] Tännsjö (1990) p. 91. He now seems to have changed his mind about this issue.

[30] Cf. Tännsjö (1990) p. 79.

AU's explanation of S1 being intrinsically better than S2, is that S1 contains more utility (pleasure over pain, or wellbeing over ill-being) than does S2. Let us, for the sake of the argument, assume that AU provides the best explanation of the content of my judgement[31].

Return once again to my coffee drinking. The sourness of my mouth slowly fades away. Over a period of 15 minutes I continue to pass judgements on the intrinsic value of my successive state of mind. "S3 is intrinsically better than S2", "S4 is intrinsically better than S3" and so on. As the pain goes away, things are getting better. I have assumed a hedonistic understanding of intrinsic value. A basic idea of AU is that the more intrinsic value realized, the better. AU is a maximising theory. I think that my judgements during the 15 minutes lend some support to AU. The fact that I pass these judgements, without any conscious *reasoning* going on, seem to support the idea that the less suffering, and the more pleasure, the better.

Suppose that we can, at least sometimes, e.g. in a cool hour, be independently justified in judging one state of mind intrinsically better that another. To the extent that these judgements are best explained by AU, this can be taken to constitute evidence for AU. Now, could this be made part of a justification for believing that AU's criterion of rightness of actions is true? If my judgements are such that every time that I judge a state of mind, S, to be better than a state of mind, S', it also holds that S contains a greater ratio of pleasure over suffering than does S', then my judgements are consistent with AU's maximising principle. Because SOM follows from AU's evaluative part, providing evidence for SOM will also provide evidence, indirectly, for AU's evaluative part. And because this part can be formulated in a way that coincides with AU's normative formulation, this can also be taken to support AU's criterion of rightness.

6.0 Concluding remarks

It is time to sum up the conclusions of this chapter. AU does not sit that well with some constructivist accounts of the grounds of morality. This has to do with AU's problem of guiding actions in practical situations of choice. Because AU's criterion of rightness of actions cannot guide our actions in practice, AU does not live up to the very purpose of constructing a morality, i.e. helping us to coordinate our actions and moderate our selfish natural impulses in a way that is to everyone's advantage. The point of morality, on

[31] I do not pretend to show that I, or that somebody else is for that matter, is justified in believing that AU is true. In order to show this, I would have to show that AU is in fact the best explanation of (most of) my, or our, judgements of intrinsic value, in imaginary cases as well as in actual cases. This is obviously a task of considerable difficulty. My task is more modest. I want to show that although AU leads to practical scepticism, this does not show that AU is false. And, furthermore, it does show that a belief in AU can be tested against relevant evidence, i.e. our considered judgements. I point to a *possible* way of justifying a belief that AU is true. Therefore, I need not show that rivalling theories, such as satisficing utilitarianism or 'prioritarianism', provides a better explanation of our considered judgements than does AU.

this account, is to have effective methods of moral decision-making. An abstract criterion of rightness is, according to this view, merely a *chimera*.

But, on a realist account of moral judgements, the versions of constructivism that appeals to the practical 'function' of morality in order to establish its putative truth, are the real *chimeras*. I have argued that the direct attempt to justify AU's criterion of rightness of actions, suggested by Tännsjö, along the lines of a coherentistic, or reflective equilibrium, idea of justification fails. A more indirect justification of AU's main ideas is however possible, or at least this is what I have argued. The suggested way of justifying a belief that AU is true, presented in this section, was based on an inference to the best explanation of some particular considered judgements of intrinsic value.

7.0 Summary

Most of the discussion on utilitarianism has focused on formulating the most theoretically satisfying version of the theory. The effort to formulate and refine ethical theories in order for them to stand up to challenges from hypothetical thought-experiments and to meet the theoretical desiderata of simplicity, generality and universality has created a gulf between theory and practice. This fact is in no way surprising. In a way, it is not even unfortunate. What is unfortunate, however, is that the gulf is sometimes not even noticed. It is not unusual to see ethical theories being referred to in defence of different positions regarding practical moral problems. This is sometimes quite legitimate. But sometimes the defences are not very plausible. Utilitarianism is often used when defending different standpoints in debates on, e.g. abortion and euthanasia. Which standpoints on these issues does utilitarianism support? Well, as I hope to have shown in this thesis, this is not an easy question to answer. In order to answer this question we would have to have a sufficiently practicable and valid method of decision for AU. This, unfortunately, we do not have.

