The London School of Economics and Political Science

Doing the best one can (while trying to do better)

Ittay Nissan-Rozen

A thesis submitted to the Department of Philosophy, Logic and Scientific Methods of the London School of Economics for the degree of Doctor of Philosophy, London, June 2011

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Abstract

The thesis explores the question of how should a rational moral agent reason and make choices when he finds himself accepting inconsistent moral judgments. It is argued that it is both conceptually and psychologically justified to describe such an agent as suffering from uncertainty. Such uncertainty, however, is not uncertainty regarding the truth of some descriptive claim, but rather uncertainty regarding the truth of a normative claim. Specifically it is uncertainty regarding the truth of a moral judgement. In the literature this is sometimes called “moral uncertainty”. Two different lines of philosophical literatures that explore the idea of moral uncertainty are discussed. The first line – the one that originated from David Lewis’ argument against the “Desire as Belief Thesis” – explores the mere possibility of moral uncertainty, while the second line explores the question how ought a rational moral agent choose in face of moral uncertainty. The discussion of these two lines of research leads to the conclusion that a consistent account of moral decision making under conditions of moral uncertainty *that will be applicable to the kind of cases that the thesis explores*, must make use of degrees of beliefs in comparative moral judgements (i.e. judgements of the form “act a is morally superior to act b”) and of them alone. Specifically, no references to *degrees of moral value* should be made. An attempt to present such an account in the framework of an extension of Leonard Savage’s model for decision making is carried out. This attempt leads to a problematic result. Several implications of the result to ethic and meta-ethics are discussed as well as possible ways to avoid it. The conclusion is partly positive and partly negative: While a plausible account of moral decision making under conditions of moral uncertainty is presented, an account

of moral reasoning that aims at finding a complete moral theory (i.e. a moral theory that gives a prescription to every possible moral choice) is shown to be a very difficult – if not impossible - aim to achieve.

Acknowledgements

I would like to thank my supervisors, Richard Bradley and Alex Voorhoeve, for their intellectual as well as personal support. Working with Richard and Alex was not only an extremely stimulating and enriching intellectual experience, but also an inspiring lesson on how to create a social environment that enables a true philosophical dialogue. Richard once told me that he takes supervision to be the most influential part of the academic work. In the last four years, through the example of his and Alex's work with me and with some of my fellow PhD students, I have learnt both that he was completely sincere in making this claim and that he was right in making it. I have no doubt that any future work I might produce will owe a significant debt to both of them.

I have found a rare combination of a vibrant philosophical community and a supporting social environment at The Department of Philosophy, Logic and Scientific Method at LSE. In their fostering of the unique atmosphere here, the faculty, staff and graduate students have contributed a great deal to this dissertation. Special thanks go to Tom Chivers, the postgraduate program administrator, whose impact on the department's life and especially on the life of the department's graduate students is invaluable.

Several members and associates of the department have discussed parts of this dissertation with me. Especially helpful were Ken Binmore, Nick Baignet, Katie Steele and Franz Dietrich. I would like to thank also Nancy Cartwright, Luc Bovens, Roman Frigg and Christian List for spending time discussing with me some of the side projects I was involved in during my work on the dissertation.

My fellow PhD students in the department have given me many comments on the dissertation. Especially helpful were Foad Dizadji-Bahmani, Conrad Heilmann, Chris Thompson, Stan Larski, Seamus Bradley, Hlynur Orri Stefansson, Bengt Autzen, Esha Senchaudhuri, Roberto Fumagalli and Ben Ferguson. Ben also kindly agreed to edit this dissertation – a job that only highly virtuous people can hope to be successful at – and he was extremely successful in this respect.

I have also discussed parts of this dissertation with several people outside of the department. For their comments, suggestions and general support I am indebted to Noam Nisan, Shiri Cohen, Rami Yeari and Yair Levi.

Writing this dissertation would not have been possible without the financial support I have received from the following sources: The ORS fellowships scheme, the LSE's Research Studentship Scheme, The Jacobsen Fellowships and AVI Fellowships.

The real motivation for writing this thesis was, as with so many other things I do, the hope of impressing Aliza, my wife. Thank you Aliza for not being impressed by it, and still loving me in the way you do.

Final thanks go to my three sons, Dor, Aner and Yahav, for a special kind of support, the nature of which I do not know how to express in words.

This thesis is dedicated to my parents, who taught me both how to reason about morality and how to be moral

Content

4. The Triviality result

5. (Being fair as) Doing the best one can

List of tables

Introduction

“Morality is made for man not man for morality”

William Frankena (1973. p.116)

"…an individual making a moral value judgment must follow, if possible, even higher standards of rationality than an individual merely pursuing his personal interests”

John Harsanyi (1978, p.226)

This thesis is about moral decisions. This means that it is about decisions *of a special kind*, but it also means that it is about a special kind of *decisions.* The two quotes from Frankena and Harsanyi correspond to these two points of view. I believe each contains a deep truth regarding morality and, in a sense**,** this thesis is an attempt to explore whether it is possible to give an account of moral decision making that is faithful to these two truths.

My conclusion will be negative. I will try to follow in this thesis what seems to me the most promising route one should take in order to develop such an account, and will show that it leads to a dead end. This does not rule out the possibility that by taking another route one might be able to develop a satisfactory account of the kind that I am seeking. I have not given up hope of developing such an account yet, but the story that will be told here is a story of failure. However, as is the case with many failures in philosophy, it creates, so I will argue, some new and exciting opportunities.

Of course, all of this is somewhat premature. As a first step, I must present specific interpretations for the two quotes. Indeed, large parts of this thesis will be dedicated to presenting and motivating these interpretations. However, as the ideas expressed in these quotes are the cornerstones of the thesis, it is appropriate to give now a very general sketch of what I take them to mean.

Firstly, and more straightforwardly, it seems to me - Harsanyi's demand: When we make moral decisions, we must obey the same principles of rationality we obey when we make decisions generally (or even more demanding principles). Throughout this thesis I will assume that these principles are the principles of what is sometimes called instrumental rationality, i.e. principles of consistency in one's judgements and behaviour, not principles that refer to the content of one's judgements. What follows are a few remarks regarding this demand.

Harsanyi writes that an "individual making a moral value judgment *must* follow…", but which sense of "must" is used here? Is it "must in order to be rational"? I think not. Choosing such an interpretation does not seem plausible, since by doing so, one actually reads the quote as a tautology (which it is clearly not): "in order to be rational one must … follow the principles of rationality…". However, maybe what Harsanyi meant is simply that rationality demands the same things in non-moral and moral contexts. This seems too weak, though. Harsanyi's claim is clearly not a claim about rationality and its scope. It is a claim about how moral judgements should be made. It seems to

suggest that people should, in some sense, be rational in their moral judgements.

Thus, I believe the right way to read the quote is by replacing "must", with "ought", i.e. to read the quote as claiming that it is a moral obligation to be rational in one's moral judgements. I believe this is indeed what Harsanyi had in mind, but even if this is not the case, this reading is the one that will be the centre of my attention in this thesis. Notice that this claim itself is a moral judgement. It is a moral judgement regarding the way one makes one's moral judgements, i.e. a second order moral judgement, but, nevertheless, a moral judgement.

Why should we accept this moral judgement? I think there are good reasons for this (some of which are discussed by John Broome in his 1991 book), but I also believe that a full justification for this judgement must involve giving an answer to the question "why be rational?", and as is clear from current discussions in the literature (see Kolodny 2005 and the various responses to it), there is still a long way to go before a satisfactory answer to this question will be found. In any case, I will discuss neither this question, nor any other reasons we might have to obey Harsanyi's demand. Rather, I take this demand as an assumption in this thesis. The thesis is an attempt to explore the possibility of reconciling two demands that are at least initially plausible. Any conclusion that I might reach regarding this possibility, might then help to justify each one of the demands.

Secondly, Harsanyi's quote includes two qualifications**.** The first one is expressed by the phrase "…if possible…" and the second one is expressed by the phrase "…even higher standards…". It is not entirely clear what Harsanyi means by both of these qualifications, as he does not give any examples, either of cases in which it is not possible to follow the usual rationality postulates, or of postulates that constitute "higher standards of rationality".

In chapter 4 of the thesis I will argue that sometimes it is indeed impossible for a moral agent to follow the standard principles of rationality and will characterise a set of such cases. In chapter 5 I will suggest a different principle of rationality that should be followed when moral agents find themselves in such situations. Does this principle constitute a higher standard of rationality? I am not sure what "higher" as used in this context means, but in any case it is a different principle of rationality. The important point here is that it is possible to understand Harsanyi's claim not as an inclusive requirement, but rather as one that leaves some room for flexibility.

Moving on to (the more problematic to explicate) Frankena's quote, I wish to present an interpretation which is clearly different to the one Frankena had in mind when he chose it to close his book. Frankena wrote: "…society… must remember that morality is made to minister to the good lives of individuals and not to interfere with them any more than is necessary. Morality is made for man, not man for morality." (Frankena, 1973, chapter 6). So it seems that what Frankena requires is that a moral theory not be too demanding in the sense of asking people to sacrifice too much in terms of their personal interests for the

sake of morality. Frankena's reason for this demand seems to be that by violating this requirement, a moral theory fails to serve its goal which is to "minister the good lives of individuals", when the emphasis here is on the expression "the good lives": Lives in which people have to sacrifice too much in terms of their personal interests are not good.

I agree. However, another reading of Frankena's claim is possible. A moral theory can be too demanding in the sense of asking people to sacrifice too much in terms of their personal interests, but it can also be too demanding in the sense of asking them to sacrifice too much in terms of acting against what they judge to be the right thing to do; *in the sense of asking them to sacrifice too much in terms of their moral interests*. A moral theory that does that, fails to minister the good lives of individuals not because the lives it aims to minister cannot be good, but rather because it cannot hope to direct these individuals' lives. In other words, a moral theory, I argue, should have enough motivational force to be accepted by people.

Again, I do not intend to argue for this claim; rather, it is an assumption that I rely on in this thesis. However, I think a prescriptive theory that violates this demand does not constitute what I used to call (until I wrote this thesis) a "moral theory". When I used the term "moral theory", part of what I meant was a theory that can guide human decision making. It may be that this is not the right way to use the term. In fact, I will briefly discus this possibility in chapter 4, but it is very appealing.

Accepting the suggested interpretation for Frankena's quote as a demand on moral theories does not preclude also accepting the demand that a moral theory should be motivational in the standard sense of not asking us to sacrifice too much in terms of our personal interests. I also accept this latter demand. Taking account of this demand in the thesis, however, would add many complexities to the discussion and so I have chosen to exclude it. Thus, in this thesis I deal with idealised moral agents who have no personal interests apart from doing the morally right thing. Such agents do not exist, of course, but I think that the conclusions I will reach regarding this kind of agents will also be relevant for the case of real moral agents.

What is it for a theory to have a motivational force on an idealized moral agent? Intuitively, all I mean by that is what I wrote above: it is for this theory to be composed of judgements that the agent will be willing to act on. For my discussion, though, I will need a more precise characterisation. The question of the appropriate way to do that will be discussed in the first three chapters of the thesis. I will present and motivate a number of increasingly more accurate characteriszations of the demand, and will argue that we should take all of them to express the same idea (in different levels of specification).

Specifically, I will argue that for a theory to have a motivational force on an agent, it must be the case that, regarding each one of the judgements of the form "in situation x you ought to do y" derived from it, the agent must believe it is, more likely than not, true. This is equivalent, I will argue, to demanding that the theory is composed of judgements that are in a state of wide reflective

equilibrium for the agent, and to claiming that the agent must believe, regarding each one of the judgements derived from the theory, that he is justified in accepting it.

I mention this now because I want the reader to have some idea of the direction in which I am heading, but, of course, none of the claims made in the last paragraph is self-evident, and they will be both clarified and justified.

So, these are the two main assumptions I am going to rely on in this thesis. I will call them “the rationality demand” and the “motivational demand”. The possibility that a tension between these two demands would arise should be apparent. This is because while the rationality demand is not sensitive at all to the content of one’s judgements, the motivational demand is. The motivational demand requires that the prescriptions a moral theory gives will not be too detached from one’s actual moral judgements, and the rationality demand requires that these prescriptions will obey the rationality postulates. Thus, if one’s actual moral judgements do not obey the rationality postulates, we can expect that any possible moral theory will have to violate at least one of the demands. This, I believe, is the essence of the main problem that I explore and try to solve in this thesis.

To make things clearer, it might be helpful to consider here the paradigmatic case I will discuss in the thesis. This is the case of an agent that finds out that he holds moral judgments that violate the requirement of transitivity.

There are two kinds of moral judgements; judgements concerning how valuable from a moral point of view a given act or outcome is on its own (let us call them non-comparative moral judgements) and judgements concerning which one of two or more acts or outcomes is morally superior to the other(s) (let us call them comparative moral judgements).

Comparative moral judgments, I assume, ought to be consistent in the decision theoretic sense. The “ought” here is, as explained, rational – i.e. they ought to be consistent in order for them to be rational – but it is also a moral “ought”. This is so, because we are morally obligated (so I assume) to be rational when acting as moral agents.

What, however, should an agent that accepts that comparative moral judgements ought to be consistent do when he realizes that he holds inconsistent comparative moral judgements? For example, what should a moral agent who realizes that he holds comparative moral judgements that violate transitivity do? There are two different questions here: 1) how should such an agent choose, when acting as a moral agent and when a decision must be made? 2) How should such an agent change his judgements so that they will become transitive, in case there is no need for an immediate decision? In other words, how should such an agent reason himself to rationality?

Most of this thesis will be dedicated to the exploration of the second question, not the first one. Thus, most of this thesis will concern moral reasoning, not moral decisions. However the kind of moral reasoning that will be discussed is a

special kind of moral reasoning: it is moral reasoning that aims at choice. The end-state of this type of moral reasoning should be a set of choice recommendations which is both internally consistent in the decision-theoretic sense and have the power to motivate (at least ideal) moral agents. So a complete answer to the second question must include an answer to the first question for the special case when the agent has reached the end-state of his moral reasoning process. This is why I wrote in the opening paragraph of the thesis that it is about moral decisions. It is about moral decisions, but it is about moral decisions that are made by agents that have done all the reasoning they can.

An agent that realizes that he holds inconsistent judgments, and so tries to change some of them to gain consistency, can be describe – I will argue - as an agent who is involved in the process of arriving at a reflective equilibrium (RE). I will discuss this notion more deeply – from an unusual perspective – in chapter 1, but even before doing so, I think it will be justified to point to three conditions that it will be desirable if they will be satisfied in a RE:

1. The comparative moral judgements that the agent accepts are consistent in the decision theoretic sense.
2. The non-comparative moral judgements the agent accepts are consistent with the agent's comparative moral judgements in the sense that the agent judge one act to be morally superior to another iff the expected moral value of this act is higher than that of the other (notice that acts can have uncertain consequences even in a RE).

3. The agents' moral preferences (i.e. the preferences according to which he acts, when acting as a moral agent) are identical to the agent's comparative moral judgements.

The first two conditions follow from my commitment to the rationality demand. If the first condition is satisfied then it is possible to describe the agent's comparative moral judgements – so tell us the different representation theorems of decision theory – as constituting an order that ranks the acts according to the level of their expected moral value, for some value functions. The second condition demands that the value function the agent adopts in a RE is one of these functions.

The third condition is not, strictly speaking, one of the conditions of RE. RE, as usually characterized in the literature, imposes demands on the agent's set of judgements, not on the relation of these judgments to other attitudes the agent holds. However, I think that this condition expresses the motivation behind the RE idea: we are interested in RE only because we believe that the judgments one accepts in a RE are the ones that ought to direct one's choices. Thus, the third condition is the one that expresses the motivational demand: it demands that the moral judgements an agent ends up accepting in a RE, are the ones that direct his behaviour. I will discuss this point more deeply in chapter 3.

As attractive as these three conditions are, my conclusion will be that there might be good reasons to give up on at least one of them. We have many more steps to take, however, before reaching this conclusion.

A central claim that I am going to rely on in this thesis is the following one: when an agent realizes she holds inconsistent comparative moral judgements, she usually becomes uncertain regarding which of the inconsistent judgements she holds she ought to reject.

This claim is partly descriptive and partly conceptual. The descriptive part is the observation that people usually experience something that they choose to describe as “uncertainty” in this kind of situations. The conceptual part is the claim that this attitude is the same type of attitude people refer to when they say that they are not sure if it will rain tomorrow, for example[1].

I believe that most people will accept the descriptive component of the claim and I will discuss it in length in chapter 1.

The conceptual component is, I think, the more controversial one as by accepting it, one commits oneself to accepting the moral cognitivist position, i.e. the position according to which moral judgements are beliefs. Moreover, as I will suggest that one’s degrees of beliefs in comparative moral judgements should constrain the way one decides which one of several inconsistent comparative moral judgments one ought to reject in face of the inconsistency, I am committing myself not only to moral cognitivism, but also to anti-Humeanism, the rejection of the Humean thesis that beliefs cannot constrain desires (and

[1] The claim also has a normative component: the demand that one’s degrees of moral beliefs will be probabilistic. I will not discuss this claim at length in this thesis. One can find in the literature many arguments for this demand in a non-moral context (for example see Joyce 1998) and there seems to be no reason that the demand will lose its normative force specifically in the moral context. I will, however, make a few comments on the issue in chapters 3 and 4.

through them preferences) in any way that is not captured by the standard axioms of decision theory[2].

Neither of these commitments is uncontroversial. The task of defending each one of them against all the objections one can find in the literature is not one that I can hope to achieve in this thesis. In chapters 2 and 4, I will argue, however, that in the context of this inquiry, adopting either the non-cognitivist position or the Humean one is not an attractive move to take.

If one does accept that when one realizes that one holds inconsistent comparative moral judgements one might become uncertain regarding which of these judgments one should reject and which of these judgments one should hold on to, then a natural strategy for dealing with the problem of reasoning oneself to rationality suggests itself.

The strategy is the following one. We should try to formulate conditions that connect one's degrees of beliefs in comparative moral judgements to the comparative moral judgements one holds in a RE. If we are able to formulate such conditions that ensure that an agent that follows them ends up accepting a set of comparative moral judgements that obey the rationality axioms, and if these conditions can be satisfied not only in trivial cases, we will have a thesis that can be used by an agent that holds inconsistent moral judgements and yet

[2] This is at least one way the Humean position is formulated. In particular, this is how John Broome (1999) formulated the thesis and this is the formulation implicit in David Lewis' discussion of the Desire as Belief Thesis that will concern me in chapter 2. Lewis' argument against the desire as belief thesis aims to show that there can be no non-trivial anti-Humean thesis that respects the rationality demand. I will try to construct such a thesis. For some discussions regarding what the Humean position is see Lewis (1988), Smith (1987), Broome (1999) Chapter 5.

wishes to modify them in order to gain consistency[3]. As I will argue in chapter 1, it also seems reasonable to argue that the judgements the agent will end up accepting in such a RE, have a motivational force on the agent.

Adopting a set of such conditions can be interpreted as a Bayesian formulation of the (wide) reflective equilibrium method. Almost nothing was said in the literature about the way an agent who is involved in the process of achieving a RE decides which initial moral judgements to keep and which to reject and what (if any) constraints should guide such an agent in his reasoning. In the absence of such constraints, the wide RE method seems to be just a *characterisation* of any reasoning. Reasoning is, in a sense, just the process of achieving coherence among one's judgements. Viewed in this way, the method of RE seems somewhat trivial.

This point was raised and discussed by T.M. Scanlon (2003) and others (for example Singer 2005). In chapter 1 I will argue that the strategy outlines above is also a way to save the RE method from this triviality. What is missing, I will argue, from the current characterisations of the RE method is a set of criteria that tell us how an agent, engaged in a process of achieving a RE, should choose which judgements to keep and which to amend. In order for the method of RE to have any bite at all, we must add something to it. This "something", I suggest, is a set of conditions that describe the way in which the reasoner's degrees of belief in moral judgements are related to the judgements he ends up accepting.

[3] We will also have a non-trivial consistent anti-Humean thesis.

I will also argue that even though it has never (as far as I know) been discussed at length in the RE literature, this idea may be what many of the defenders of RE had in mind when they wrote about the matter.

Thus, if successful, the route I am going to take in this inquiry should lead me both to a new understanding of the RE method and to a prescriptive theory of how to reason oneself to a moral theory which is both rational and motivational[4]. I will follow this route in the first three chapters of the thesis. Chapters 1 and 2 will prepare the ground for the discussion that will follow in chapter 3, by examining and drawing connections between some apparently unrelated discussions in the philosophical as well as psychological literatures. In chapter 3 – while relying on the conclusions of the discussion in the previous two chapters – I will present an account of moral reasoning of the kind I am looking for, in the framework of a formal model.

Chapter 1 will have three main aims. The first one will be to present an initial characterization of the motivational demand, using the RE idea, and to motivate this characterization. The second one will be to present some psychological findings regarding inconsistent judgements and choices that will serve me in chapters 2 and 3, and to clarify their relevance to the question this thesis explores. The third one will be to draw connections between these findings, the RE idea and the idea if moral uncertainty and to suggest –based on these

[4] It should be clear by now, but maybe it is still worth mentioning, that what I am after in this inquiry is not a first-order moral theory. Rather, I am looking for a normative epistemological account of a reasoning procedure that can lead to accepting a first-order moral theory that respects my two demands.

connections – a general strategy for dealing with the problem of reasoning oneself to rationality.

In chapter 2 I will move on to discuss more deeply the notion of moral uncertainty. I will first address an argument presented by David Lewis that can be interpreted either as a refutation of the claim that moral uncertainty is possible, or as a refutation of the claim that moral uncertainty can be used to restrict the choices a rational moral agent makes. Lewis' argument is not usually discussed in the literature on moral uncertainty, but I will show that it must be dealt with.

I will deal with it. By this I do not only mean that I will find a way to block the threat it poses to the mere possibility of using the notion of moral uncertainty in a philosophical inquiry, but also that I will take advantage of the important lessons we can learn from it, as well as from the literature that has discussed it.

The most important lesson is, I will argue, the following. Current accounts of moral decision-making under conditions of moral uncertainty treat moral uncertainty in much the same way that decision theory treats uncertainty regarding the state of the world, i.e. by demanding that in face of moral uncertainty one should maximise expected moral value (see Lockhart 2000 for example). Such a demand, however, is based on the thought that the following two assumptions are true. First, that one can reduce the moral uncertainty one suffers from regarding the question which one of the available acts is the

morally right act to choose to uncertainty regarding the question which one of several moral theories or general moral claims is correct.

Second, that one is able to tell how good or how bad every possible act available to one is, according to each one of the moral theories one believes might be true, and one is able to compare these values across theories. Drawing on some of the conclusions of chapter 1, I will argue that both of these assumptions should be rejected, at least in some cases.

I will then suggest that in cases in which at least one of these assumptions should be rejected, one has no alternative but to make one's decisions based solely on one's beliefs regarding which act is the morally right act to choose, i.e. without making any reference to degrees of moral value. This conclusion will allow me to present in chapter 3 a formal version of the problem, which was described here in an informal way.

As will be clear from the formal presentation of the problem, it is just an instance of the much discussed lottery paradox in which a rational agent finds himself in a situation in which he must "accept as true" an inconsistent set of judgements. Since the judgements in question, in this thesis, are judgements regarding what one ought to do morally, the importance of the notion of "acceptance" here cannot be dismissed. This is because by accepting a moral judgement, one commits oneself to act on it.

Having presented the problem in a formal way, I will try to follow the general strategy I have suggested for dealing with it: when one finds oneself in a lottery paradox in a moral context, one should try to escape this situation by changing one's degrees of belief. The formal model that I will use will allow me both to give this strategy more structure and to examine whether it can – in principle – lead an agent to accept the kind of moral theory I am interested in.

In practice I will try to use the model in order to prove a representation theorem according to which, if one respects plausible axioms regarding moral decision-making in terms of moral uncertainty, it is possible to represent one as maximising the expectation of some value. If such a theorem holds in the case of some distributions of degrees of beliefs over the set of possible moral judgements an agent can hold, it is possible to argue that these distributions constitute the set of possible degrees of beliefs an agent can have in a reflective equilibrium that respects Harsanyi's demand. Thus, this could be seen as a solution to the problem.

When I first formulated the problem in this way, my hope was to prove such a representation theorem. I was indeed able to do that. However, as will be discussed in chapter 4, the set of distributions in which it holds turns out to be trivial. Thus, instead of constituting a solution to the problem, the result, I will argue, should be interpreted as implying that the problem is inescapable. More accurately, it should be interpreted as implying that whenever one suffers from moral uncertainty then, except in trivial cases, one must either hold intransitive

moral preferences, or one must sometimes act against one's own moral judgements.

The novelty of the result is that it shows that there is no way an agent can change her judgements (or degrees of belief in the propositions which are the objects of them) to help her escape the problem (except in trivial cases). In other words, it shows that lottery paradoxes are not only possible in the moral domain, but are, in fact, inescapable.

I will discuss possible interpretations for this result, as well as some possible ways to avoid it. However, my tentative conclusion will be that what the result shows is that any plausible complete moral theory (i.e. a theory that gives prescriptions for every possible choice problem) cannot be wholly motivational even for ideal moral agents (i.e. agents who are only motivated by moral considerations).

As negative as this conclusion is, it does have some advantages. Some of them are explanatory and others are more ethically substantive. I will discuss some of these advantages tentatively in chapter 4 and in the conclusion and will move on, in chapter 5, to a more rigorous discussion of one of them.

The discussion in chapter 5 begins by considering the possibility of relaxing the transitivity of preferences axiom. As explained above, this is "technically" not a violation of Harsanyi's demand, at least as interpreted here, since it can be argued that an agent must indeed have transitive moral preferences when

possible. However, the result presented in chapter 4 shows that sometimes it is indeed impossible (at least if the agent obeys two other axioms that I will present) for an agent to have transitive moral preferences.

Although relaxing transitivity is generally extremely unattractive, I will argue that in the context of actual moral decision-making this is less so. This is because all that the result shows is that an agent that follows the axioms in the model must have intransitive preferences among some *possible* acts. However, these might be acts that the agent only considers hypothetically, not acts that are really available to the agent. I will argue that such hypothetical inconsistency in one's moral judgement is not a strong enough reason for an agent to act against her own judgements regarding acts that are actually available to her.

Moreover, in those cases where the agent actually has intransitive preferences over acts that are available to her, relaxing transitivity in the model turns out to open the way for an elegant explanation of what makes a lottery sometimes the right act to choose. Relaxing transitivity leaves the question open as to which act the agent should choose. I will demonstrate that if we allow the agent to use mixed strategies, i.e. if we demand that the set of possible acts available to the agent is convex, then there always exist a mixed strategy such that the agent believes it is more likely or equally likely morally better than any other act available to her. Thus, choosing one of these acts seems the only rational thing to do, for an agent who finds it impossible to have transitive moral preferences.

I will argue that this phenomenon not only explains the rightness of lotteries, but also offers a good and new explanation of which lotteries are justified and in which situations.

Although, for the reasons given, in the context of actual moral decision-making it might not be so worrying that an agent has intransitive moral preferences, when it comes to a moral inquiry that aims at finding the correct moral theory, I find such intransitivity unacceptable. Thus, one tentative (tentative because I still hope to find a way to avoid it) conclusion I will reach will be that we have reasons to be sceptical of the possibility of rational moral reasoning that aims at a complete moral theory.

There is no need to explain why this is a very worrying conclusion. However, it has also some positive implications. The most important one, it seems to me, is the following. Since the triviality result that will be presented holds for any set of beliefs, not only for beliefs that are based on intuitive judgements, the sceptical conclusion that might follow from it has nothing to do with intuitions either. In fact, the result can be taken to *explain* why our moral intuitions are sometimes inconsistent.

The explanation it offers shows that the inconsistency does not arise as a result of some contingent circumstances that made human beings develop in a certain way. Rather, the inconsistency is a necessary by-product of the combination of two features moral agents usually have: that they can be uncertain regarding their moral judgements and that they want their moral judgements to be

consistent. That is, to the extent that moral agents can suffer from moral uncertainty, they will sometimes have to accept inconsistent judgements, *no matter what their reasons for accepting these judgements are.*

Now, moral agents that never suffer from moral uncertainty are immune to this threat, but why would such moral agents want to get involved in a moral inquiry in the first place? After all, they are absolutely sure of their judgements regarding all moral questions, so why should they try to reason about any of these questions?

It seems, then, that it might not be our intuitions that we should blame for the inconsistent judgements we sometimes hold. The inconsistency would arise (and will arise) given any set of reasons or moral evidence one might use in one's moral inquiry. Thus, to the extent that rational moral reasoning is possible, there is no reason to exclude our moral intuitions from playing a role in it. Since, as I have argued, the result does not threaten the possibility of rational moral reasoning in the context of specific moral decision-making problems, it seems that in such contexts, we should take our intuitive moral judgements seriously. However, when it comes to moral inquiry that aims at the correct moral theory, this might not be the case.

However, many scholars who have discussed the psychological findings regarding the way moral judgements are produced, suggest that the main lesson that should be learnt from these findings is the exact opposite. It is usually argued that when it comes to moral decision making, or to practical

ethics, little weight should be placed on moral intuitions, as they are unreliable (in a sense that will be discussed in chapter 1). At the same time, however, it is generally acknowledged that, ultimately, there is no escape from using some moral intuitions in an ethical inquiry that aims to find the correct moral theory (even among harsh critics of the reflective equilibrium method).

Indeed, in the literature, sometimes such a position is used in order to relax the tension between the competing claims that moral intuitions are unreliable and that we must, nevertheless, use some intuitions in our moral inquiry. Thus, when discussing the implications for ethics of psychological findings that question the reliability of moral intuitions, Peter Singer draws the following conclusion: “We need to think about what our underlying values are, and then distinguish these values from the moral intuitions that merely have a heuristic role in furthering them” (Singer 2005, p. 561).

The picture drawn by Singer has a feature that can be very attractive both to philosophers and to social scientists. It enables, although it does not necessitate, a kind of a “division of labour” between them: philosophers are free to explore what “our underlying values” are without worrying about the implications of the theories they develop in specific cases, while social scientists are free to give policy recommendations on the basis of a given set of values without having to trouble themselves with questioning these values on the basis of their implications for policy purposes.

My claim, on the contrary, is that such a "division of labour" is unjustified. The result presented in chapter 4 constitutes, I believe, strong evidence for my position. This result can be interpreted as showing that moral inquiry that aims at a complete moral theory must lead to some recommendations that a reasoner will find hard to accept, *no matter what the reasoner's reasons are*. It follows that when it comes to a moral inquiry that aims at a complete moral theory, we face a problem *with or without relying on our moral intuitions*. However, when scientists face a specific question of policy recommendation, people's intuitions regarding what is the right act to choose must play a restrictive role in their reasoning. This is not because these intuitions are a reliable guide for the correct moral theory (they might or might not be), but rather because by ignoring these intuitions, the policy recommendations are unlikely to be accepted.

In an interview with Alex Voorhoeve (Voorhoeve 2009), Daniel Kahneman raises a similar point. Referring to the judgement that it is impermissible to push the bystander in the Fat-Man version of the trolley problem (which will be discussed in chapter 1) he argues: "…since it's also an extraordinarily powerful intuition, you should not have a rule that ignores it. That is, if anyone had a system that would condone pushing the bystander to save the five, then that system would not be viable; that system would not be acceptable. On practical terms, it would not be a sensible moral system…".

It is true, of course, as Voorhoeve stresses in his response to Kahneman's point, that this is a pragmatic consideration that should not play a role when

considering the question as to what *really is* the right thing to do in the situation. However, when it comes to moral decision-making, pragmatic considerations should play a major role. Thus, it seems that although we do not have a justification for the use of intuitions in the case of a moral inquiry that aims at a complete moral theory (but we do not have a reason to preclude them from playing such a role, either), we *do* have a justification for the use of intuitions in a moral inquiry that aims at providing a policy (or action) recommendation for a specific moral problem. I take this to be a positive result[5].

So where does all of this leave us? While I was still in the process of developing the argument of this thesis, I thought of giving it the title "Doing the best one can, while trying to do better". The first part of title was supposed to refer to the context of a specific moral decision faced by an agent. What ought the agent to do in such a context? Regardless of what the decision is, I wanted to claim, he

[5] It is positive in the sense that it can help solving highly important problems in practical ethics. An example for a case like this is the debate regarding the appropriate measures that should be taken in order to prevent the possible damage caused by climate change. One of the central issues that is discussed in both the economic and philosophical literature on the subject is the question of how much is it justified to discount the welfare of future generations relative to the welfare of the current generation. An examination of the different positions, that one can find in the literature regarding this question, reveals, I think, that at least part of the disagreement arises as a result of different scholars adopting different approaches regarding a more fundamental issue, which is not what the discount rate should be but rather how should it be determined. Leaving aside the positions of those who refuse to accept that moral debate should play any role in these decisions (see Weitzman 2007 for example), one can still find different methodological approaches to the ethical question. Some scholars (for example Stern 2007 chapter 2) choose a methodology that seems in line with Singer's remark, i.e. they start from an abstract philosophical discussion regarding values and then, based on this discussion, assign values to the ethical parameters that determine the discount rate. Others, (see Dasgupta 2007) argue that these values should not be determined on a philosophical a priori ground without paying attention to our intuitive judgements regarding the actual implications (in terms of policy recommendations) that adopting such values leads to. Finally, still others (see Baron 2000) question our ability to deliver consistent intuitive moral judgements regarding the matter. Without discussing in length the implications for the matter of the conclusions of this thesis, it is easy to see that they will be according to the general line of (a) Letting intuitive judgements regarding specific policy recommendations play a role in the ethical inquiry and (b) Letting them play such a role in a way that will be sensitive to how reasonable it is to expect decision makers to actually follow these recommendation. For further discussion see Broome (2008).

must try his best, given the moral and non-moral information available to him, to do the right thing. The second part of the title was supposed to refer to the context of a moral inquiry that aims at a complete moral theory. This context is the one, I wanted to argue, in which the agent can try to do better than the best he can, i.e. he can try to enrich and improve the quality of the moral information available to him.

I think I was successful with the first part of the title: this thesis does provide, I believe, at least a partial explication of the expression “doing the best one can” in the moral context. As mentioned, however, I was less successful with the second part of the title. Nevertheless, the negative conclusion I have reached may help further inquiry on the matter, at least in the sense of excluding some apparently attractive possibilities. This is the reason I decided not to get rid of the second part of the title, but rather to put it in brackets. I believe we ought to do the best we can to do the right thing, and I think I understand what this generally means. I also believe that we ought to try to do better, and although I am not sure what *this* means, I am pretty sure that I know what it *does not* mean. Having this knowledge is a first step towards figuring out what it *does* mean.

Chapter 1: Reflective Equilibrium and Moral Psychology

Introduction

My aim in this chapter is to prepare the ground for the arguments that will follow in the next chapters. This will be done by a discussion of both the conceptual and empirical relations between the central ideas, phenomena, principles, and concepts I will use in this thesis. The discussion will serve me both to set the terminological landscape of the thesis and to present and argue for 8 claims that will serve me in later chapters.

In section 1 I will do two things. Firstly, I will use Peter Singer's famous "child in the pond" argument in order to clarify the relations between some of the central concepts I will use in the thesis. I will discuss the relations between moral judgements and moral beliefs and between acceptance and quantitative belief. I will also explain what I take to be the role of each one of these concepts in the reflective equilibrium method.

Secondly, I will argue for the flowing claim:

1. The right way to understand the motivational demand is not in terms of the level of intuitiveness of a moral theory's recommendations, but rather in terms of the level of their fit with the moral judgments an agent accepts in a reflective equilibrium.

In section 2, using two examples of a popular criticism on the reflective equilibrium method, I will argue for the following three claims:

2. The right interpretation to accept for the reflective equilibrium method in our context is the wide one.
3. The wide interpretation of the reflective equilibrium method threatens to make it too trivial to be useful in our inquiry.
4. One way to adopt the wide interpretation for the reflective equilibrium method while saving it from triviality is to specify consistency conditions on the way the reasoner chooses which judgements should he keep and which judgements should he reject in face of inconsistency.

In section 3, I will make use of some of the literature on the psychology of judgement and decision making in order to argue for the following two claims:

5. Psychological research tells us that when people do not have a direct access to degrees of moral value, it is likely that they will find themselves accepting comparative moral judgements that violate the rationality axioms.
6. Psychological research tells us that when people face moral decisions in which more than one morally relevant dimension in involved they will find it hard to assign exact degrees of moral value to the different acts available to them.

These two claims will serve me in two different ways. Firstly, in this chapter, I will use them in order to argue against a common claim among philosophers that hold a consequentialist ethical approach (according to which psychological research on the way people produce their moral judgements pose a problem only to non-consequentialist approaches).

Secondly, and more importantly, I will use both of these claims again in chapters 2 and 3 in my discussion of the question of moral decision making under conditions of moral uncertainty.