It has been claimed that the question of whether or not AU is applicable in practical situations of choice is not a philosophical question. This is perhaps true. *But*, and this is an important rider, the practical applicability or inapplicability of ethical theories can be taken to have different implications for issues that are philosophical.

It is time to sum up the conclusions of the foregoing chapters. This thesis has been an investigation into the subject of act-utilitarianism and the problem of practical action-guidance. In chapter II, I provided the stage of the discussion to come. I defined AU, and argued in favour of upholding a sharp distinction between criterion of rightness of actions and methods of moral decision-making. The desiderata of *practicability* and *validity* was presented and defended. In chapter III, the principle of maximising expected utility was examined. More particularly, it was examined as a part of a method of

decision for AU. I introduced the distinction between the *unrestricted deliberative strategy* and the *restricted deliberative strategy*. It was argued that no proponent of AU as a criterion of rightness of actions has defended the unrestricted deliberative strategy as a method of decision for AU. In chapter IV, some suggestions of methods of decision-making for AU defended by thinkers in the utilitarian tradition were examined. The general question of the place of secondary rules within a method of decision for AU was discussed. Although different utilitarian thinkers have defended different suggestions on this question they all chare roughly the same basic structure. It was argued that defending a particular set of secondary rules is very hard. The reasons for favouring one particular set above another are weak and we are hard pressed to justify them.

Is AU a plausible ethical theory? In order to answer this question we need to supplement AU with a theory of what is intrinsically valuable. AU is more of a *structure* than a full-fledged ethical theory. Different versions of preferentialism and hedonism are theories of value that can give content to AU. It is the structure of AU together with the nature of the good things that leads to Practical scepticism. Because AU prescribes maximisation of intrinsic value *and* because of the fact that this value is scattered over vast distances in time and space, (and therefore very hard to determine the exact quantity of) determining whether or not the outcome of an action satisfies AU's criterion of rightness is practically impossible for human beings.

Appendix

Do other normative theories face similar problems regarding practical action-guidance as does AU? If they do, the relevance of this for the plausibility of these theories depends, just as for AU, on how we conceive of the foundations of morality. Again, if realism is true then the lack of practical action-guidance constitutes no problem for normative theories. But if one would defend some versions of moral constructivism, stressing the practicability of normative theories, then many other theories, besides AU, face the same *kind* of problem. For a moral constructivist, of the sort I have been discussing in the previous chapter, it is dubious to take AU's lack of practical action-guidance to constitute a very strong reason to prefer these other approaches to AU.

Several well-respected normative theories either presents criteria of rightness that are hard to apply in practical situation of choice, or they contain clauses which, in order for us to be justified in our beliefs that they are satisfied, we would need much more information than we are in a position to get. I will discuss these problems with regard to one version of virtue ethics, Kant's normative theory, Robert Nozick's theory of entitlement and John Rawls theory of justice of identifying the worst-off group. My tentative conclusion is that these theories are also hard pressed to provide us with practical action-guidance, although the problem might well be less radical for (some) of these theories than it is for AU. But it transpires that the problem is one of degrees, not of kind.

1.0 Practical scepticism and other normative theories

In Frazier (1994), Robert L. Frazier presents what he calls "A revised impracticality argument", where he argues that either is AU false or we are justified in believing in Practical scepticism:

> (1) Act utilitarianism cannot have a practicable, validated ethical decision procedure.
> (2) If act utilitarianism cannot have a practicable, validated ethical decision procedure, then we cannot be justified in believing of any act that it satisfies act utilitarianism's conditions for moral permissibility.
> (3) If we cannot be justified in believing of any act that it satisfies act utilitarianism's conditions for moral permissibility, then either act utilitarianism is not true, or we cannot be justified in believing of any act that performing it is morally permissible.
> (4) Therefore, either act utilitarianism is not true, or we cannot be justified in believing of any act that performing it is morally permissible.
> (Frazier, 1994, p. 43-44)

The upshot of the arguments of chapters II-IV is that we are justified in believing that (1) is true. If (2) follows from (1), then we are justified in believing that Practical scepticism is true.