In section 4, I will use the conclusions of the first three sections in order to argue for the following two claims which together constitute the central conclusion of this chapter:

7. When people find themselves accepting inconsistent moral judgments they might become uncertain regarding which one of the judgments they accept they ought to reject.
8. Using people's degrees of beliefs in the propositions which are the objects of their moral judgements in order to formulate consistency conditions on the way they choose which judgements they ought to accept and which judgements they ought to reject in face of inconsistency is a promising route to take in order both to save the reflective equilibrium method from triviality and to find an account of a reasoning procedure that can lead one to a moral theory that respects my two demands.

The motivational demand and reflective equilibrium

Consider Peter Singer's classical argument from "Famine, Affluence, and Morality". It starts with a moral judgement that most people find hard to reject:

"If I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing." (Singer 1972, p.231).

It ends up with a moral judgement most people find hard to accept:

"…we ought to give until we reach the level of marginal utility-that is, the level at which, by giving more, I would cause as much suffering to myself or my dependents as I would relieve by my gift. This would mean, of course, that one would reduce oneself to very near the material circumstances of a Bengali refugee." (Singer 1972, p. 241).

In the middle, there are other moral judgements. Some of them are second order moral judgements, like the judgements that some considerations (physical distance, the fact that other people do not obey their moral duties) are morally irrelevant for the context of use, and some of them are first order moral judgements that are either generalisations, or slight modifications of the two judgements quoted in the first paragraph. For the sake of simplicity, let us concentrate on the three following moral judgements:

1. In the “child in the pond” story, the agent is morally obligated to save the child, even at the cost of ruining his shoes.
2. Wealthy people are not morally obligated to donate the money worth of a new pair of shoes for the purpose of saving dying children in under-developed countries, even when by doing that they will save, for sure, the life of one child.
3. There are no morally relevant differences between the situation described in the “child in the pond” story and the situation described in judgement 2 above.

Now, it seems inconsistent to accept all of these three judgements at the same time. On the other hand, for most people, both rejecting the first judgement and rejecting the second judgement, seem to be too drastic a move. So what remains is to put the blame for the inconsistency on whatever is going on in the middle, i.e. on the third judgement.

Singer recognised that this is a plausible reaction: “It may still be thought that my conclusions are so wildly out of line with what everyone else thinks and has always thought that there must be something wrong with the argument somewhere.” (Singer 1972, p. 238) and indeed many philosophers do react to Singer’s argument in this way. They look for ways to reconcile the strong intuition most people have that the first judgement is correct with the strong intuition most people have that the second judgement is correct.

As mentioned, the natural way to do that is to reject the third judgement, i.e. to look for differences between the situations to which the first and the second judgements refer, which are, plausibly, morally relevant. Maybe, it is not the physical distance, for example, that is responsible for the difference between the judgements, but rather something that is correlated with it, like cultural connections or political connections (for example see Walzer 1983**)** .

Singer, on the other hand, took another path. It is true, argued Singer, that the second judgement is "…one which we may be reluctant to face". However, he continues, "I cannot see, though, why it should be regarded as a criticism of the position for which I have argued, rather than a criticism of our ordinary standards of behaviour" (Singer 1972, p. 238).

I find Singer's comment not entirely fair. It is not only that rejecting the second judgement is very demanding, in the sense of asking us to sacrifice too much in terms of our personal interests that bothers us about it. It is also that it is very unintuitive. Of course, it is not unintuitive that it is good to donate money for the purpose of saving dying children in under-developed countries. What's unintuitive is that it is obligatory to do so, as long as one's standards of living exceed a very low threshold. For many people it is as unintuitive to accept that it is obligatory to do so as it is to accept that we are under no obligation to help the drowning child in the pond. Thus, it is not clear why they should reject the former and accept the latter. A reasonable reaction by such people seems to be trying to save both of the judgements somehow, even if this involves telling

complicated stories, for example pointing to some differences between the situations and explaining why these are morally relevant.

Such a strategy is very common in ethics today and was common long before John Rawls came up with the name “Reflective Equilibrium” (Rawls 1971) to refer to a specific version of it. Before discussing the notion of Reflective Equilibrium (RE) in more depth, I want to use the inconsistency that stands in the heart of Singer’s argument in order to make some important conceptual clarifications.

I wrote “it seems inconsistent to accept all of these three judgements at the same time”, but what exactly does “accepting a moral judgement” mean? Contemporary philosophers use the world “judgments” in two different ways. Sometimes they use it to refer to acts (mostly - but not exclusively - verbal acts) that express mental attitudes of agents, and sometimes they use it to refer to the attitudes themselves. This is specifically true regarding moral judgments. Thus, on the one hand, Martha Nussbaum likes to promote the thesis that “emotions are judgments” (see for example Nussbaum 2001, p.37) and supports this thesis with observations such as that “judgments come in varying degrees of confidence” (Roberts, 1999, p. 794), which clearly indicates that they are mental attitudes, but on the other hand, Robert Solomon that promotes the same thesis claims that “An emotion is a *judgment* (or a set of judgments), something we *do…”* (Solomon 1976 p.185, Solomon’s Italics) and Nussbaum herself argued that “…the appearance has become my judgment and that *act* of acceptance is what judging is” (Nussbaum 2001, p.37, my Italics).

In the same way, Alan Gibbard wonders whether moral judgments “are factual beliefs of some kind or something else” (Gibbard, 1990, p. 130). Gibbard indeed rejects the thesis that they are factual beliefs, but his choice to contrast his notion of moral judgments with the notion of beliefs (which are mental attitudes) indicates that he takes his account as treating moral judgments as mental attitudes as well (though different from beliefs).

If taking to be acts, producing moral judgments is a kind of a decision as the agent must decide which act of judgment to perform out of all the possible judgments available for him. Decisions can be rational or irrational, of course, but their (ir)rationality cannot be assessed solely in terms of their consistency with other decisions the agent makes. Some information regarding the agent’s mental attitudes must be used in such an assessment. For example, it is not hard to imagine a situation in which it is perfectly rational to perform the verbal act of judging that all three claims I have used above in order to describe Singer’s argument are true. If the agent knows that performing such an act is the only way for him to win a reward that he desires then performing this act seems like a rational decision to make (even in case the agent realizes that it is impossible for all three claims to be true).

The irrationality involved in judging the three claims to be true is not, thus, an irrationality of decisions. Rather, it is an irrationality of attitudes. Under any circumstances, so it seems, it is irrational to judge – in the mental attitude sense

– that all three claims are true. Thus, in the context of this inquiry the right way to use the word “judgements” is in the mental attitude sense.

If taken to be mental attitudes, the question as to what kind of mental attitude moral judgements are, arises. The literature here is vast (for an overview see Van Roogen 2008). For our purpose the important distinction is between cognitivist to non-cognitivist approaches. Very crudely, cognitivist approaches takes judgements generally and moral judgements in particular to be beliefs and non-cognitivist approaches deny that[6]. Although different cognitivists about judgements proposed different accounts of more complex relations between judgments and beliefs[7] that seem to be more suggestive than analytic, the important point is that according to the cognitivist position judgments are propositional and the propositions which are the objects of judgements are taken by the agents to have truth values. Thus, according to the cognitivist position, to judge A is at least – for a reflective agent - to believe that A is true.

My assumption in this thesis is that cognitivism regarding moral judgements is correct. In the literature one can find many arguments for and against moral cognitivism. In chapters 2 and 4 I will discuss some of these arguments and will defend moral cognitivism against some non-cognitivist objections that arise specifically in the context of the inquiry carried on here. For now, we will just take it as an assumption.

[6] Different non-cognitivist approaches take judgements to be different types of mental attitudes. See Van Roogen (2008) for a discussion.

[7] Kant, for example, took judgments to be a cognitive relation prior to belief and necessary for belief formation (see, for example Hanna 2004). Others take belief to be prior to judgments in the sense that when an implicit belief becomes conscious for an agent and he gets mentally committed to it, it becomes a judgment (see for example Roberts 1999).

Taking moral judgements to be beliefs, the question arises as to what kind of beliefs are they – qualitative or quantitative? There are three conceptually possible approaches here. On the one hand, one can argue that moral beliefs are always qualitative, i.e. that the mere idea of believing in a moral propositions to varying degrees is incoherent, meaningless or at least normatively insignificant. On the other hand, one can go the other way round and argue that moral beliefs (maybe just like any other belief) always come with degrees. That is, that the mere idea of having a binary belief is either incoherent or meaningless or at least normatively insignificant. Finally, one can admit that there are two kinds of beliefs, qualitative ones and quantitative ones and the two kinds should (so it seems) be related to each other in a systematic way.

The literature regarding qualitative vs. quantitative beliefs in non-moral contexts is very large. I will discuss some if it in chapter 3. When it comes to the moral context, however, not much was written about the matter[8]. In this thesis I will adopt the third possibility mentioned above. On the one hand, in this chapter and the next one, I will argue that thinking about moral beliefs in quantitative terms has many advantages but on the other hand, in chapter 3 I will argue that – specifically in the moral context – there is an important role for a binary concept of belief[9] that cannot be played by a quantitative beliefs.

[8] A notable exception is Michael Smith 2002 paper in which he argues for moral cognitivism on the basis of its ability to account for degrees of confidence (what Smith calls "certitude") in moral judgements.

[9] For convenience I will use the term "acceptance" in order to refer to the qualitative belief attitude. Also, "to accept judgement A" will be used as a shortening to "to accept the proposition which is the object of judgement A". In the same way, one's "degree of belief in judgement A" will be used as a shortening to one's "degree of belief in the proposition which is the object of judgement A".

For now, the important thing to notice is that the inconsistency that stands in the heart of Singer's argument must be – for a cognitivist – an inconsistency among qualitative beliefs. While it is irrational to accept all of the three claims presented in the beginning of this chapter, there is nothing irrational about attaching high subjective probability to each one of them.

Although rationality dictates assigning value 0 to the conditional probability p("wealthy people are not morally obligated to donate the money worth of a new pair of shoes for the purpose of saving dying children in under-developed countries, even when by doing that they will save, for sure, the life of one child" | "in the "child in the pond" story, the agent is morally obligated to save the child, even at the cost of ruining his shoes" ∩ "there are no morally relevant differences between the situation described in the "child in the pond" story and the situation described in judgement 2 above"), it is perfectly rational to attach high unconditional probability to all three claims.

In fact, it seems that ascribing this kind of probability distributions to people succeeds in capturing the mental attitudes toward the different propositions involved in Singer's argument, of many of them. It also succeeds in explaining why being exposed to this argument and even finding it very appealing do not usually lead people to change their expressed judgements about the matter.

So Singer's argument should be understood as operating on the level of acceptance, not on the level of quantitative beliefs and the same must hold for any attempt to avoid it by changing one's judgments regarding the moral

relevancy of different features of the two situations Singer uses in his argument. The inconsistency is supposed to be resolved by making some changes in the set of judgements we accept, not by denying that there was an inconsistency in the first place. Such attempts, I have mentioned, can be seen as an involvement in a process of achieving a reflective equilibrium. Indeed. The reflective equilibrium method, as usually characterized in the literature, works with binary attitudes, not with attitudes that come with degrees. In the next section I will suggest that reformulating the method in such a way that degrees of beliefs will get a definite role in it, can save the method from some recent objections. First, however, we have to introduce the method.

In the literature, the concept of reflective equilibrium is used in different ways by different scholars and there is some discussion regarding the question as to what is the best way to use it (for an overview see Daniels 1996). In this thesis, I am going to use the concept in a very general form that will be discussed in this section and the next one.

The first characterisation of the idea was made by Nelson Goodman. In his “Fact, Fiction and Forecast”, Goodman suggested this approach, without naming it “reflective equilibrium”, as a way to justify inductive inferences. Goodman’s idea was that it is possible to justify induction in the same way that we implicitly justify principles of deductive inference, which is, according to Goodman, “by their conformity with accepted deductive practice” (Goodman 1965, p.63).

The obvious objection that arises is that this argument seems to lead to circularity, as we justify the general rules of induction by their conformity with specific inductive inferences, which themselves are justified only by virtue of being special cases of the general rule of induction. Goodman is well aware of this circularity, but does not take it to be an objection to his approach. Indeed he regards it as a virtue: “The point is that rules and particular inferences alike are justified by being brought into agreement with each other. *A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.* The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either” (Goodman1965, p.64, Goodman’s Italics).

In the process that Goodman describes, one starts with a set of judgements (about both inference rules and specific inferences), performs some operations on them (i.e. changes some of them in such a way that will make the whole set consistent), and ends the process with a new set of judgements. It would be helpful to make a distinction between the method of achieving a reflective equilibrium, which Goodman refers to, for example, by the expression “*the process of* justification”, and the state of being in a reflective equilibrium, which Goodman refers to using, for example, the expression “the agreement achieved”.

In our context, two different but related questions can be asked regarding both the method and the state of (being in) a reflective equilibrium: why is it justified

to use this method as a method for moral reasoning? Can it guarantee that the set of judgements achieved by using the method will have motivational force for an agent? Let us start with the second question. Of course, the answer to this question depends on the exact interpretation one gives to the expression “motivational force”. Here I want to use a weak interpretation, according to which a judgement has motivational force for an agent if the agent believes he is justified in accepting this judgement, where by “accepting a judgement” I mean intending to act upon it. Thus, if for judgements to be in a reflective equilibrium, for an agent, is for them to be justified for her, they also have, according to this interpretation, motivational force for the agent that holds these judgements.

This interpretation might seem inadequate because, for most people, it constitutes neither a necessary nor a sufficient condition for a judgement to have a motivational force on them. An agent can find a judgement justified and still not be motivated by it as it is too demanding in the sense of asking the agent to sacrifice too much in terms of her personal interests, and an agent can be motivated by a judgement she finds unjustified, if this judgement is appealing to the her on other grounds. However, for my purposes here, using this condition will suffice. This is because, as mentioned in the Introduction, I am concerned here with an idealised moral agent who is motivated only by moral reasons.

For such an agent, as per the definition, accepted moral judgements are the only source of motivation and all accepted judgements are motivational and this

is true regardless of the question as to how intuitive these judgements are. It may be that just in the same way we that sometimes accept factual judgements that we intuitively (that is, on the basis of what our senses tells us) find implausible and, moreover, base our decisions on the assumption that these judgements are true, we should sometimes accept very unintuitive moral judgements and base our moral decisions on them.

Whether this is the case or not is, of course, an open question, until an answer to the question of whether the reflective equilibrium method is justified is answered. The point is, however, that if the method of reflective equilibrium is justified, then it seems that although sometimes our moral intuitions are inconsistent, we have a way to get rid of these inconsistencies while still keeping motivational the moral theories that we accept. They are motivational in the sense that we believe it is justified to accept them, i.e. that we are willing to act upon them.

So how can one justify the reflective equilibrium method? Goodman seems to suggest that the concept of a reflective equilibrium is an explication of the term “justification”, and so the judgements an agent accepts in a reflective equilibrium are justified on semantic grounds. Without elaborating too much at this stage, I wish to point out an implicit assumption used in the above argument: it does not follow directly from the claim that the judgements that the agent accepts in a reflective equilibrium are justified, that the method of reflective equilibrium is justified. One must assume also that the method of

reflective equilibrium can lead one to a state of reflective equilibrium. Later on I will question this assumption.

This is not, however, the problem that most critics of the reflective equilibrium method highlight. The main problem discussed in the literature concerns not the reflective equilibrium method, but the mere claim that for a judgement to be in a reflective equilibrium with all other judgements is for it to be justified. Being in a reflective equilibrium, it is sometimes argued (see for example Stich 1988), lacks some of the necessary features of being justified.

Of special interest to us is one version of this criticism that will be discussed in the next section.

Scepticism about intuitions and wide reflective equilibrium

The method of reflective equilibrium gives intuitive judgements a central role in determining the set of judgements an agent ends up accepting. Consistency is indeed a restriction, but other than that it seems, prima facia, that the only thing that has to be taken into account by an agent that is involved in the process is how intuitive she finds some judgements.

It is not at all clear how exactly the agent should choose which intuitions to reject and which to keep, when she has conflicting intuitions. Indeed, this question is one of the central questions I address in this thesis. However, even before one addresses this question, there is another challenge that must be

met. This challenge was explicitly presented as a criticism of the reflective equilibrium approach by Stich, Nichols and Weinberg (Weinberg et al. 2001) in the context of a theory of knowledge.

Weinberg et al. conducted a series of simple experiments that supported the claim that people's intuitive judgements regarding normative epistemic questions, for example about whether the young man in Dretske's Zebra-in-Zoo story knows or only believes that the animal he sees is a Zebra[10], are strongly influenced by variables that seem to be irrelevant to epistemological questions, from a normative point of view, for instance, by the cultural background of the subjects.

If this is the case, they argued, the "explication of justification" justification of the reflective equilibrium method, or of a wider set of methods they call "intuition driven romanticism"[11], seems to fail, as the justification itself is not in a reflective equilibrium with our intuitive judgements about which features a justification ought to have. Although they do not explicitly state this, it seems that they assume that one of these features must be that *if a judgement is justified then, for any true proposition 'p' that is normatively irrelevant to the question whether the judgement is justified, it would still be justified had not p*.

[10] In this story a young man, visiting the zoo, sees a zebra, points to it and say, "That's a zebra". We are told, however, that it is possible for the zoo authorities to disguise a mule to look like a zebra in such a way that this young man, had he seen it, would have thought that it was a zebra. Although this is possible, this is, we are told, not the case here and the animal that the young man pointed to is really a zebra. This story is, it is sometimes claimed, a counter example to the justified true belief definition of knowledge, as the young man's belief in this story is both true and justified, but intuitively, according to this argument, he does not know that the animal is a zebra, but merely believes it.

[11] That is any method that takes epistemic intuitions as input and produce normative claims as output in a way that depends on the specific intuitions that were taken as input.

I will shortly try to investigate more carefully how the argument can be stated more precisely, but before doing so, it is important to say something about what the argument is not. The argument is not that we cannot trust our intuitions because they sometimes mislead us. This argument is a bad one, everybody agrees. It assumes that we have some independent access to the truth against which we can test the validity of our intuitive judgements. The problem is, of course, that ultimately we have no direct access to the truth. This is exactly the problem that the reflective equilibrium method was supposed to solve: how to justify a theory when we have no Archimedean point to turn to.

The argument is rather a different one. But what is it exactly? Let us first examine an argument with a very similar structure; this time made in an ethical context. The argument has many versions, but I think the most philosophically developed one is the one made by Joshua Greene in his dissertation and a related article (Greene 2002 and Greene 2007)[12].

The argument starts with a descriptive claim about the causal mechanisms that are responsible for the production of our intuitive moral judgements. In Greene's version, this is the claim that our intuitive moral judgements can be produced by one of two causal mechanisms, one in which our emotional reaction to the situations we evaluate plays a significant role, and one in which this is not the case. What determines, according to Greene, which mechanism will be responsible for a specific moral judgement are some structural properties of the

[12] For another example see Baron 1995.

situation evaluated. For example, situations in which the act that has to be morally evaluated is close and personal to the evaluator tend to trigger the emotional mechanism, and situations in which the act is detached and impersonal are more likely to trigger the cognitive mechanism.

Greene presents an impressive body of evidence in support of this descriptive claim, but of course not all psychologists agree with him. Some present other dual-process accounts in which either the emotional mechanism (see Haidt 2000, for example) or the cognitive mechanism (see Bucciarelli et al. 2007 for example) has a more prominent role. Others (see Mikhail 2007, for example) are committed to a single mechanism that makes use of both cognitive and emotional inputs. In any case, one feature that is common to almost all of the prominent accounts is that they are committed to the claim that some of our intuitive moral judgements are produced by causal mechanisms which are influenced by variables we usually take to be morally irrelevant. In most, but not all, cases these are our emotional reactions to the situations we evaluate. For convenience we can use, therefore, Greene's specific version of this claim that was previously mentioned.

The second step in Greene's argument is the claim that one consequence of the operation of a dual process is that in some cases people will tend to produce intuitive moral judgements that are inconsistent with their second order moral judgements regarding which variables are morally relevant. This indeed seems to follow from the first claim, since if our intuitive moral judgements are causally sensitive to variables we judge to be morally irrelevant, then for no

inconsistencies to arise between the moral judgements we have and our judgements regarding which variables are relevant, an unlikely coincidence must happen.

Greene takes Singer's drowning child example to be a case in which such an inconsistency occurs[13], but he, as well as others, was able to design experiments that produced other such inconsistencies. Perhaps the most well known is the difference between the patterns of responses to the famous trolley problem when presented in two different versions. In the first version, an agent can save the life of 5 people who are about to be run over by a trolley by pushing a button that will cause the trolley to move to a side track where it will only kill one other person. In the second version, the agent can push a man onto the track, thus causing his death, but stopping the train from running over the 5 people. While 90% of the participants in the experiment (in Mikhail's 2007 report) held that it is permissible to push the button in the first version of the dilemma, only 10% held that it is permissible to choose the analogous option in the second.

Greene takes this pattern to support his dual-process account, as he argues that while in the first scenario, the act of pushing the button is detached and impersonal, and thus the cognitive mechanism is more likely to be triggered, in the second scenario the act of pushing a man is close and personal and thus the mechanism that is more likely to be triggered is the emotional one. Others (like Mikhail himself) disagree and provide different explanations. In any case,

[13] This is so since in the case of the drowning child the emotional mechanism is likely to be triggered, while in the case of dying children around the world, this is less likely to happen.

as explained, it seems that according to any of the serious candidates for being the right explanation for the phenomenon, what *causes* the difference in the responses to the two versions is not something that most people judge to be morally relevant.

The next step in Greene's argument is, thus, the claim that since what causes the apparent inconsistency is a variable which we judge to be morally irrelevant, then even if one succeeds in resolving the inconsistency by pointing to another difference between the scenarios, that one is willing to accept as morally relevant, this is just a post–hoc justification for one's judgements.

Here is how Greene makes the point: "…according to Judith Jarvis Thomson (1986, 1990) and Frances Kamm (1993, 1996)… there is a complicated, highly abstract theory of rights that explains why it is okay to sacrifice one life for five in the *trolley* case but not in the *footbridge* case, and it is *just so happen,* that we have a strong negative emotional response to the latter case but not to the former" (Greene 2006, p.68, Greene's Italics).

What is wrong with post-hoc justifications? one might ask. Well, if one believes that justifications have a constitutive role in ethics, than there seems to be nothing wrong with them. However, Greene's argument is directed at those who do not believe that. Specifically, it is directed at rational deontologists, regarding which Greene writes "They can't say that our emotional responses are the *basis* for the moral truth… because they are *rationalists.* So they are going to have to explain how some combination of biological and cultural evolution managed to

give us emotional dispositions that correspond to an independent, rationality discovered moral truth that is not based on emotions" (ibid, p.68).

I think that the similarities between Greene's argument and Weinberg's et al. argument are clear enough, but let me now try to point out the structure that is common to both of them. Both arguments play on the inconsistency among the following five (schemas of) claims:

1. Judgement A is justified.
2. Judgement A is justified on the basis of the strong intuition that A.
3. The intuition that A is caused by C.
4. The question of whether C or not C is normatively irrelevant to the question as to whether A is justified or not.
5. A necessary condition for a judgement to be justified is that, for any true proposition p that is normatively irrelevant to the question whether the judgement is justified, it would still be justified had not p.

Numbers 1 and 2 are judgements of the kind that whoever uses the reflective equilibrium method must make sometimes. For Weinberg et al. these will be the judgements, for example, that the agent in Dretske's Zebra-in-Zoo case only believes, but does not know, that the animal he sees is a Zebra and the judgement that this is a justified claim since one has a strong intuition that this is so. For Greene, these would be, for example, the judgements that the difference between the responses to the Fat-Man version and the Side-Track

version of the trolley problem is justified and the judgement that this is so because this pattern of responses is very intuitive.

Number 3 is a judgement regarding an empirical matter and the evidence presented by Weinberg et al. and by Greene was supposed to support this judgement. In order to generate an inconsistency between the five judgements one must, of course, be committed to an analysis of causality according to which, at least in the cases in question, if A causes B then had not A, not B[14]. The counterexamples to this condition are well-known and although I do not see how any of them should apply to the case of Weinberg et al. it does seem that they can be applied to the case of Greene.

This is so since one can argue that although the different emotional reactions the subjects experience in response to the two versions of the trolley problem cause the difference between their judgements concerning each one of the versions, had the emotional reactions been the same, the responses to the two versions would still be different because there are normative reasons for giving different answers to the two versions.

However, as Greene notes, in order to escape the inconsistency between the five judgements, one must insist that this kind of a reply will be available in every case of apparent inconsistency that is caused by different emotional reactions to structurally different situations, and - leaving God out of the picture - there seems to be no reason to expect that this will be the case.

[14] If one is committed to this, then the inconsistency arises in the following way. Had not C, one would not have the intuition that A, and thus (from 1 and 2) A would not be justified. However, this contradicts the conjunction of 4 and 5.

Numbers 4 and 5 are normative judgements. For Weinberg et al. judgement 4 will be is the judgement that whether one's cultural background is Eastern or Western, for example, is normatively irrelevant to the question of whether one is justified in the judgement that the agent in the Dretske case only believes - but does not know – that the animal he sees is a zebra. For Greene it would be, for example, the judgement that what emotional reaction one has to a specific situation is morally irrelevant to the question as to whether one is justified in the judgement that one ought to, or ought not to, push the man onto the track in the second version of the trolley problem.

Judgement 5 is a general normative claim regarding the nature of justification. It is best viewed, I think, as a necessary condition for normative relevance. That is to say that it requires that for any proposition p and judgement A, if 5 is violated by p in regards to A, then p is normatively relevant to A.

Now, Greene and Weinberg et al claim, independently of one another, that in the set of cases they point to, one ought to reject 2, keep 3, 4 and 5 and either keep or reject 1, based on other considerations. The remarkable thing about this claim is that it is based on a kind of argumentation that is at least very much like the reflective equilibrium method. Even if one takes 3 not to be supported (at some point) by intuitive judgements, but rather by some kind of intuition-free science, and even if one takes 5 to be supported not by epistemological or linguistic intuitions, but on the basis of some kind of intuition-free conceptual

analysis (two positions that I believe will be hard to maintain), still it is clear that 4 is accepted in virtue of its intuitive appeal.

Weinberg et al. seem to accept the last claim, as regarding the possibility of rejecting judgement 4 they only write "...that we take to be quite a preposterous result" (Weinberg et al. p.35) which seems to be just another way to say "we have a strong intuition that this is not the case". Greene's reaction to this possibility is quite similar. In a reply to Mark Timmons (Timmons 2007) that raised, among other things, the possibility of being a deontologist while accepting the claim that one's emotional reactions to different situations are morally relevant (what he calls "sentimentalist deontology"), Greene only writes that "Kant was opposed to emotion-based morality because emotions are fickle and contingent in oh-so-many ways... About that he was right" (Greene, 2007, p. 116). No further argument is given. Why then is it the case that "about that he was right", one might ask, and again it seems that the only possible answer Greene can offer is that it is so because it is highly intuitive that this is so.

I do not intend this to be a criticism of either Greene's position or of Weinberg's et al. position. Although I think a more explicit recognition of the implicit use they make of some intuitions would be appropriate, it is clear to me that their arguments are valuable. They teach us that when we are engaged in a process of achieving a reflective equilibrium we must make use, not only of our intuitive judgements regarding the specific question at hand, but also of other judgements: judgements that are based on scientific knowledge, second order

judgements about the relevance of different matters to the question at issue, and so on.

In the literature this is sometimes described as using a method of a “wide reflective equilibrium” (see Daniels 1979, Rawls 1974, for example). When taken to the extreme (i.e. when making use of all the judgements an agent has) this seems to be just a *characterisation* of any reasoning. Reasoning is, in a sense, just the process of achieving coherence among one’s judgements. Viewed in this way, the method of reflective equilibrium, while indeed justified on semantic grounds, seems somewhat trivial.

This point was raised by T.M. Scanlon (Scanlon 2003). Referring to the wide interpretation of the method he writes: “It becomes simply the truism that we should decide what views about justice to adopt by considering the philosophical arguments for all possible views and assessing them on their merits” (Scanlon 2003, p.151). Scanlon, however, does not take this as constituting a problem for the reflective equilibrium method. He admits that “This charge of emptiness seems… to be largely correct” (ibid), but points to two restrictions he believes the reflective equilibrium method does impose on moral reasoners, thus saving it from being vacuous.

I find Scanlon’s characterisation of these restrictions quite puzzling, so it is better to quote his exact words and then discuss them. Referring to the reflective equilibrium method, he writes “…the method is not vacuous because it is incompatible with some views about these sources. It is incompatible, first,

with the idea that any particular class of judgements or principles can be singled out in advance of this process as justified on some other basis and, second, with the idea that any class of *considered* judgements should be left out of this process…" (ibid, Scanlon's italics).

Both of these restrictions, it seems to me, require a reasoner *not to restrict,* in some specific ways, the set of judgements he uses in his reasoning. As such they save the *claim* that the reflective equilibrium method is the right method to use in moral reasoning from being empty, but they do not save the *method* from being merely a characterisation of any kind of reasoning. This still leaves us with the conclusion that the method has no bite.

As Peter Singer writes regarding Daniels' formulation of the concept, "That approach renders the model of 'reflective equilibrium' relatively innocuous by making it so all-embracing that it can include any grounds for rejecting intuitions." (Singer 2005, p.561). Triviality, however, does not imply falsity. In most cases, the opposite holds. Indeed, Singer's conclusion is that "In that form, there is no need to object to reflective equilibrium[15]".

Thus, adopting the wide interpretation of the reflective equilibrium method saves it from objections that are based on scepticism about intuitions but only for the price of making it biteless. However, it also highlights what is missing from the

[15] An observation: The phrase "in that form" in the quote seems to refers to the previous sentence that says "...Now, the 'data' that a sound moral theory is supposed to match have become so changeable *that they are no longer a barrier to the acceptability of utilitarianism"* (my italics). This seems to imply that Singer is willing to accept the reflective equilibrium method on the condition that it will not rule out utilitarianism. In other words, Singer's moral views come before his view regarding which methodology is appropriate for ethics. I will return to this point in the conclusion.

current characterisations of the reflective equilibrium method. This is, as was mentioned, a set of criteria that tells us how an agent, engaged in a process of achieving a reflective equilibrium, should choose which judgements to keep and which to amend. In order for the method of reflective equilibrium, understood in the general form presented here, to have any bite at all, we must add something to it. A set of consistency conditions regarding the process of accepting and rejecting judgements may play this role.

Inconsistency between first order moral judgements

The inconsistency among people's moral judgements that was discussed in the previous section was essentially an inconsistency between one's first order moral judgements and one's second order moral judgements. Specifically, it was an inconsistency between a set of judgements regarding which act is morally superior to the others in different situations and judgements regarding the moral relevance of different aspects of these situations.

As we have seen, this kind of inconsistency was used by Greene and others to argue against non-consequentialist approaches in ethics. This criticism usually comes accompanied by strong support of some consequentialist approaches. Thus, it seems that the critics assume consequentialism is immune to their criticisms.

The reason for this is, I think, quite simple: consequentialist approaches usually provide us with a very clear guide as to how to escape the kind of

inconsistencies discussed in the previous section by telling us exactly which aspects of a situation are morally relevant. However, I think that the emphasis in the literature on the implications of the findings in moral psychology discussed in the previous section on the consequentialism versus deontology debate, has prevented philosophers and psychologists from paying attention to a more fundamental problem that these and related psychological findings pose for consequentialist and non-consequentialist approaches alike.

This problem is the result of another kind of inconsistency among moral judgements: having sets of first order moral judgements that violate the axioms of Bayesian decision theory. There are many versions of Bayesian decision theory. For convenience, throughout this thesis, the discussion will be made mainly in the framework of Leonard Savage's (Savage 1972) version, which is still the most widely accepted theory among economists. Although from a philosophical point of view it has some limitations, I do not believe any of these limitations will play a role in this thesis[16].

Savage makes use of many assumptions in his book. Some of them are supposed to express conditions of rationality, while others play different roles, and it is not always easy to say which is which. In this context, it is convenient to concentrate on three requirements of rationality that the psychological literature usually deals with. These are the demands that the agent's beliefs will obey the axioms of probability (this is not, strictly speaking, an axiom Savage uses, but in this context it will be more convenient to take it as an axiom), that

[16] This claim will be discussed again in the next chapter.

the agent's preferences will be complete and transitive and that they will obey the Sure-thing Principle.

It is almost common knowledge today that, in spite of their normative appeal, most people systematically violate each one of these requirements. The psychological literature that investigates these violations is not unrelated to the psychological literature regarding moral judgements, discussed in the previous section. Indeed, referring to one of the central papers in this literature, Haidt (2000), Daniel Kahneman, one of the founding fathers of the psychology of choice literature, writes "…the psychology of judgement and the psychology of choice share their basic principles and differ mainly in content…" (Kahneman 2003, p.717), and suggests that although " A general framework such as the one offered here is not a substitute for domain-specific concepts and theories…" it is important to try and accommodate similar findings in different fields under the same conceptual framework as "…broad concepts such as accessibility, attribute substitution, corrective operations, and prototype heuristics can be useful if they guide a principled search for analogies across domains, help identify common process, and prevent overly narrow interpretations of findings" (Kahneman 2003, p.717).

Indeed, examining the psychological literature regarding judgements in the framework of the psychology of choice naturally leads one to suspect that we should also expect to observe some violations of the rationality axioms in the moral domain. This is not surprising, of course, as moral decisions are, after all, a special kind of decision, but the point I wish to make in this section is that

these violations cannot be avoided by adopting a consequentialist approach. The main reason is that although consequentialist approaches usually do tell us which aspects of a situation are morally relevant, they do not usually tell us how important each one of these aspects is (compared to the others). Thus, in situations in which the relative weight of different morally relevant considerations matters, consequentialist approaches will be in no better a position than non-consequentialist approaches in terms of their ability to deal with inconsistencies among first order moral judgements.

The paradigmatic example for such cases is a set of choices between alternatives with varying levels of independent morally relevant features. In the simple case, there are only two such features. These can be, for example, the number of lives saved and the quality of these lives, the quality of a life saved and the chances of it being saved, the quality of a life saved and the length of this life, and so on (and note that all of these values are consequentialist values).

In all such examples, although most people will probably find it easy to judge which one of two alternatives is morally superior, for most combinations of levels of the different features, they will find it very hard to say exactly how much better or worse one alternative is from another. Most people judge, for example, that saving the life of another person is better than slightly improving his wellbeing, but it is really hard for us to say exactly how much better it is. Thus, when we have to decide between saving the life of one person and

slightly improving the wellbeing of many people, we may reach a point at which it will become really hard for us to judge which of the options is morally superior.

In such cases, the psychology of choice tells us, we should expect to observe violations of Savage's axioms. To see why this is the case, something has to be said about the concept of heuristics, which usually goes together with the related concept of framing but - for our purposes - concentrating on just one of them will be sufficient. The concept was first used in the context of decision theory by Kahneman and Tverskey in a series of experiments they conducted during the 1970s. As Kahneman himself notes, during the early days of their joint research, no explicit definition of the term "heuristics" was offered and heuristics "…were described at various times as principles, as processes, or as sources of cues for judgement" (Kahneman 2003, p.707). In 2002, however, Kahneman and Frederick (Kahneman and Fredrick 2002) offered such an explicit definition they believe succeeds in capturing most of the cases described in the literature as "heuristics".

According to this definition, the term "heuristics" should be applied to processes in which an agent, who assesses a target attribute, does this by substituting this attribute by another attribute that is easier to assess. This definition highlights what Kahneman takes to be the main insight of the heuristics and biases literature: when people are confronted with a difficult task, they tend to approach it by focusing on the aspects of it that are more accessible to them.

For example, when people have to assess the probability of an event or a proposition, they sometimes substitute it by another attribute of the event or the proposition that is more accessible to them, that is, by how well they represent the relevant class of propositions or events (this is what is called in the literature the “representativeness heuristic”).

Over the years, many different heuristics have been identified and tested by psychologists, but as Kahneman writes, "The idea of an affect heuristic… is probably the most important development in the study of judgement heuristics in the past few decades" (Kahneman 2003, p.710). Since this kind of heuristics fits nicely with Greene’s dual process account of moral judgements, it will be helpful to use it in the remaining discussion. However, very similar stories to the one I am about to tell, can be told using other heuristics (and other accounts in the psychology of moral judgements). What is doing the job is not the specific attribute substituted, but rather the process of attribute substitution itself.

The idea of an affect heuristic was proposed by Slovic et al. (Slovic et al. 2007) as a way to integrate the findings of Antonio Damasio (Damasio 1994) regarding the role emotions play in decision-making into the heuristics and biases literature. Damasio studied the decision-making abilities of patients with damage to the ventromedial frontal cortices of the brain, which are responsible for people's emotions. These patients have perfect capacities to reason, remember and calculate, but still they achieve very poor results in decision-making tasks.

Damasio's explanation for this phenomenon employs the concept of "somatic markers". Somatic markers are "…feelings generated from secondary emotions…" that "…have been connected, by learning, to predicted future outcomes of certain scenarios" (Damasio 1994, p.174). Thus, when a normal agent takes a decision, his preference ordering is determined, at least in part, by these somatic markers that enable him to predict his experienced utility from different possible outcomes. When people lose the ability to use these markers, their decision-making ability is affected. Interestingly, the patients Damasio examined also performed very "badly" in moral judgement tasks (that is, they expressed judgements that are usually taken to be immoral).