Premise (3) of Frazier's argument can be taken as an instance of a more general principle:

> (3') If we cannot be justified in believing of any act that it satisfies Theory T's conditions for moral permissibility, then either Theory T is not true, or we cannot be justified in believing of any act that performing it is morally permissible.

What about other normative theories, then? Is AU special in that it leads to practical scepticism? Even if other theories do not entail practical scepticism, they are often hard pressed in providing practical action-guidance. It is often *very* hard to justify a belief to the effect that one has satisfied the theory's criteria of permissibility of rightness. That AU exhibits this flaw is not unique. *If* this flaw is considered to constitute a telling criticism of AU, then it is also a telling objection to other moral theories of a generalist kind. Let us turn, now, to a discussion of (3') in relation to other ethical theories.

1.1 Virtue-ethics

Do virtue-ethics provide us with practical action-guidance? Suppose virtue-ethics could be formulated as a criterion of rightness of actions:

> An action is right iff[1] it is what a virtuous agent would do in the circumstances[2]. (Hursthouse, 1991, p. 225)

This criterion of rightness must be supplemented by a method of decision. How should we go about making our decisions if we want to satisfy this criterion? In order for us to satisfy this criterion intentionally, we would need to answer several difficult questions. One would have to have justified beliefs about *who* is virtuous, presumably some actual individual[3]. To be justified in a belief about the virtuousness of an agent seems to require a great deal of information with regard to the person's character-traits. Furthermore, we would need to be justified in our hypothesis about what this agent *would do in the circumstances*. Another complication is, perhaps, that if different virtuous

[1] This is an abbreviation for 'if, and only if'.

[2] Also Cf. Oakley (1996). He writes: "The first and perhaps best-known claim, which is central to any form of virtue ethics, is the following: (a) An action is right if and only if it is what an agent with a virtuous character would do in the circumstances. […] A right action is one that is in accordance with what a virtuous person would do in the circumstances, and what *makes* the action right is that it is what a person with a virtuous character would do here."(Oakley, 1996, p. 129-130)

[3] Of course, we cannot define a virtuous agent as the agent who performs right actions. This would be to give a circular definition. A more plausible version of virtue ethics would claim that the virtuous agent can determine which properties of a situation is of moral import. It is these properties that determine rightness, not the fact that a virtuous agent would have chosen it.

persons did different things in the same situation, we would have the problem of choosing which person to follow.

Virtue-ethics, at least if it is interpreted in this way, seems to be troubled by problems similar to those which trouble AU. If we take the doctrine to state a criterion of right action, i.e. that an action is right if, and only if, it is the action a virtuous person would have performed in the situation, much of the same problems, due to our cognitive limitations, arise for this theory. Finding *examples of virtuous persons* is no easy task. The degree to which a person is virtuous seems open to controversy, different reasonable and relatively well-informed persons will hold different persons to be virtuous in different degrees. When I try to come up with an example of a virtuous person I always fail[4].

Finding out what a virtuous agent would do is not any easier, it seems. Finding out what the other person would have done, perhaps being a non-virtuous person (or a virtuous one!?) myself, involves complex processes of calculation, imagination and conjectures. Part of these problems lies in determining the proper weight between virtues such as courage and temperance. The hypothesis that this version of virtue ethics may well lead to practical scepticism seems rather plausible. I will not, however, try to argue in favour of this point any further. So this hypothesis remains highly speculative.

1.2 Kant's theory

What about deontological theories, then? I will take Kant's theory as an example of a deontological theory. Because he provides a ground, or reason, for the duties in his system and not just a list of forbidden, permissible and obligatory actions, problem much like those arising for AU due to the distinction between criterion of rightness and method of decision, arises also with this theory. Does Kantianism imply practical scepticism? Kant maintained that it is impossible to know 'with complete certainty' whether a particular action has moral *dignity*, or moral *worth*:

> In actual fact it is absolutely impossible for experience to establish with complete certainty a single case in which the maxim of an action in other respects right has rested solely on moral grounds and on the thought of one's duty. It is indeed at times the case that after the keenest self-examination we find nothing that without the moral motive of duty could have been strong enough to move us to this or that good action and to so great a sacrifice; but we cannot infer from this with certainty that it is not some secret impulse of self-love which has actually, under the mere show of the Idea of duty, been the cause genuinely determining our will. (Kant, 1948, p.71-72)

[4] I have tried to come up with an example several times, never feeling satisfied. There are always flaws in the person's personality that seems to exclude the person from being virtuous even to an extent that approximates the optimal degree.