Slovic et al. (2007) developed Damasio's ideas and suggested that people form their preference orderings by using affect heuristics, in which, in order to assess the level of desirability (or the amount of future experienced utility) of an outcome, instead of judging the outcome in light of a list of relevant attributes, they substitute these attributes with the emotional reaction associated with the description or image of the outcome. Using this idea, they were able to explain and predict phenomena from many different fields, including, most famously, the phenomenon of preference reversal.

The phenomenon of preference reversal can be demonstrated in the following way. Consider the following three decisions (In this example, I have placed the phenomenon in a moral context in order to highlight its relevance to this inquiry, but of course it is not restricted to moral contexts):

Decision 1: you are offered the choice between the following two lotteries. In case you win any of them, the money goes to charity. Lottery 1: a chance 0.95 to win £1,000, Lottery 2: a chance of 0.1 to win £10,000.

Decision 2: you are told that a lottery ticket that gives a 0.95 chance of winning £1,000 is donated to charity, but it is possible to return the tickets to the bookie and instead to transfer a sum of £x to the same charity. What is the minimum x such that you will be willing to accept the offer?

Decision 3: you are told that a lottery ticket that gives a 0.1 chance of winning £10,000 is donated to charity, but it is possible to return the tickets to the bookie and instead to transfer a sum of £y to the same charity. What is the minimum y such that you will be willing to accept the offer?

Experiments, such as those of Slovic and Lichtenstein (1974)[17], show that a significant proportion of the population prefer lottery 1 to lottery 2, while giving y a higher value than x. Since most people do prefer donating more money to donating less money, intransitivity occurs.

Slovic's and Lichtenstein's explanation for this phenomenon, which was later accommodated by Slovic et al. (2007) into the general affect heuristics framework, is the following one. When subjects are asked to give a money value for a bet, they use the affect heuristic and substitute the money value attribute with the affect attribute, but when they are asked to choose between

[17] This is one of the most studied phenomena in the psychology of judgement and decision making and in behavioural economics. The results are extremely robust. For a meta-analysis see moffatt and Bardsley (yet unpublished).

two bets, they do not use this heuristic and either choose by following another heuristic, or by using some kind of conscious reasoning. The difference between the two processes that are responsible for the two different tasks is what causes the reversal in the agent's expressed preferences (i.e. the fact that they choose a bet to which they assign a lower money value when offered two bets).

Of course, the operation of this heuristic seems to fit nicely with the claim that emotions play an important causal role in the production of moral judgements. It can be argued that when coming to assess the moral status of a given act, agents, who find this task difficult, use the affect heuristic and substitute the attribute of being morally wrong or right with the attribute of what kind of emotion the action produces. It does not matter why the agents find the task difficult. They may find it difficult because there are no such properties as wrongness and rightness of acts, or they may find it difficult because, although there are such properties, they are not directly accessible to us. The important point is that sometimes we do find this task difficult and, nevertheless, we are determined to perform it.

The claim that affect heuristics are responsible for many of our moral judgements fits nicely also with other findings in the psychology of moral judgements literature. An interesting example is a claim made by Monin, Pizarro and Beer, in two different papers (Monin et al. 2007, a and b). Monin et al. observed that in the psychological debate concerning which factor is more dominant in the production of moral judgements, reason or emotions, each side

bases its arguments on different kinds of experiments. While the "emotional camp" uses mainly experiments in which the subjects are asked to react to stories about moral choices other people make, for example to answer questions of the form "X did such and such. Is this morally wrong?", the "reason camp" uses mainly experiments in which the subjects are presented with a moral dilemma and asked to choose between the available actions.

This observation can be accommodated in the psychology of choice literature using the concept of affect heuristics in the following way. When confronted with a reaction question, what the subjects are asked to do is to assess the moral status of a given action, and in order to do so, they may use the affect heuristic and substitute the moral status attribute with the emotional reaction attribute. But, when they are confronted with a moral dilemma question, the subjects are required to do something else: to choose between two alternatives or to make a pair-wise comparison between them.

Thus, it might be that in the moral dilemma problems, the subjects are more prone to using conscious reasoning in order to come up with a judgement or, alternatively, they employ another heuristic, a cognitive one, in order to come out with a single judgement. In fact, this kind of an explanation follows almost step-by-step the explanation given by Slovic and Lichtenstein (1974) for the preference reversal phenomenon that was outlined above.

The similarities between the explanation offered by Slovic et al. for the preference reversal phenomenon and the explanation I have suggested for the

observations of Monin et al. leads to the natural hypothesis that we will find similar phenomena in moral contexts that do not involve placing money values on lotteries, as well[18].

Maybe another example will be useful here. In order to give further support to the claim that inconsistencies among people's comparative moral judgements are likely to occur when there is no direct access to the degrees of moral value of different acts, I will use an example that makes use of another (this time cognitive) heuristic identified in the literature, the "similarity-based decision making" heuristic.

Here is how Alex Voorhoeve (following Ariel Rubinstein) characterizes this heuristic:

"When deciding between multidimensional alternatives, say bundles of pain intensity and the time it must be endured (pi, ti) and (pj, tj), a decision-maker goes through the following three-stage procedure:

Stage 1: The decision-maker looks for dominance. If pi < pj and ti < tj, then bundle (pi,ti) is preferred to bundle (pj, tj).

Stage 2: The decision-maker looks for similarities between pi and pj and between ti and tj. If she finds similarity in one dimension only, she determines her preference between the two pairs using only the dimension in which there is

[18] Evidence for this can be found in a yet unpublished experiment conducted by Binmore and Voorhoeve (2008).

no similarity. For example, if pi is similar to pj while ti is not similar to tj, and ti< tj, then bundle (pi, ti) is preferred to bundle (pj, tj).

Stage 3: The choice is made using an unspecified different criterion." (Voorhoeve 2008, p.289).

Notice that such a decision rule falls under Kehneman's and Fredrick's definition for "a heuristic", as instead of answering the difficult question of "which one of the alternatives is better (or preferred?)", one answers a simpler question which is "which one of the alternatives is better (or preferred) along the dominant dimension". In other words, one substitutes the attribute of "being better" or "having greater value" with the attribute of "being better along the dominant dimension".

It is easy to see how using such a heuristic can lead one to accept intransitive comparative moral judgements. Consider for example the seven acts described in table 1below.

Dimension of moral value / Act	Total number of children going to school	Expected number of students continuing to university level education	Proportion of females
a	400	390	0.5
b	415	335	0.45
c	430	265	0.4
d	445	195	0.35
d	460	125	0.3
f	475	55	0.25
g	500	0	0.2

Table 1

Suppose that without any intervention each one of the children gets no access to any kind of school-level education and that all children that are sent to school do learn reading and writing, basic mathematical skills and so on.

Psychological experiments (for an example see Tversky 1969) show that a significant proportion of the population tend to express intransitive preferences in this kind of scenarios. These people prefer (or judge to be better) act a to act b, act b to act c, c to d, d to e, e to f, f to g, but also g to a. The explanation is straightforward. When comparing each one of the first six acts to its adjacent one, the difference along the “number of children sent to school” dimension

seems insignificant and thus people ignore this dimension when making the pair-wise comparisons. However, when comparing act g to act a, this difference does seem significant and as the number of children sent to school is judged by many people to be the most important dimension, they judge g to be better than a.

It is important to note that the reason that people use the heuristic in such cases is that the comparison they are asked to make is a difficult one. In what sense is it difficult? Well, it is difficult in the sense that in order to assess the level of moral value gained by performing each one of acts they have to make cross-dimensional comparisons of moral value[19]. Thus, the source of the observed inconsistency in people's judgements is the lack of access to exact degrees of moral value.

The conclusion of the last discussion is the following one. The psychology of judgements and decision making tells us that when people face hard questions they tend to substitute them with easier ones. When this happens, we should expect to observe violations of Savage's axioms. A very common "hard question" of this sort is the question of how valuable a specific act is. This question is hard either because it is hard to weigh against each other different types of value or because the agent has no direct access to the values of different acts even according to one dimension. Thus, when - in order to make comparative moral judgements – one must use such weighing or must make

[19] Note that these comparisons are not among the moral values different moral theories assign to the same act. Rather, it is among the moral value gained by one act according to different dimensions that are all judged to be morally significant by a (possibly implicit) theory an agent holds.

use of information regarding the exact levels of moral value of different acts, we should expect to see violations of Savage's axioms.

What should a morally motivated agent who accepts the rationality demand do when she realises that her intuitive moral judgements violate this demand, for example, when she realises she has intransitive moral judgements? The natural answer is that she has to change some of these judgements. As argued, doing that by using the reflective equilibrium method, defined in the broad way suggested in the previous section, does not mean giving up the motivational demand. However, since no consequentialist theory gives us a complete guide for assigning exact degrees of moral value to available acts, it seems that in the face of this kind of inconsistency, consequentialists have no more resources to use in the reflective equilibrium method than non-consequentialists.

First formulation of the problem

First step: An agent finds out that his intuitive moral judgements violate some of Savage's axioms. For convenience, let us assume that he finds out that his judgements are intransitive. Thus, concerning three acts a, b, and c he judges a to be morally superior to b, b to c and c to a. The agent is, however, committed to the rationality demand, that is, he holds the second order judgement that his first order moral judgements ought to be transitive.

Since I am interested in the possibility of reconciling the rationality demand with the motivational one, we can assume that this second order moral judgement is not one which the agent is ready to give up, under any circumstances.

Second step: The agent, realising that his system of judgements as a whole is inconsistent, looks for ways to change some of these judgements. Specifically, he uses his judgements, derived from the psychological knowledge he has, regarding the mechanisms that are causally responsible for his moral judgements, as well as any other set of evidence he finds relevant to the question (for instance he takes into consideration the opinions of people he respects on the matter, he reads some philosophy books and so on), to see if he has reasons to reject one of his initial moral judgements.

Third step: After using all such information, and changing some of his initial moral judgements, he still finds himself in a position in which some of his judgements are intransitive. The discussion in the previous section was supposed to show that this is not only possible, but also very likely to happen in situations where more than one aspect, which is judged by the agent himself to be morally relevant, plays a role.

The agent, trapped in this situation, knows that something is wrong with his set of judgements as a whole, but he does not know where exactly to put the blame. He knows he has to change some of his judgements, but he does not know which one(s). This description sounds to me like a description of someone

that suffers from uncertainty: the agent knows that he has to change some of his judgements but he is uncertain which ones.

This is the way I choose to describe my mental state when thinking of myself in this kind of situation, and I believe the same holds for many other people. This is not a good enough reason, though, to actually adopt the technical concept of uncertainty when discussing these mental states. In the next chapter I will discuss this issue more seriously and will present arguments for the claim that it is appropriate to do so, but for now all I want to do is to push the intuition a little more.

Consider the matter differently: an agent realises that a moral judgement that he holds, regarding what is the right thing to do in a specific choice situation, conflicts with a moral rule he endorses. Thus, he decides either to modify the moral rule or to reject the moral judgement. Whichever of these two options he chooses, it seems reasonable, for the agent, to ask himself: have I made the right choice? It also seems reasonable for him to be certain, to varying degrees, of the answer he gives to himself. It does not matter what the agent takes to be constitutive of the “right” answer to this question. As long as he does take such an answer to exist, he can form beliefs regarding what the right answer to the question is and these beliefs can come in degrees.

If it is indeed appropriate to describe the mental states of agents in such situations in this way, then the problem of reconciling the motivational demand with the rationality demand is reduced to the problem of presenting an account

of dealing with the kind of uncertainty from which agents suffer in these situations.

In a reflective equilibrium that respects Savage's axioms, both the motivational demand (as in my definition) and the rationality demand are satisfied. Now, the problem is to find a method to reach this state. At the end of the second section, I argued that in order for the method of reflective equilibrium to have some bite, some conditions of consistency on the way one changes one's judgements must be specified. Now, we might have the resources to formulate such conditions: they should make (some) use of the agent's degrees of beliefs in his judgements.

Even though it has never (as far as I know) been discussed at length in the reflective equilibrium literature, this idea may be what many of the defenders of reflective equilibrium had in mind when they wrote about the matter. Thus, when Nelson Goodman writes: "The process of justification is the delicate one of making mutual adjustments...", adjusting one's degrees of belief may be what he is referring to by "delicate" and by "mutual adjustments".

Scanlon (2003) also made some remarks that can be understood on this line. Thus, in several places in his discussion, he describes the attitudes that a reasoner, who is involved in a process of achieving a reflective equilibrium, has toward different considered judgements using quantitative terms like "confidence" (p.139) and "uncertainty" (p.144), and at one point (p.148) he explicitly argues that the interaction between considered judgements and

possible general moral rules is not decisive. When this interaction is not decisive, it is plausible that the agent will be uncertain about which judgements he should reject and which judgements he should accept.

Norman Daniels was more explicit in his commitment to the use of degrees of belief in the reflective equilibrium method, but he did not develop the idea at all beyond simply stating it. In his 1979 paper he wrote (referring to Richard Brandt's characterisation of the reflective equilibrium method) "We begin with a set of initial moral judgements or intuitions. We assign an *initial credence level* (say from 0 to 1 on a scale from things we believe very little to things we confidently believe). We filter out judgements with low initial credence levels to form set of considered judgements. Then we propose principles and attempt to bring the system of principles plus judgements into equilibrium, allowing modifications wherever they are necessary to produce the system with the highest over-all credence level" (Daniels 1979, p.268, Daniels' italics).

In a footnote, Daniels added "Presumably, we could use fairly standard treatment of degree of belief, rooted in probability theory, to formalize what is sketched here. This formalization might give particular content to the assumption that persons are rational, imposing certain constraints on revisability and acceptability..." (ibid, p.268, footnote 18). However, Daniels only mentioned this idea as an introduction to a discussion of the justifiability of the reflective equilibrium method and assumed that this kind of formalisation can be carried out. In chapter 3 I will try to actually do what Daniels has proposed.

Conclusion

Here is an overview of what I have done in this chapter: from the general observation that sometimes people hold inconsistent judgements, I moved to a discussion of the concept of reflective equilibrium and explained how, in light of the observed inconsistencies, this concept can help us to understand the motivational demand more clearly. Specifically, it was argued that just in the same way that we sometimes accept extremely unintuitive judgements regarding factual matters and act on them, we can sometimes accept extremely unintuitive moral judgements and act on them. Thus, the motivational demand is properly understood not as a demand regarding the level of intuitiveness of the recommendations of a moral theory, but rather as a demand regarding their fit with the set of judgements we believe to be justified, all things considered.

I then moved on to discuss an important criticism of reflective equilibrium, which helped me present a more pluralistic interpretation of the concept, in line with the wide reflective equilibrium approach. According to this interpretation one's moral judgements should be consistent not only with each other, but also with one's other judgements. Under this interpretation, it was argued, the concept of reflective equilibrium is so broad that it can be seen as a characterisation of any kind of reasoning: reasoning as a process in which one aims to achieve coherence among one's judgements.

I have raised the question, which I believe has not been adequately addressed in the literature, of which consistency conditions, if any, should guide an agent

in his decisions to reject or accept specific judgements when he is involved in a process of achieving a reflective equilibrium. Providing a satisfactory answer to this question, I have argued, may help to add some bite to the reflective equilibrium method while still keeping it as an uncontroversial characterisation of reasoning.

Having reached this point, I stopped pursuing this line of investigation for the time being, and moved on to discuss what I have argued is the most troubling kind of inconsistency among people's moral judgements. Using some insights from the psychology of choice literature, I have argued that this kind of inconsistency, namely the violation of Savage's axioms by people's judgments regarding what ought to be done in some situations, cannot be avoided merely by taking account, in the process of achieving reflective equilibrium, of some factual judgements. I have pointed out the fact that according to existing psychological knowledge, we should expect this kind of inconsistency to occur particularly in situations when the agent does not have a direct access to degrees of moral value. This happens, for example, when the choice involves more than one morally relevant aspect.

When an agent who is involved in the process of achieving reflective equilibrium finds himself facing this kind of inconsistency, it is natural, it was argued, to claim that this agent suffers from *uncertainty* regarding the question which judgements should he accept and which judgements should he reject. This claim was not, though, fully defended. This conclusion, however, brings us back to the point we were at the end of section 2, only now we have more material to

work with in order to look for consistency conditions for the method of reflective equilibrium: the agent's degrees of beliefs in his judgements.

In the next chapter I will discuss the claim that it is appropriate to describe such an agent as being uncertain regarding his judgements, and will examine some implications of accepting this.

Chapter 2: Moral Uncertainty

Introduction

At the end of the previous chapter, I suggested that it is appropriate to characterise as suffering from uncertainty an agent who realises that she has intransitive moral judgements and who is determined to change some of them in order to restore transitivity, and who is unable to do so by incorporating psychological knowledge into her reasoning. Following Ted Lockhart (Lockhart 2000), I will use the term "moral uncertainty" to refer to this kind of uncertainty: it is uncertainty regarding moral claims. I have also suggested that we can then use the agent's degrees of beliefs in the propositions that are the objects of her judgements in order to formulate consistency conditions on the way she chooses which judgements to keep and which to amend.

The first part of the claim nearly amounts to a commitment to moral cognitivism, the thesis that moral judgements are beliefs[20]. The second part amounts to a commitment to anti-Humeanism, the rejection of the Humean thesis that beliefs cannot constrain desires in any way that is not captured by standard rationality axioms[21]. Why exactly this is the case will be discussed in the second section.

[20] "nearly" because one can reject moral cognitivism, but accept the weaker thesis that we can treat moral judgements as beliefs in the sense of assuming they come in degrees that respect the laws of probability. This weakening of the claim, however, will play no role in this thesis.
[21] This is at least one way the Humean position is formulated. In particular, this is the formulation implicit in Lewis' discussion of the Desire as Belief Thesis that will concern me in section 1. I suppose some philosophers that call themselves Humeans will be willing to settle for a less restrictive formulation. Such Humeans, however, have no special reason not to accept my suggestion and so there is no need to argue against their positions. For some discussions regarding what the Humean position is see Lewis 1988, Smith 1987, Broome 1999, Chapter 5, Rosati 2006.

As mentioned in the introduction, neither of these commitments is uncontroversial. The task of defending each one of them against all the objections one can find in the literature is not one that I can hope to achieve in this investigation. However, it will be appropriate to discuss one objection that arises specifically as a result of accepting the rationality demand. This objection, suggested by David Lewis, applies directly to anti-Humeanism, but some scholars have also taken it to challenge moral cognitivism.

In the first section of this chapter, I will present Lewis' argument and discuss some of the anti-Humean and cognitivist replies to it. Lewis' argument has the form of a triviality result. Lewis formulated an apparently plausible anti-Humean thesis and showed that it is consistent with the rationality demand only in trivial cases.

The general structure of all of the replies to Lewis' argument is identical. It is argued that Lewis' attack indeed succeeds in showing that a specific anti-Humean thesis is incompatible with standard rationality axioms, but that this specific thesis is implausible in any case. Thus, the conclusion is, there is still hope for other anti-Humean theses.

In the second section I will argue that although all of these replies are successful in blocking Lewis' attack, none of them points in the direction of a specific anti-Humean thesis that is applicable in the context of this investigation, i.e. that can be used as a guide for an agent who realises he holds inconsistent

moral judgements and wants to change some of them in order to restore consistency.

The discussion in this section will be undertaken in the context of another literature that has discussed uncertainty regarding moral questions, and that has evolved independently of the literature that discusses Lewis' result. Unlike the latter, in which the discussions were mainly made in the framework of Richard Jeffrey's decision theory, the former seems to be implicitly committed to Leonard Savage's framework. I will explain the significance of this difference and its implications for my project.

The conclusion of this chapter will be that what is needed in order to progress on the route I have pointed to at the end of the previous chapter is an anti-Humean thesis that makes use of an agent's degrees of beliefs regarding comparative moral judgements *and makes use of them only*. Specifically, one should not make reference to different hypotheses regarding degrees of moral value.

The Desire as Belief Thesis Controversy

David Lewis (1988, 1996) presented, only in order to reject, an anti-Humean thesis he called "The Desire-as-Belief" Thesis (DBT). The thesis is formulated in the framework of an atomistic version of Richard Jeffery's system, as introduced in his 1965 book. For our purpose, as we will soon see, the relevant feature of this system is that it uses the same kind of objects, i.e. propositions, to be the

carriers of both desires and beliefs. This is in contrast to Savage's system, in which states of the world are the objects of beliefs and consequences are the objects of desires.

According to the version of this thesis that is usually discussed in the literature, an agent's desire for a proposition, A, should be[22] equal to his degree of belief in another proposition, A*, that can be interpreted as the proposition that says that A is good or desirable, and this should be so after any redistribution of his degrees of belief over the set of all propositions[23].

Lewis showed that this thesis is consistent with another requirement that he found reasonable, the invariance requirement (IR), only when the agent's degrees of belief in A* or in A are either 1 or 0. The IR says that the agent's degree of desire for a proposition, A, should not change after his degree of belief in A changes. Intuitively this means that one's desire for A is independent of one's belief in it.

[22] In the first section of his1996 paper, Lewis made some remarks that seem to imply that he took the DBT to be (also) a descriptive thesis, i.e. that it is not (only) that an agent's degree of desire for A should be equal to his degree of belief in A* in order for him to be rational, but also that it is the case for (typical) people. The descriptive interpretation is, however, implausible. Descriptively, people's degrees of belief do not obey the probability axioms (and the other axioms of decision theory) even regarding non-normative propositions and there is no reason to assume that particularly when it comes to normative propositions, they will start behaving more rationally. Thus, in the current discussion I will stick to the normative interpretation which is, in any case, the one that most scholars that have discussed the DBT seem to assume.

[23] Although Lewis did not explicitly discuss it, the idea behind the demand that the constraint will still hold after any redistribution of the agent's degrees of belief is presumably that if one takes some constraint on an agent's attitudes at a given point in time to be normatively appealing, then after the agent changes some of these attitudes in a normatively permissible way (whatever the norm is: rationality, morality or anything else) the constraint must still hold, as a set of attitudes that is normatively permissible that has been updated in a normatively permissible way must lead to a normatively permissible set of attitudes. The same idea plays a significant role in Lewis' discussions of the Principal Principle and Adam's thesis (see Lewis 1976 and 1980). I will return to this point later on in this section.

To see why this is the case, let u(A) denote the agent's desirability for A, and p denote a probability distribution over the set of all propositions and notice that IR together with DBT implies that A and A* must be probabilistically independent, since from DBT we get u(A)=p(A*) and from IR we get u(A) = p(A*│A) and so p(A*) = p(A*│A) . However, if both A and A* are above 0 and below 1 and the agent learns, for example, that B = ¬(A∩A*) then his new probability distribution after learning B, p', gives p'(A*│A) = 0 and p'(A*)>0, which contradicts the IR. But p' was obtained from p by Bayesian updating and so DBT is violated[24].

Granted that one accepts Jeffery's decision theory and the IR condition, Lewis' result can be interpreted in three different ways when it comes to the moral domain. According to the first interpretation, it implies that although moral judgements (of the form "A is good" or "A is right") are (or might be) beliefs, one cannot rationally hold degrees of beliefs that are different than 0 or 1 in the propositions that are the objects of these judgements, i.e. that moral uncertainty (regarding such propositions) is impossible. According to the second

[24] It is worth mentioning that the thesis' apparent commitment to degrees of desire that range from 0 to 1 is not what at issue here. It is easy to see that the same result holds in the more general case in which an agent has several hypotheses (to which he gives a positive probability) regarding the degree of desirability or goodness of a proposition.

In such a case the thesis will be formulated in the following way:

$u(A) = \sum_X p(\text{'}g(A)=x\text{'})x$

Here g(.) is a function that assigns to propositions degrees of goodness or desirability. The thesis says that the degree of desirability the agent ought to assign to a proposition is equal to the expected goodness of the proposition (relative to the agent's degrees of belief).

It is easy to see that the thesis introduced earlier is a special case of this version. All that is needed is to assume that there are only two possible degrees of goodness, and these can be normalised to 0 and 1. For simplicity, we can stick (most of the time) to this simple case, since it makes things easier to follow and nothing really hangs on it.

interpretation, it implies that although moral judgements are beliefs (which might be non-trivial), they cannot constrain, for a rational agent, the degrees of moral value he attaches to different acts, i.e. that this version of anti-Humeanism is false. According to the third interpretation, it implies that moral judgements are not beliefs, i.e. that moral cognitivism is false.

Although, it is not entirely clear which one of these interpretations Lewis had in mind, it seems most likely that it was the second one, as he clearly indicates that he takes the result to refute the anti-Humean position. Others including Oddie (1994) and Weintraub (2007) seem to adopt the third interpretation. Their rationale, although not explicitly indicated, seems to be something like the following; if moral judgements are beliefs, possibly with non-trivial degrees, then it must be that they constrain, for a rational agent, desires in some way or another. This is so since normative beliefs, by their nature, are action-guiding (at least for an ideal agent) and, for a rational agent, actions are connected to desires in a systematic way. Thus, at least through their mutual connection with actions, desires and normative beliefs do constrain each other. Now, by modus tollens, if one accepts the second interpretation for the result, i.e. if one accepts that the result shows that anti-Humeanism is false, one must also accept the third interpretation, i.e. one must also accept that moral judgements are not beliefs.

This argument depends, of course, on the assumption that normative beliefs are, indeed, necessarily action-guiding. As mentioned in the Introduction, I am sympathetic to this assumption. However, it surely stands in need of

justification. The question as to whether it is a justified assumption or not will not be discussed here[25]. The point I want to make, however, is that it is possible, but not necessary, to trace the failure of the DBT back to the moral cognitivism it is committed to. I will return to this point later on.

In any case, by adopting any of the three interpretations, one must reject either my first claim, i.e. that it is justified in some cases to characterise rational agents as suffering from uncertainty regarding normative propositions, or my second claim, i.e. that we can use the agents' degrees of beliefs in normative propositions in order to formulate consistency conditions on the way they determine their moral preferences.

It is clear why this is the case regarding the first and third interpretations which pose a direct threat to my first claim. These two interpretations, however, are not dictated by Lewis' result. One can be a cognitivist that accepts that moral beliefs can be non-trivial and still be a Humean by rejecting the claim that normative beliefs are, by their nature, action guiding. Thus, in order for me to defend my two claims against Lewis' argument, it is enough to attack the second interpretation. A survey of how this has been addressed in the literature will follow. In the fourth chapter, I will go back to the first and third interpretations and will argue that even a non-cognitivist, who accepts that normative beliefs can be non-trivial, can accept my two claims, with slight modifications.

[25] Although, as explained in the Introduction, this thesis is at least partly about trying to answer the question whether normative beliefs of a special kind (comparative moral judgements) *can* be action guiding in a strong sense.

What about the second interpretation, then? To see why this interpretation poses a threat to my second claim, notice that for a rational moral agent, moral preferences must accord with the degrees of moral value he attaches to propositions, in the sense that he ought to prefer any proposition to which he attaches a higher level of expected moral value to any proposition to which he attaches a lower level of expected moral value. In other words, the degrees of moral value a rational moral agent attaches to different propositions constrain his moral preferences. The DBT can be seen, thus, as an attempt to give some substance to my second claim; it is an attempt to formulate, using degrees of belief in moral propositions, a constraint on one's moral preferences.

Lewis has shown that this specific attempt fails, at least if one accepts the IR, but he also made a stronger claim according to which the reason for the failure is not the specific form that this attempt takes, but rather the mere idea that beliefs can constrain desires, i.e. the whole anti-Humean position. The general idea is quite simple: if normative beliefs constrain desires, then instead of using desires as a guide for decisions, one can use one's beliefs regarding the appropriate desires to have in light of one's normative beliefs for the same purpose. However, in decision theory, beliefs and desires behave differently[26]; they have different constraints operating on them. Thus, when trying to reduce one to the other, we should expect to lose something. We should expect to lose all that is gained from the interaction of these two different mathematical objects, which will probably mean that we will have to give up on at least some of the axioms of decision theory.

[26] In Lewis' result the feature that does most of the work is that they are updated differently: beliefs are updated according to Jeffery's conditionlisation (of which usual Bayesian updating is a special case) and desires are updated in accordance with Jeffery's desirability axiom.

This, of course, is not an argument, but I believe it is the kind of general consideration that led Lewis to try to formulate the anti-Humean position so that a real argument against it can be put forward. He did that by choosing the DBT as a plausible formulation of the anti-Humean position, but he was well aware that refuting the DBT does not amount to refuting anti-Humeansim. In his 1996 paper, he extended his argument to other possible formulations of the anti-Humean position, but admitted that although "A systematic survey of all possible versions, including versions not yet invented, would be nice", he is, of course, unable to provide one and, thus, "...we shall settle for less" (Lewis 1996, p.307).

In order to save the anti-Humean position, therefore, one must place the blame for the failure of the DBT, as well as the other versions Lewis considered, on some specific feature of it that is plausibly not shared by all possible formulations of the anti-Humean position. In the literature this has been done in several different ways. A good place to start in order to consider some of these attempts and to evaluate them is John Broome's short discussion of Lewis' result (Broome 1991).

Broome begins his discussion with the introduction of another thesis, "the Desire as Expectation Thesis", which says the following:

"Suppose there are a number of degrees of goodness, g_1, g_2, and so on, and let G_j be the proposition that our world is good to degree g_j. The expected goodness of any proposition *A* (the expectation of good from *A)* is

(1) $g_1 p(G_1 \mid A) + g_2 p(G_2 \mid A) + \ldots,$

where $p(G_1 \mid A)$ is the probability of *Gj* conditional on *A.* (This conditional probability is defined as the ratio of probabilities *P(Gj & A)/P(A)).*Teleology, let us suppose, says a rational person desires *A* to a degree equal to (1). Let us call this the *Desire-as-Expectation Thesis."* (Broome 1991, p.398, Broome's italics).

Both Humeans and anti-Humeans can accept the "Desire as Expectation Thesis" (DET), claimed Broome. The difference between them is that while the Humean takes the degrees of goodness, to which Broome refers by the different g_is, to be constituted solely by the agent's desires, the anti-Humean rejects the necessity of this claim.

To see that the DET is not necessarily an anti-Humean thesis, note that (by elementary manipulations of Jeffery's desirability axiom) the DET is equivalent to the following thesis: for all propositions G that say that the world is good to degree g, and for any other proposition, A, $d(G \cap A) = g$. This formulation of the thesis makes it clear that it says nothing about beliefs and so, it has no relevance to the Humean/anti-Humean controversy.

In the simple case when there are only two possible degrees of goodness, either 0 or 1, the DET can be reduced to the following formula:

$U(A) = p(G \mid A)$

Here, G is the proposition that says that the world is good or desirable. The expression $p(G \mid A)$, claimed Broome, does not refer to the agent's degree of belief in any proposition. It is true, of course, that it refers to a conditional probability (the probability of the proposition "the world is good" conditional on A obtaining), but, according to Lewis (1976) himself, conditional probability is not a probability of any proposition. Thus, even in the simple case, a commitment to the DBT does not follow from a commitment to the DET.

The failure of the DBT, claimed Broome, is not, therefore, to be blamed on the anti-Humean position generally, but rather on a feature of the version of anti-Humeanism that Lewis chose. This feature is, according to Broome, the commitment of such positions to yet a third thesis, which is usually called in the literature "Adams' Thesis" (see Bradley 1999 and 2000 for a discussion and an overview of the literature), according to which the conditional probability of one proposition A, given another proposition B, is equal to the probability of the conditional $B \rightarrow A$.

If Adams' Thesis is correct (and if conditionals express propositions) then, at least in the simple case, the DBT can be derived from the DET, since from

Adams' Thesis we get $p(G \mid A) = p(A \rightarrow G)$ and from the DET we get $U(A) = p(G \mid A)$ and so by stipulating $A^* = A \rightarrow G$ we get the DBT[27].

This is a bit hard to swallow, as we are asked to believe that the question as to whether Adams' Thesis is true or false, and whether conditional express propositions or not, can settle the question of whether anti-Humeanism is possible or not. It follows from Broome's analysis that if he were convinced that Adams' thesis is true, and that conditionals express propositions, he would stop being an anti-Humean (which he claims to be; see Broome 1999).

This is, no doubt, a radical conclusion and I believe this is so regardless of the question as to whether anti-Humeanism is true or not. Both Humeans and anti-Humeans should resist the claim that what they are really arguing about is the nature of conditionals. Indeed, I think they are not obligated to accept this, since Broome's extremely useful analysis of Lewis' argument, using the DET, points to other possible directions one might take in order to avoid Lewis' result.

What Broome's DET suggests is that it is not the case that one's degree of desirability for a proposition should be equal to the expected goodness or desirability one attaches to the proposition, but rather, that it should be equal to the expected goodness or desirability one attaches to the world, in case the proposition is true.

[27] Of course, one can accept the DBT without accepting Adam's Thesis. The point is, though, that it is possible to accept the DET without accepting the DBT by rejecting Adams' Thesis.

The shift from the former claim to the latter expresses, I believe, a shift in points of view. While the former claim seems like a natural assumption to make when thinking of the issue in the framework of Jeffrey's system, the latter seems natural from the point of view of Savage's system. In fact, the DET is best viewed, I think, as an attempt to express in Jeffrey's system the only possible treatment of moral uncertainty (and uncertainty generally) of which Savage's system (with no modifications) is capable.

In order to see this more clearly, let us formulate in Savage's framework the simple case of moral uncertainty that we have discussed, i.e. when there are only two possible degrees of goodness, either 0 or 1. We can also assume, to make things even simpler, that there is no uncertainty regarding any other issue, only regarding degrees of goodness.

In Jeffrey's system, propositions play roles that are played by three different mathematical objects in Savage's system. First, they play the role of acts in the sense that they are the objects of the preference relation. While in Jeffrey's framework, preferences are determined by the desirability levels of the propositions - in the sense that a proposition with a higher level of desirability is preferred to one with a lower level - in Savage's system preferences are determined by the expected utility of acts. Thus, what we should seek is the expected utility of an act under conditions of moral uncertainty. Let us consider two such acts, a and b.

Second, propositions play the role of events in the sense that they are the objects of uncertainty. Since we have assumed that the only uncertainty the agent suffers from is uncertainty regarding the levels of goodness of the acts and since we have assumed there are only two such levels, either 0 and 1, we have to consider two possible events for every act: “a is good”, “a is not good” and “b is good”, “b is not good”. Events, in Savage’s framework, are sets of states, thus we have to consider four possible states (no more, since we have assumed that uncertainty regarding the levels of goodness of the acts is the only uncertainty the agent suffers from): a state in which both a and b are good (that is, the degree of goodness of both acts is 1), a state in which a is good and b is not, a state in which b is good and a is not, and a state in which neither is good.

Thirdly, propositions play the role of consequences in the sense that they are the objects of desirability, or “utility” in Savage’s terms. In Savage’s framework, there is a consequence for every act in every state. Since we have assumed there are only two possible degrees of goodness, either 0 or 1, we can summarise the utility distribution over the set of consequences in the following table:

	ω_1 (a good, b good)	ω_2 (a good, b not)	ω_3 (a not, b good)	ω_4 (a not, b not)
a	1	1	0	0
b	1	0	1	0

Table 2

Now it is easy to see that the expected goodness of act a is $p(\omega_1) + p(\omega_2)$ and the expected goodness of act b is $p(\omega_1) + p(\omega_3)$. Notice also that since the event "a is good" is the set $\{ \omega_1, \omega_2 \}$ and the event "b is good" is the set $\{ \omega_1, \omega_3 \}$, the probability of the former is equal to the expected goodness of a and the probability of the latter is equal to the expected goodness of b. Thus, it is true in this case, under any probability distribution, that the expected goodness the agent attaches to an act is equal to his degree of belief that the act is good, and yet the degrees of belief can be non-trivial. Thus, it seems that for an agent that takes utility to be goodness, i.e. to the idealised morally motivated agent we are interested in, Lewis' result does not hold in Savage's framework[28].

This has led many scholars to put the blame for the failure of the DBT, not on the thesis itself, but rather on Jeffrey's framework. This was done in two different ways. Firstly, some scholars have pointed to the fact that while in Savage's system consequences have utility, they do not have expected utility; only acts have expected utility. However, in Jeffrey's framework, in virtue of his

[28] See Oddie (2001), Byrne and Hajek (1997) and Weintraub (2007) for similar points. Also, notice that the IR condition cannot be expressed in Savage's framework since act a is not a part of the events' algebra. An attempt to introduce it into the algebra will necessarily violate one of Savage's axioms. This will be discussed in more detail soon.

desirability axiom, the desirability of a proposition equals the proposition's expected desirability. This feature makes desirability behave in a way that is quite different to the way that utility behaves in Savage's framework.

Specifically, it means that if a proposition is preferred to its negation, as the probability of it increases other things being fixed, its desirability decreases. To illustrate this point, consider Broome's DET in the simple case and notice that: $d(A) = p(G \mid A) = p(A \mid G)p(G)/p(A)$. Keeping both the agent's degree of belief that the world is good and his degree of belief in A conditional on the world being good, fixed, as the agent's degree of belief in A increases, he desires A less and less.