Even the possibility of knowing whether an action is *right*, according to the categorical imperative, is not an easy thing[5]. For an agent to have a justified belief regarding whether an action, which the agent contemplates whether to perform it or not, satisfies the categorical imperative, is very hard. Consider the categorical imperative, in one of its formulations:

> Act only on that maxim through which you can at the same time will that it should become a universal law. (Kant, 1948, p. 84)

When is an agent justified in her belief that her maxim satisfies this criterion? How specific is the maxim supposed to be? What would the world be like if a particular maxim had become a universal law? These are very hard questions to answer. Perhaps it is relatively easy to determine whether some, particularly hideous actions, are morally forbidden. The torturing of innocent persons for no particular reason, say. But this does not take us very far. How much help is required from us in assisting people in distress? Can I tell a lie in order to save the lives of one hundred persons?

There is also the problem of the relation between actions and maxims. Because it is the maxims that are the morally relevant things, they are either compatible or incompatible with the categorical imperative and because the relationship between a maxim and an action is of a many-many kind, there is a problem regarding the moral status of actions. An action can be performed on different maxims and a maxim can give rise to different actions. We need relevant act descriptions and a criterion of relevance of maxims. For example, I might act on the maxim "To save the person in the water from drowning" by tossing in a life-buoy, jumping into the water, push another person into the water (in order for the drowning person to have something to cling to). I might even shoot the person in the water (this saves her from *drowning*)!

Another problem with applying Kant's theory is the problem of conflicting duties. Suppose that I promise you that I will commit suicide tomorrow at noon. (For the sake of argument I shall interpret Kant's theory as implying that breaking promises and committing suicide is always wrong.) The time of my promised suicide is closing. What shall I do? Kant would be hard pressed to give an answer here.

Again, what method of decision should an agent use in order to satisfy this criterion of rightness? Different methods have been proposed, and Kant himself seems to have thought that he could formulate a body of secondary rules justified by the categorical imperative. In Kant's case, the trouble arises out of his particular way of grounding the duties. The problem does not arise for every deontological system. A deontology that contains only one rule, a prohibition on lying say, is highly practicable. The problem is, of course, that

[5] Cf. Nell (1975).

we would like a reason for following this set of duties. Not just any set of rules would do, it seems. Kant's theory is a plausible moral theory partly because it has a foundation (the categorical imperative). A mere list of duties lacks this foundation. It is the fact that a theory has a foundation that gives rise to this particular kind of practicability problem. It seems that practical scepticism is a possible consequence of Kant's ethics as well.

1.3 Nozick's 'principle of rectification'

Robert Nozick presents his libertarian normative theory in *Anarchy, State and Utopia*. The question is whether this theory can be taken to provide practical action-guidance. I will concentrate on Nozick's *entitlement theory*. This theory has three main principles:

1. A person who acquires a holding in accordance with the principle of justice in acquisition is entitled to that holding.
2. A person who acquires a holding in accordance with the principle of justice in transfer, from someone else entitled to the holding, is entitled to the holding.
3. No one is entitled to a holding except by (repeated) applications of 1 and 2. (Nozick, 1974, p. 151)

How does Nozick's theory give rise to practical scepticism? In order to determine people's entitlements it is necessary to determine whether or not peoples present holdings rests on historical injustices. Nozick:

> The existence of past injustice (previous violations of the first two principles of justice in holdings) raises the third major topic under justice in holdings: the rectification of injustice in holdings. If past injustice has shaped present holdings in various ways, some identifiable and some not, what now, if anything, ought to be done to rectify these injustices? […] How far back must one go in wiping clean the historical slate of injustices? […] I do not know of a thorough or theoretically sophisticated treatment of such issues. Idealizing greatly, let us suppose theoretical investigation will produce a principle of rectification. This principle uses historical information about previous situations and injustices done in them (as defined by the first two principles of justice and rights against interference), and information about the actual course of events that flowed from these injustices, until the present, and it yields a description (or descriptions) of holdings in the society. The principle of rectification presumably will make use of its best estimate of subjunctive information about what would have occurred (or a probability distribution over what might have occurred, using the expected value) if the injustice had not taken place. If the actual description of holdings turns out not to be one of the descriptions yielded by the principle, then one of the descriptions yielded must be realised. (Nozick, 1974, p. 152-153)

Obviously, the *practical* difficulties in determining people's entitlements are considerable. Without *extensive* historical information, as well as *justified* beliefs about alternative world histories, we cannot determine people's entitlements. And before we can do that we cannot determine what people can rightfully do with their present holdings. This problem cuts deep into Nozick's normative theory, because what one may rightfully do, depends to a large extent on what one is entitled to. If my great, great, great, great grandmother built her wealth on an unjust holding and my present holdings are a result of hers, then I am not entitled to my present holdings. I may not do what I want with it. But if my distant relative had not committed the injustice, history would (presumably) have taken another course. Perhaps there is a true answer to the question of which holdings I would possess in this possible history. Then these are the holdings to which I am entitled. But how do we determine this? Perhaps my distant relative would not have conceived her only child, at the time when she actually did, if it wasn't for the fact that she unjustly gained her holding. Then I would not even exist in the 'just' history, making the question over my entitlements rather queer.

Nozick hints as an *ad hoc*, solution to this problem involving appeals to considerations of equality to determine the entitlements where no definite answer can be given as to the 'real' entitlements of persons[6]. This, surely, does not square well with the original idea.

Satisfying the principle of rectification is a necessary condition for making Nozick's theory normatively reasonable. If past injustices were not rectified, these injustices would spill over on present entitlements making them unjust. For a historical theory of entitlements like Nozick's this would be devastating. It seems that we cannot justifiably believe that we are entitled to our present holdings. If we cannot know this we cannot know whether our present transfers are just either. Thus, it seems that Nozick's entitlement theory might well lead to practical scepticism.

1.4 Rawls: The greatest benefit for the worst-off

According to John Rawls, institutions should be designed so that their workings secure the *greatest* benefit for the *worst-off group*. Determining in practice whether or not there is an alternative set of institutions whose workings would secure a greater benefit for the worst-off group seems extremely hard. We can imagine that small modifications in the way institutions work, would have consequences for the benefit of the worse-off group. Providing good reasons for the claim that a certain institutional set is actually *maximising*[7] the life prospects of the worst-off group seems very hard.

---

[6] Ibid. p. 153.
[7] Cf. Rawls, 1971, p. 83.

Another practical difficulty stems from the fact that it seems difficult to determine which group of people that actually are worst-off.

> Here it seems impossible to avoid a certain arbitrariness. One possibility is to choose a particular social position, say that of the unskilled worker, and then to count as the least advantaged all those with the average income and wealth of this group, or less. The expectation of the lowest representative man is defined as the average taken over this whole class. Another alternative is a definition solely in terms of relative income and wealth with no reference to social position. Thus all persons with less than half of the median income and wealth may be taken as the least advantaged segment. (Rawls, 1971, p. 98)

Rawls own suggestions seem rather rough. We would like, in order to make the theory normatively plausible, to have good reasons for believing that these alternative ways of determining the group actually identifies the persons who possesses the smallest share of 'primary social goods', i.e. 'rights, liberties, opportunities and powers, income and wealth and a sense of one's own worth' (Cf. Rawls, 1971, p. 92). Accomplishing this in practice is surely very hard. Even Rawls' theory, it seems, might lead to practical scepticism.

2.0 Conclusion

I have argued that other important normative theories does not seem to fare much better with regard to action-guidance, than AU does. The (tentative) conclusion is that an adherent of AU is not in that much worse a position than an adherent of its most popular contestants, when it comes to the problem of practical action guidance. *If* there are any differences between the theories, it is surely a difference of degree rather than of kind. We have seen that alternative normative theories also exhibit this tension between theory and practice, between criterion of rightness of actions and moral decision-making. They are also difficult to apply in hard practical moral cases. Because of the fact that different candidates of methods of decision-making surface for every one of them and because it seems to be an empirical question as to which of the possible methods is the most effective means of realizing the end set by the different criteria of rightness (or of justice), our cognitive limitations and shortcomings as moral agents makes it hard for us to justify a belief to the effect that a particular method of decision actually is the best way of meeting the standards of these theories.