This feature can be interpreted as a consequence of the view that is sometimes attributed to Jeffrey[29] according to which propositions should be viewed as "news items": as the agent's degree of belief in A increases, A becomes less of a news for the agent and so he desires it less and less.

However, argued Piller (2000), Weintraub (2007), and Daskal (2010), in slightly different ways, for this reason Jeffrey's concept of desirability cannot serve the role of goodness, or moral value. The moral value or the level of goodness we attach to propositions stays the same regardless of how probable we think they are. So the problem with the DBT, so they argued, is not the anti-Humean position to which it is committed, but rather the specific version of anti-

[29] Jeffrey has explicitly presented this view as one possible way – suggested to him by Savage – to interpret his 'propositions'.

Humeanism to which it is committed in virtue of its commitment to Jeffrey's framework, according to which goodness always equals expected goodness.

Some scholars have made use of another difference between Jeffrey's and Savage's frameworks in an attempt to "save" the DBT. In Savage's framework, acts are not the objects of uncertainty; only states are. However, in Jeffrey's framework, the analogue mathematical objects are, again, propositions and they are the object of uncertainty. This feature of Jeffrey's framework allows agents to treat their acts, the decisions they make, as evidence and to update their beliefs on them; an operation the agent is incapable of under Savage's framework.

This difference is actually the one that is usually emphasised in the literature as it is the one that makes it possible to express, in Jeffrey's framework, the division between evidential decision theory and causal decision theory. Thus, it has been suggested (see for example Byrne and Hajek 1997) that it is evidential decision theory which is to be blamed for the failure of the DBT and not the thesis itself. It has been argued, in other words, that it is not anti-Humeansim in general that is responsible for Lewis' result, but rather the specific anti-Humean position Lewis chose to attack, according to which evidential decision theory is the correct decision theory.

However, a closer examination of the formal structure of the result suggests that what drives it is neither Jeffrey's framework's commitment to the claim that

desirability always equal expected desirability nor the mere commitment to evidential decision theory.

To see that the first claim is not correct, notice that, as it is clear from the short proof of Lewis' result presented earlier, and as Lewis himself pointed out , in his 1996 paper, the triviality result holds whenever the condition $p(A^*) = p(A^* \mid A)$ holds. This condition follows from the DBT and the IR, but even if both are violated, and the condition still holds, the result holds. In other words, Jeffrey's desirability axiom plays no role in the result.

To see that the second claim is not correct, we have to examine more closely the differences between Savage's and Jeffrey's frameworks that result from the fact that in the former the probability function is not defined over acts while in the latter it is, when propositions play the role of acts.

Notice first that since acts are not part of the events' algebra in Savage's framework, it is meaningless to speak of the probability of any proposition conditional on them obtaining, as they are. Thus the condition $p(A^*) = p(A^* \mid A)$ cannot be expressed in this framework. However, one might want to add them artificially into the algebra, by adding more states to the states set. By making this move, the number of states in our original example will increase to eight, assuming one must choose either act a or act b, in the following way. Each world can now be seen as an ordered triple, where the first element can take either the value "a is good" or the value "a is bad", the second element can take either the value "b is good" or the value "b is bad" and the third element can

take either the value “a is chosen” or the value “b is chosen”. Now, we can treat the set of states in which the third element takes the value “a is chosen” as the proposition that says that a is chosen and we can update our beliefs on this proposition, which is now part of the algebra.

However, by doing that there will necessarily be states in which one of the acts does not lead to any consequences. This will be the case for each one of the acts regarding each one of the states in which this act is not taken. For instance, in a world where the third element takes the value “b is chosen”, act a has no consequence since we have assumed that only one of the acts can be chosen. Allowing acts to have no consequence in some states is prohibited in Savage’s theory. On an immediate level, this cannot be the case because acts are defined as functions from the set of states to the set of consequences, and a function must assign a value to every element in the domain, but more generally it is a violation of what Broome calls “the rectangular field assumption” which is an integral part of Savage’s theory[30].

So the condition $p(A^*) = p(A^* \mid A)$ cannot be expressed in Savage's framework, but, in a sense, this is so because it is written into its structure: the expected utility of an act is the expected utility of the world given that the act is chosen, so it might be more appropriate to say that *while in Jeffrey’s system the condition can be violated, in Savage’s system it cannot.*

[30] It will be discussed again later on in this chapter.

This limitation of Savage's framework makes it impossible to express in it the conditions Lewis has assumed for his result. Thus it "saves" the DBT in a trivial way, but it should be clear now that since this is so, the same kind of trivial reply to Lewis' argument is available also within Jeffrey's framework. All that one has to do is to replace the demand that one's degree of desire for a proposition ought to be equal to one's degree of belief that the proposition is good, by the demand that one's degree of desire for a proposition ought to be equal to one's degree of belief that the proposition is good, given that the proposition is true.

By adopting the latter demand, one adopts, within Jeffrey's framework, the Savageian restriction that the expected utility of an act is calculated always under the assumption that the act is taken. Indeed, Huw Price (1989) has argued exactly for this demand as a more plausible anti-Humean thesis. Lewis discussed Price's suggestion in his 1996 paper, admitting that it is consistent, but rejecting it in any case. However, his reasons for rejecting it are not clear. I was not able to find in the literature any discussion regarding the issue, but from what Lewis wrote it seems that he rejected Price's thesis on the ground that it is descriptively implausible[31].

As mentioned earlier, this seems odd since the DBT as well as the DET and Price's thesis are obviously false as descriptive hypotheses. Their initial appeal comes only when considering them as normative theses. However, some of the things Lewis wrote on the matter seem to suggest that he rejects Price's thesis

[31] See Lewis' (1996) discussion of the "Desire by necessity" thesis in section 2 and note that in section 5 he argues that Price's thesis is equivalent to the "Desire by necessity" thesis.

not merely because of its descriptive implausibility, but rather because it is trivial in some sense.

Although I hesitate to attribute the following argument to Lewis himself, there is one sense in which Price's thesis as well as the DET and all the replies to Lewis that are based on the rejection of the appropriateness of Jeffrey's framework for a formulation of an anti-Humean thesis, are trivial.

Firstly, notice that Price's thesis, just like the DET, is not necessarily an anti-Humean thesis. This is so since just in the same way that it is possible to reformulate the DET without referring to the agent's beliefs; if one accepts Jeffrey's desirability axiom, it is also possible to do this in the case of Price's thesis. It is easy to see that the following thesis is equivalent to Price's thesis: for any two propositions, A, and A*, where A* is the proposition that says that the level of goodness or desirability of A is x, $d(A \cap A^*) = x$[32].

Since this is so, it is clear that Price' thesis, as well as the DET, does not, in practice, restrict the agent's degrees of desires using the agent's degrees of beliefs. When the agent updates his beliefs, his desires are automatically updated accordingly so that Price's thesis will still hold. The same does not hold in the case of the DBT. In virtue of the IR, changes in the distribution of degrees of beliefs limit, for an agent that respects the DBT, the way he updates his degrees of desires.

[32] It is clear, thus, that by stipulating $A^* = A \cap G$, Price's thesis follows from the DET.

The triviality becomes even more apparent when considering the formulation of the DBT in Savage's framework. Here, a closer inspection reveals that the DBT is not only consistent with Savage's axioms, but actually follows from them. Since we have constructed the states in such a way as to make sure that in each world that belongs to the event "the goodness of act a is x", a indeed brings the degree of goodness x, the DBT cannot be violated if the agent maximises expected goodness.

Thus, the anti-Humean commitment of the DBT as formulated in Savage's framework is not due to any further assumption (that is, above Savage's original axioms) made in the model regarding the connection between the agent's probability and utility functions, but rather it is due to the assumption that the agent's utility function is identical to his goodness function. The latter is part of our interpretation of the model we have built. It is not expressed as an axiom in it. This in turn, leads to the conclusion that in order for us to formulate an anti-Humean thesis in a Savageian model, so that we can examine its consistency, we must enrich it in some way.

All of this does not show that Price's thesis, the DET, or the Savageian formulation of the DBT, are false. It does show, however, that they all avoid the question that lead us to consider Lewis' argument: the question of whether it is possible to use degrees of beliefs in normative propositions in order to constrain one's desires and through them one's preferences, or in our case, moral preferences.

The original DBT, however, does not do that. It does offer us a real, yet flawed, restriction on one's desires, using one's beliefs. As we have seen, it does this using the IR condition. Indeed, some scholars have put the blame for Lewis' result not on the DBT, but rather on the IR (see for example Bradley and List 2009). It is not anti-Humeanism in general that is responsible for Lewis' result, they have argued, but rather the specific anti-Humean position Lewis attacked according to which the IR condition must be kept.

This reply can be understood in two different ways. First, it can be understood as placing the blame on Bayesian conditionalisation. It can be argued, that is, that although in most cases updating one's beliefs using Bayesian conditionalisation is a rational thing to do, in some cases it is not. Adopting this position may also shed new light on another triviality result Lewis (1977) has proved, the triviality result for Adam's Thesis. Formally, the two results are very close and so it makes sense to argue that what they show is that Bayesian conditionalisation is not always rational.

Second, it can be argued that it is not the updating process that is responsible for the triviality, but rather the application of it on objects that are not propositions. In other words, it can be argued that sentences of the form "A is good to degree x" do not express propositions. This stance was taken by Allan Gibbard (1981), for example, regarding Adams' Thesis; conditionals, he has argued, are not propositions.

Adopting this line also regarding the DBT may seem to be a rejection of moral cognitivism, but it is actually not. It is, in fact, a rejection of moral cognitivism regarding judgements of the form “A is good to degree x”. Not all moral judgements, however, have this form. In the next section I will argue, moreover, that these judgements are not the ones we should concentrate on in any case.

So who is right? Where should we really put the blame for Lewis’ result? Which one of the replies to Lewis that I have presented should we accept? Well, I think all of the replies that we have considered succeed in blocking Lewis’ argument against anti-Humeanism in general. The anti-Humean who Lewis attacked is a very special anti-Humean; he is an anti-Humean committed to evidential decision theory, who respects the IR, who rejects Adams’ Thesis, and believes that goodness always equals expected goodness, and not all anti-Humeans are like that.

However, the fact that Lewis’ argument fails to refute anti-Humeanism in general, does not imply that Lewis’ general worry is unsound. Desires and beliefs do behave differently and thus there is a reason to suspect that trying to reduce one to the other might cause problems. In order for us to be convinced that this is not the case, we must introduce a specific anti-Humean thesis that is consistent with decision theory and that is plausible also in other respects.

Our discussion of the different replies to Lewis can help us establish this task in a way that will be immune to similar objections. The discussion has revealed, I hope, that the DBT fails not because it is an anti-Humean thesis. It is clear that

whatever is wrong with the DBT, it is not its mere commitment to a necessary connection between desires and beliefs. It is either the specific connection it requires (which can be accepted, as the reformulation of the thesis using Jeffrey's desirability axiom shows, also by Humeans) or its inconsistency with the IR.

The anti-Humean thesis I will present in the next chapter will take this into consideration. It will do so in several ways. It will not assume cognitivism regarding judgements of the form "A is good to degree x", it will be silent regarding the question of the appropriate way to update one's beliefs and it will not be committed to evidential decision theory. Nevertheless, it will not be trivial, that is, it will put some real restrictions on the agent's moral preferences using the agent's degree of beliefs in normative propositions.

The trick will be to work directly with beliefs about preferences. This move is also conceptually important for reasons I will explain in the next section.

Moral Uncertainty, Comparative Moral Judgements and Degrees of Moral Value

Quite independently of the literature that has responded to Lewis' argument, another line of philosophical inquiry has explored moral uncertainty. While the first set of literature was mainly concerned with the mere possibility of uncertainty regarding normative propositions having any normative significance to decision making, the latter deals specifically with moral uncertainty and has a

more pragmatic aim; it tries to suggest normative constraints on the way one ought to choose under conditions of moral uncertainty.

Of course, the two issues are not entirely distinct. The DBT, for example, can be seen also as a prescriptive thesis in that it demands that rational agents always choose an act which they believe is a good one. However, the same cannot be said regarding the literature discussed in this section. Philosophers that belong to this school of thought have ignored the question of the mere possibility of moral uncertainty having any normative significance to decision making. Rather, they have assumed that moral uncertainty does have such normative significance and have tried to answer the question of how morally motivated rational agents ought to choose under conditions of moral uncertainty[33].

Since the discussion in this literature is implicitly conducted in Savage's framework, and so is freed from the problems discussed in the previous section, examining its treatment of moral uncertainty will be useful for our purposes. It will allow us to focus our attention on an issue that was ignored by the literature that evolved around Lewis' result and that does not arise only in Jeffrey's framework.

[33] Interestingly, these philosophers have almost always expressed their surprise that the issue was not discussed at length earlier. Thus, Frank Jackson and Michael Smith (2006) write, referring to this issue, "Our concern in this paper is with an issue that seems to have slipped under the radar" (p.267) and Ted Lockhart dedicates a whole section in his book to the discussion of possible explanations for the absence of philosophical discussions of the issue. However, it seems clear to me that the literature that has evolved around Lewis' argument does discuss this issue. I tend to attribute their failure to notice that to their (implicit) commitment to Savage's framework (that will be demonstrated soon) in which the problem, as we have seen, does not arise.

The starting point of the philosophers involved in this body of literature is similar to my starting point in this thesis, in one respect, and different from it in another. It is similar to it in the sense that it chooses to take seriously the phenomenon of people *feeling* uncertain regarding moral propositions. It is different from it in the sense that it is not limited to the uncertainty that comes from realising that one holds inconsistent moral judgements. This difference, I will argue, is significant.

The general structure of the situation, in any case, is familiar; you must make a decision between, say, two possible acts available to you. You are acting as a moral agent, that is, you want to do the right thing, but you are not sure what the right thing to do is.

The approach taken by virtually all the scholars that have discussed this kind of case is to try to trace this uncertainty to uncertainty about a different issue. These "different issues" are of two kinds. Firstly, it might be that you are uncertain about what is the morally right thing to do because you are uncertain about the consequences of your acts. That is, you would know which one of the acts available you ought to chose, if you had no uncertainty regarding their consequences, but since you have such uncertainty you are not sure what the right thing to do is. My assumption in this thesis is that in these types of case one should just follow the recommendations of Bayesian decision theory for decision making under risk, as Harsanyi claimed one should. Although some scholars have doubted the plausibility of some of the Bayesian axioms, particularly in the context of moral decisions-making (see Diamond 1967 for example), I will ignore this issue at this point.

However, it has been argued by some philosophers, and I agree, that sometimes the uncertainty that we experience regarding moral decisions cannot be reduced to uncertainty about the consequences of our actions. For example, Jackson and Smith (2006) argued that in some cases the uncertainty is not about any factual matter, but is rather about a moral question. Specifically, it is about how to apply a moral theory to a real life situation.

Their example (which is used by Lockhart [2000] too) is the uncertainty that some people experience about the question as to whether abortions are permissible or not. One can hold a position according to which killing an innocent person is prohibited regardless of the consequences of such an act, for example one can believe that it is not permissible to kill an innocent person even in order to save the lives of many other people, while still being uncertain about the question of whether an early stage foetus is a person. To be sure, the question of whether an early stage foetus is a person is a moral question, not a non-moral one. Whatever biological findings one may have, one must still take the further step of deciding what the ethical relevance of these findings is.

Even if one is a consequentialist, or one believes consequentialism might be true, one might experience moral uncertainty without being able to trace it back to uncertainty about the consequences of one's possible acts. One might merely be unsure about what morality requires one to do. This can happen if the moral theory one happens to hold does not cover the particular decision one is facing, if one is unsure what the recommendation of the theory is in the context

of one's decision, or if one is unsure which one of several competing moral theories is the right one.

In his 2000 book, Ted Lockhart addresses exactly this kind of case. Lockhart employs a wide range of arguments to show that in such cases, a moral agent must choose an act that maximises expected moral rightness. In order to do so, however, the agent must not only know the degrees of moral rightness of every possible act, according to every theory to which he assigns a positive probability, but the agent must also be able to compare the degrees of moral rightness that different moral theories (i.e. the theories the agent thinks might be the right ones) assign to each one of the acts available to him. In order to do this, Lockhart adopts a principle which he calls the "Principle of Equity among Moral Theories" according to which, when comparing between degrees of rightness assigned to a specific act by different theories, one should give equal weight to every theory, that is to say that one should measure the degrees of moral rightness on a single scale for all moral theories.

Lockhart presents some arguments in support of this principle[34] and demonstrates how it can be used in particular cases, but it is important to see that neither Lockhart's characterisation of cases of moral uncertainty, nor Jackson and Smith's, succeeds in capturing the kind of moral uncertainty that motivates my discussion.

[34] And Andre Sepielli questions its plausibility and offers a modified version of it (Sepielli 2008).

My starting point in this thesis is the phenomenon of people holding inconsistent moral judgements. I argued in the previous chapter that in order to get rid of such inconsistencies we must use the method of reflective equilibrium. I also argued that some of these inconsistencies cannot be eliminated by incorporating into the process of achieving a reflective equilibrium some psychological knowledge regarding the way our judgements are produced. Specifically, this will be the case when the judgements in question are about choice situations in which more than one morally significant aspect is involved and thus, in order to make a comparative judgement, the agent must implicitly assign weights to the different aspects.

It was argued that moral theories do not typically give us such weights and thus, Lockhart's rule that is based on the assumption that the agent has access to the exact levels of rightness or wrongness of the acts available to her, according to each moral theory, does not seem to be applicable in such cases. As mentioned, much of the discussion in the moral uncertainty literature addresses the question of how to compare degrees of moral value among theories. However, in light of the discussion in the previous section, it seems that there is a much more pressing problem, which is how to get the exact levels of moral value *from a specific theory*.

My intention is not to argue that the moral uncertainty literature is useless. Sometimes people face moral choices that do not involve more than one morally relevant dimension and sometimes people face moral choices that do not require assigning exact degrees of moral value. In such cases, the

framework Lockhart has suggested may be helpful. However, I hope I have shown that the kinds of cases I am dealing with are not insignificant.

How wide is the range of cases I am dealing with? Well, it is as wide as violations of Savage's axioms by moral agents are frequent. Now, the success of the psychological literature in identifying systematic deviations from Savage's theory – as well as the wide exposure it gets in popular culture - tends to make people forget that these deviations are relatively rare. Most of the time people do obey Savage's axioms, both in moral contexts and in non-moral contexts. So the range of cases I am dealing with is small relative to the entire moral domain.

However, when one is in search of a complete moral theory one must consider the entire moral domain. Thus, I believe there is no escape from the need to give an account of moral decision making under conditions of moral uncertainty that is applicable also in the kind of cases that has motivated my discussion, i.e. cases in which the moral uncertainty people experience is the result of their lack of access to a well-defined value function. In such cases, Lockhart's framework, which is based on the assumption that the agent does have such an access (at least according to a given theory), does not seem to be applicable.