Furthermore, the difficulties of predicting future outcomes of actions, which troubles AU, affects every theory that places any weight on the value of these outcomes. Shelly Kagan has argued that the problem of practical action-guidance haunts every plausible ethical theory. This problem

> […] threatens not only consequentialism, but indeed all plausible normative theories. For if it is in fact impossible to get a grip on the consequences of an act, then this problem will be inherited by all theories that give this factor any weight at all and this will be virtually all theories. For […] all plausible theories agree that goodness of consequences is at least *one* factor relevant to the moral status of acts. (Kagan, 1998, p. 64)

So this problem is a general one. Where does *this* fact put us? It is certainly a discomforting fact and a rather depressing one too, one could think. For several plausible criteria of rightness, one cannot have sufficiently good reason to believe that one has succeeded in satisfying them even if one tries.

References

Audi, Robert (1995) (General Editor), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, 1995.

Austin, John (1832) *The Province of Jurisprudence Determined*, (ed. W. E. Rumble, Cambridge University Press, 1995.

Bales, R. E. (1971) "Act-Utilitarianism: Account of Right-making Characteristics or Decision-Making Procedure", *American Philosophical Quarterly* 8, 1971.

Bennett, Jonathan (1995) *The Act Itself*, Clarendon Press, Oxford, 1995.

Bentham, Jeremy and Mill, John S. (1987) *Utilitarianism an Other Essays*, (ed. Alan Ryan) Penguin Books, 1987.

Bergström, Lars (1966) *The Alternatives and Consequences of Actions*, Acta Universitatis Stockholmiensis, Stockholm Studies in Philosophy 4, 1966.

Bergström, Lars (1996) "Reflections on Consequentialism", *Theoria*, LXII, 1-2, 1996.

Brink, David O. (1986) "Utilitarian Morality and the Personal Point of View", *The Journal of Philosophy*, vol. 83, No. 8, 1986.

Carlson, Erik (1995) *Consequentialism Reconsidered*, Kluwer Academic Publishers, 1995.

Crisp, Roger (1997) *Mill on Utilitarianism*, Routledge, 1997.

Frankena, William K. (1988) "Hare on the Levels of Moral Thinking", in Seanor, Douglas and Fotion, N. *Hare and Critics, Essays on Moral Thinking*, Oxford University Press, p. 43-56, 1988.

Frazier, Robert L. (1994) "Act Utilitarianism and Decision Procedures", *Utilitas*, Vol. 6, No. 1, 1994.

Garner, R. and Rosen, B (1967) *Moral Philosophy*, New York, Macmillan,1967.

Goodin, Robert E. (1995) *Utilitarianism as a Public Philosophy*, Cambridge University Press, 1995

Griffin, James (1994) "The Distinction Between Criterion and Decision Procedure: A Reply to Madison Powers", *Utilitas*, vol. 6, no. 2, 1994.

Griffin, James (1992) "The Human Good and the Ambitions of Consequentialism", in *The Good Life and the Human Good*, (ed.) Paul, E. F., Miller, F. D., Paul, J., Cambridge University Press, p. 118-132, 1992.

Gruzalski, Bart (1981) "Foreseeable Consequence Utilitarianism", *Australasian Journal of Philosophy*, Vol. 59, No. 2, 1981.

Hare, Richard M. (1973) "Principles", *Proceedings of The Aristotelian Society*, Vol. 73, 1972-73.

Hare, Richard M.(1981) *Moral Thinking*, Oxford University Press, 1981.

Harman, Gilbert (1977) *The Nature of Morality*, Oxford University Press, 1977.

Harman, Gilbert (1986) *Change in View*, The MIT Press, 1986.

Hursthouse, Rosalind (1991) "Virtue Theory and Abortion", *Philosophy and Public Affairs*, Vol. 20, No. 3, Summer, 1991.

Jackson, Frank (1991) "Decision-theoretic Consequentialism and the Nearest and Dearest Objection", *Ethics*, Volume 101, Issue 3, 1991.

Jeffrey, R. C. (1965) *The Logic of Decision*, McGraw-Hill Book Company, 1965.

Kagan, Shelly (1998) *Normative Ethics*, Westview Press, 1998.

Kagan, Shelly (1989) *The Limits of Morality*, Oxford University Press, 1989.

Kant, Immanuel (1948) *Groundwork on the Metaphysics of Morals*, Routledge. (trans.) Paton, H. J., 1993.

Lenman, James (2000) "Consequentialism and Cluelessness", *Philosophy and Public Affairs*, vol. 29, no. 4, 2000.

Lif, Jan (2003) *Can a Consequentialist be a real friend? (Who cares?)*, Acta Universitatis Gothoburgensis, 2003.

Mackie, John L. (1977) *Ethics*, Penguin Books, 1990.

Moore, George. E. (1903) *Principia Ethica*, Cambridge University Press, 1993.

Mill, John S. and Bentham, Jeremy (1987) *Utilitarianism an Other Essays*, (ed. Alan Ryan) Penguin Books, 1987.

Miller, Dale E. (2003) "Actual-Consequence Act Utilitarianism and the Best Possible Humans", *Ratio* (new series), Vol. XVI, March, 2003.

Nell, Onora (1975) *Acting on Principle*, Columbia University Press, 1975.

Nozick, Robert (1974) *Anarchy, State, and Utopia*, Basil Blackwell Ltd,1986.

Oakley, Justin (1996) "Varieties of Virtue Ethics", *Ratio*, Vol. IX, no. 2, Sep. 1996.

Parfit, Derek (1984) *Reasons and Persons*, Oxford University Press, 1991.

Railton, Peter (1984) "Alienation, Consequentialism, and the Demands of Morality", *Philosophy and Public Affairs,* Vol. 13, no. 2, 1984.

Rawls, John (1971) *A Theory of Justice*, Oxford University Press, 1992.

Rawls, John (1980) "Kantian Constructivism in Moral Theory", The Journal of Philosophy, vol. LXXVII, no. 9, September 1980.

Resnik Michael D. (1993) *Choices*, University of Minnesota Press, 1993.

Rheinehart, Luke (1971) *The dice man*, Harper Collins Publishers, 1999.

Ross, David (1930) *The Right and The Good*, Oxford Clarendon Press, 1973.

Shaw, William H. (1995) *Moore on Right and Wrong*, Kluwer Academic Publishers, 1995.

Shaw, William H. (1999) *Contemporary Ethics: Taking Account of Utilitarianism*, Blackwell Publishers Inc., 1999.

Sidgwick, Henry (1874) *The Methods of Ethics*, (Seventh edition), Macmillan and Co., Limited, London, 1930.

Singer, Peter (1993) *Practical Ethics*, (Second edition), Cambridge University Press, 1993.

Slote, Michael (1985) *Common-sense Morality and Consequentialism*, Routledge & Kegan Paul, 1985.

Smart, J. J. C. and Williams, B (1973) *Utilitarianism for and against*, Cambridge University Press, 1973.

Smith, Holly M. (1986) "Moral Realism, Moral Conflicts, and Compound Acts", *The Journal of Philosophy*, Vol. 83, No. 6, 1986.

Smith, Holly M. (1988) "Making Moral Decisions", *Noûs*, Vol. 22, No. 1, 1988.

Smith, Holly M. (1989) "Two-Tier Moral Codes", *Social Philosophy & Policy*, Vol. 7, Issue. 1, 1989.

Stocker, Michael (1976) "The Schizophrenia of Modern Ethical Theories", *The Journal of Philosophy*, Vol. 73, no. 14, 1976.

Tersman, Folke (1991) "Utilitarianism and the Idea of Reflective Equilibrium", *The southern Journal of Philosophy*, XXIX, no. 3,1991.

Tännsjö, Torbjörn (1998) *Hedonistic Utilitarianism*, Edinburgh University Press, 1998.

Tännsjö, Torbjörn (1990) *Moral Realism*, Rowman & Littlefield Publishers, 1990.