One might try and expand Lockhart's suggestion by interpreting the term "moral theory" in a more general way. Maybe by the term "moral theory" Lockhart does not mean an actual moral theory that can be found in the literature, or in people's ethical discourse, but rather a hypothetical moral theory that the agent

should formulate for herself. Specifically, Lockhart's suggestion, under this interpretation, would be something like the following.

When an agent has to make a judgement regarding which one of several available acts is the morally right act to choose in a specific situation, she first decides which are the morally relevant aspects of the choice situation, then she formulates to herself several hypotheses regarding the exact levels of moral value of each act, in light of each one of the morally relevant aspects, and about how to weigh each one of them. Having done this, she assigns probabilities to each one of these hypotheses and uses the "Principle of Equity among Moral Theories" in order to maximise the expected level of moral value.

In this picture, the "moral theories" that Lockhart refers to are not actual moral theories, but rather are hypotheses regarding the degrees of moral values of the different acts available.

The fact that such a suggestion seems to be detached from the way actual people make moral judgements should not count as a criticism of it. People actually do make moral judgements in a way that leads them to generate inconsistencies and, thus, if we want to find a way to get rid of these inconsistencies we might have to adopt a method for making moral judgements that seems unnatural to us. Lockhart's suggestion is normative, after all, not descriptive.

In fact, it is easy to see that this reconstruction of Lockhart's suggestion is basically the "Savageian" way to approach moral uncertainty. Each hypothesis regarding the degrees of moral value can be represented as a state in the states' set (or, in cases where the agent also suffers from some factual uncertainty, as a set of all the states in which the hypothesis holds), when the numbers it assigns to consequences represent the degrees of moral value each act will bring, given that this hypothesis is true. The "Principle of Equity among Moral Theories" makes it unproblematic to compare these numbers across states, but any other principle that gives instructions for how to compare degrees of moral value across theories can play the same role. One just has to treat the numbers assigned to each consequence as weighted moral values[35].

Indeed, when economists working in the Savageian tradition tried to model situations in which the agent suffers from uncertainty regarding his own tastes, this route is exactly the one they took (See for example Loomes, Orr and Sugden 2009). In this account different "taste states" were introduced and each one of them was characterised by a well-defined utility function.

Notice the psychological assumption underlying this treatment, which is especially striking when considering that the model presented by Loomes et al. was suggested as a descriptive model: although the agent is uncertain whether he prefers one act, a, to another, b, for any lottery between a and a third act, c, regarding which he is certain that he prefers a to it, he is certain that if it is the case that he prefers a to b then he prefers/does not prefer/is indifferent between

[35] Notice also that this interpretation of Lockhart's suggestion is basically the way Broome's DET will be formulated in Savage's framework. The different theories, the different hypotheses regarding the degree of moral value of each act, are exactly Broome's different propositions G_i.

b and this lottery. In other words, the agent is uncertain regarding his preferences among two acts, but still has no uncertainty regarding his preferences over all the possible acts, given that his preferences among these two acts are one way or another.

This assumption may be justified in some descriptive contexts. As incredible as it sounds as a psychological assumption, given that using it yields good predictions and interesting explanations, it should not be rejected immediately. However, when one considers it in the context of a prescriptive theory, it is justified only as so far as (and in the contexts when) it holds for a specific agent.

I think that by examining the strategy described above more carefully in our context, it becomes clear that all that it does is to push the problem one step further. Consider an agent who tries to follow the strategy described above. She decides which aspects of the situation faced are morally relevant and formulates a number of hypotheses regarding the degrees of moral value of the acts available to her. Then she has to assign a probability value to each one of these hypotheses. On what basis can this be done? Surely Lockhart would not want to argue that she should do this arbitrarily. Arguing this is like arguing that she should choose an act arbitrarily as by assigning different probabilities to the hypotheses she can make any one of the acts the one that maximises expected moral value.

What Lockhart probably had in mind is that she should do this according to her actual degrees of beliefs in these hypotheses, which should probably be based

on what she takes to be moral evidence. However, it is not clear what can constitute evidence for a specific hypothesis regarding the exact degree of moral value of an act, other than the kind of choices that assigning such a degree to the act (together with assigning other degrees to other acts) leads to. Actual moral theories do not give us exact degrees of moral values in the kind of situations I am referring to and people do not generally have intuitions regarding such levels. What people do when they have to assign such levels, is to do so implicitly by judging, among many possible acts, which is better.

It is important to stress that this argument is independent of the question about whether degrees of moral values are "real", obsolete, or irreducible to other moral facts, if there are any. To make things clearer, it is best to look at the analogy with the debate regarding utility, preferences and the relation between them. Some scholars believe that utility (or desirability) is a real mental quantity. Others believe that it is not but rather that it is a conceptual construction that makes it easier for us to discuss and predict behaviour[36]. Either way, it is agreed by everyone that if an agent is rational in Savage's sense, the degrees of her utility must be consistent with her preferences in the sense that one act is preferred to another iff the degree of expected utility of one act is higher than that of the other.

This is what Savage's representation theorem says. It states that an agent with preferences (over a rich enough set of acts), that obeys the axioms, can be

[36] For an overview see Colander (2007).

represented as an expected utility maximiser for a unique probability function and a utility function that is unique up to affine transformation[37].

Using this theorem, one can measure one's degrees of utility from the possible outcomes of every act available. The main idea was suggested already by Von Neumann and Morgenstern: utility levels can be measured using the agent's preferences over lotteries among possible outcomes, which are equivalent to acts in Savage's framework. One can understand "measuring" here in a purely realist, purely operationalist, or a conventional way; that is one can take the measurement to be a procedure that aims at finding the real utility values, that defines these values, or that has some constitutive role in the determination of these values. In any case, one can gain access to these values through one's preferences. This is true, it is important to stress again, regardless of one's position regarding the question of ontological, conceptual or epistemological priority, of utility over preferences or preferences over utility.

Moving from personal preferences and utility to moral preferences and moral value does not change anything in this respect: regardless of the question of whether degrees of moral value are real or not, it is clear that one can gain access to them through, and that they ought to be consistent with, one's moral preferences.

[37] That is, if for a probability distribution p and a utility function u, maximisation of expected utility gives the agent's preferences, the same holds, and holds only, for p and any utility function v, such that v=au+b, when a and b are parameters.

John Broome has argued that "…goodness is actually fully reducible to betterness; there is nothing more to goodness than betterness". (Broome 1999, p.164). Although for Broome the term "goodness" is not synonymous with "rightness" (this will be discussed in more detail in chapter 5), the main idea is the same; it is not the case that degrees of moral value determine moral preferences, but rather, the other way around.

Now, there is no need for us to adopt Broome's strong claim that goodness (or rightness in our case) is reducible to betterness (or to the relation "being morally superior to", in our case); rather, it is enough to claim that in the kind of cases I have characterised in the previous chapter, the relation of "being morally superior to" has *epistemic priority* over the notion of an act being a right act to a specific degree.

This claim is not only supported by introspection, but also by the psychological evidence discussed in the previous chapter: one factor that causes the inconsistencies among people's comparative moral judgements is the fact that they do not have direct access to the levels of rightness and thus they substitute the target attribute of being right to a specific degree with the "heuristic" attribute of triggering an emotional reaction to a specific degree.

Again, this is not to say that people have no access, under any circumstances, to degrees of moral value; only to say that there are cases in which they do not and that these cases are exactly the cases in which we will expect to observe inconsistencies. It is, of course, true that most people will not hesitate to claim

that the degree of positive moral value of saving another person's life is higher than that of slightly improving a person's well-being and that, ceteris paribus, the degree of positive moral value of donating $2000 to charity is higher than that of donating $1000.

However, it is also true that most people will have a hard time assigning exact degrees of moral values to each of these acts. Thus, when having to make a choice between saving the life of one person and slightly improving the well-being of many people, or between donating $1000 that will definitely go to a good cause, and donating $2000 that may go to the same good cause and may not, there will be a point (i.e. when there are enough people whose well-being can be improved or when the probability of the money going to the good purpose is low enough) at which most people will be unsure regarding which act they morally ought to choose.

The moral uncertainty literature, it seems to me, puts too much emphasis on the problem of comparing the degrees of moral value that different theories assign to the acts available and not enough emphasis on the problem of how we get these degrees for a single theory. The first problem is the analogue of the problem of interpersonal comparisons of utility: it is derived from the fact that even when one has consistent comparative moral judgements, given a theory, these determine a moral value function that is unique only up to affine transformation and, thus, when one has to compare the degrees of moral value different consistent theories assign to a specific act, one must use some source of information beyond one's comparative moral judgements.

However, this problem only arises when one's comparative moral judgements, given different theories, are consistent. When they are not, the problem is not how to compare the degrees across theories, but rather *that there is no function that is consistent with them*. Thus, it seems to me that prior to dealing with the problem of inter-theoretical comparisons of moral value, one has to deal with the problem of the cardinality of moral value within one theory.

Now we are also in a position to see why none of the replies to Lewis' argument that I have discussed in the previous section (and indeed none of the suggestions for an anti-Humean thesis that is not under threat from Lewis' result) is satisfactory for my purpose. They are all unsatisfactory because they all keep the assumption that it is always possible to treat moral uncertainty as uncertainty regarding different hypotheses about degrees of moral value. I think that sometimes it is possible to do this, but as I have argued above, this is not the case when the moral uncertainty arises as a result of one becoming aware of an inconsistency among one's moral judgements.

The conclusion so far is the following. In order for an account of moral uncertainty that makes use of degrees of moral value (e.g. any account that is based on the idea of maximisation of expected moral value) to be applicable, it must give an answer to the question as to how people can gain access to these values. Such an answer cannot be based solely on people's comparative moral judgements, as typically these will be inconsistent when the set of available acts is rich enough, *and in order for Savage's representation theorem, or indeed any*

other representation theorem in the literature, to hold, they must think of a rich enough set.

Such an answer cannot also be based solely on introspection because, as discussed in the previous chapter, introspection does not give us such degrees. Now, Lewis' DBT seems, at a first glance, to be such a thesis, as it describes a direct connection between degrees of moral value and degrees of belief in normative propositions. However, this is due to the fact that Lewis' discussion was mainly made in the context of the simple case where there are only two possible degrees of goodness that can be normalised. When moving to the general case, degrees of moral value come back into the equation[38].

Therefore, my claim is that in order for us to formulate an anti-Humean thesis *that can be applicable to our context*, we must use degrees of belief in comparative moral judgements and these alone. As we have no direct access to degrees of moral value, no reference to such degrees should be made.

The strategy I suggest is to try to formulate consistency conditions that connect one's moral preferences and one's degrees of beliefs in comparative moral judgements. If we are able to formulate such conditions that will ensure that an agent that follows them ends up with moral preferences that obey the rationality axioms, and if these conditions can be satisfied not only in trivial cases, then we will have a consistent, non-trivial anti-Humean thesis. We will also have a thesis

[38] See footnote 24.

that can be used by an agent that holds inconsistent moral judgements and yet wishes to modify them in order to gain consistency.

As argued in the previous chapter, adopting a set of such conditions also amounts to saving the reflective equilibrium method from being just a characterisation of any kind of reasoning. Considering the issue from this point of view sheds new light on the shortcomings of the moral uncertainty literature. By restricting itself to uncertainty that can be reduced to uncertainty regarding theories, it is unable to deal with uncertainty that arises in the course of one's search for a moral theory.

The main idea behind the reflective equilibrium approach is that our moral judgements regarding specific moral questions should be taken as evidence for (or against) accepting general moral claims. However, the moral uncertainty literature demands that whenever uncertainty regarding a specific moral question arises, one should reduce it to uncertainty regarding complete moral theories. Thus, inference from the former to the latter is not allowed.

In conclusion, the commitment of the moral uncertainty literature to the standard Savageian treatment of uncertainty limits its applicability to cases in which the agent is able to reduce the uncertainty he experiences to uncertainty regarding which one of several competing hypotheses regarding the exact (unique up to affine transformation) degrees of moral value of every act available to him and every possible lottery among these acts, is correct. This prevents it from being applicable to cases in which the uncertainty arises as a result of becoming

aware of having inconsistent moral judgements. In particular, it prevents it from being applicable in a context of moral inquiry.

The same is true regarding the literature on Lewis' result. All of the different approaches in this literature share the assumption that the agent is able to formulate his uncertainty as uncertainty regarding which hypothesis about degrees of moral value is correct.

In the next chapter, I will relax this assumption and try to follow the strategy I have suggested above. In practice, I will be looking for a representation theorem that has the following general form: *an agent who obeys the following conditions regarding the way his degrees of beliefs in comparative moral judgements interact with his moral preferences ... can be represented as an expected moral value maximiser.* This formulation follows from my commitment to the rationality demand. A preference ordering that obeys the decision theoretic axioms can be derived from a maximisation of expected utility for some probability function and some utility function (unique up to affine transformation). Therefore, a set of conditions regarding the relations between one's degrees of belief in comparative moral judgements and one's moral preferences, that ensure that one's moral preferences obey the decision theoretic axioms, also ensures that one's moral preferences can be derived from the maximisation of expected moral value for some probability function and some moral value function.

However, here a methodological dilemma arises regarding the question as to which framework is more suited for the inquiry, Jeffrey's or Savage's? In Savage's framework, the probability function that represents one's uncertainty is defined only over the set of states. The agent's preferences are defined over the acts available to the agent. Since we are concerned in this inquiry only with idealised morally motivated agents, we can assume that under conditions of no moral uncertainty, the agents' moral preference orderings are identical to the orderings derived from the set of the comparative moral judgements they hold. Thus, instead of talking about uncertainty regarding comparative moral judgements, we can talk about uncertainty regarding one's moral preferences.

The only way to express uncertainty regarding such preferences in the Savageian framework is by including the agent's preferences among acts in the description of the states. However, by following such a strategy one must, in some cases, allow the agent to give positive probability to states in which although one act, a, is preferred to another, b, the consequence of b in the state, is preferred to the consequence of a in the state[39].

This happens because, in these cases, a is preferred to b not because its consequence in the given state is preferred to the consequence of b, but because its consequences in other states are preferred to the consequences of b.

[39]As will be explained in the next chapter, to say that a consequence, A, is preferred to another consequence, B, means that the act that brings A in every state of the world (the constant act of A) is preferred to the act that brings B in every state of the world. Savage requires that all the constant acts are available to the agent.

It is well-known that such cases, in which the value of a consequence depends on the consequences that the act that brings it about, brings in other states, lead to unintuitive results in Savage's framework, even when no uncertainty regarding the preference relations is involved[40]. In order to get rid of these unintuitive results, the consequences must be redescribed, using a finer individuation. However, such finer individuation usually leads to a violation of one of Savage's assumptions, "the rectangular field assumption"[41]. This assumption states that the set of acts available to the agent must include every act that can be constructed by assigning any one of the consequences to any one of the states. Thus, one might argue that Savage's framework is not the right framework to use in order to investigate uncertainty regarding comparative moral judgements.

To demonstrate that this phenomenon also occurs in our context, consider the following simple case. An agent is certain that act a is morally superior to act b, and that act b is morally superior to act c (and thus that a is also superior to c). However, since he has no access to exact degrees of moral value, there is some lottery, l, between a and c, regarding which he is uncertain whether it is superior to b or not. How should we model such a situation, under Savage's framework?

Consider first a case in which there is no moral uncertainty. In such a case all the information about the agent can be summarised in the following table.

[40] See Binmore (2009, chapter 1) and Broome (1991, chapter 5) for discussions.
[41] This is what John Broome (1991) calls it.

	ω_1	ω_2
A	A	A
B	B	B
C	C	C
L	A	C

Table 3

Here act a, b and c have sure outcomes, and act l is a lottery between a and c. In order to represent the moral uncertainty the agent suffers from, we have to split each one of the two states into two states, one in which b is preferred to l, and one in which the opposite holds (to simplify the matter we can assume that the agent is certain that the acts are not morally equivalent). This is demonstrated in the matrix below.

	ω_1 (b>l)	ω_2 (l>b)	ω_3 (b>l)	ω_4 (l>b)
A	A	A	A	A
B	B	B	B	B
C	C	C	C	C
L	A	A	C	C

Table 4

Now, ω_1 is the state in which act I brings A and b is (morally) preferred to I, ω_2 is the state in which act I brings A and I is preferred to b, ω_3 is the state in which act I brings C and b is preferred to I and ω_4 is the state in which act I brings C and I is preferred to b.

Consider now state ω_4. In this state, act I is morally preferred to act b, however, the consequence b brings in this state is preferred to the consequence that I brings in it. This happens, intuitively, because even though, in ω_4, I brings about a consequence which is morally inferior to the consequence b brings about, because of what I brings about in the other states, compared to what b brings in them, I is morally superior to b.

However, compare this to state ω_3. I and b bring about in ω_3 exactly what they bring in ω_4 and, moreover, what they bring in states different from ω_3 is exactly what they bring in states different from ω_4. Still, in ω_4 I is preferred to b and in ω_3 b is preferred to I. How can this happen? Which feature of the decision problem is responsible for the difference?

Well, the only difference between the two states is that in ω_3, b is preferred to I and in ω_4, the opposite holds. There is no other difference. The consequences the two acts bring in each of these two states are identical and, given a state, the consequences the two acts bring in all other states are also identical. This leads to the conclusion that we have failed to individuate the consequences properly. The consequence of I in state ω_4 must be different from the

consequence of l in ω_3. Otherwise, there will be no reason for the difference between the two worlds.

This conclusion is also intuitively appealing since, from a moral point of view, choosing or failing to choose the act that should have been chosen does make a difference. The consequence l brings about in ω_4 is not identical to the consequence it brings in ω_3, since the mere fact that in the first l is morally superior to b and in the latter the opposite holds, matters to the agent, and thus, changes the way he evaluates the two consequences (see Broome 1991 chapter 5 for a similar point).

Thus, one might argue that the correct way to describe the choice problem is by differentiating between the consequences of acts l and b in the different states. This can be done in the following way:

	ω_1	ω_2	ω_3	ω_4
B	B and b>l	B and l>b	B and b>l	B and l>b
L	A and b>l	A and l>b	C and b>l	C and l>b

Table 5

If this is true, however, then there can be no act that brings, for example, the consequence “C and l>b” in state ω_3 since this state is a state in which b is preferred to l. Thus, the rectangular field assumption is violated.

Thus, Savage's framework, as it is, does not seem to fit our purpose by virtue of the assumption of a strong separation of the set of states from the set of acts it relies on. In Jeffrey's system, on the other hand, this problem does not arise since there is one object, propositions, on which both the preference relation and the probability function are defined. As Jeffrey himself noted, "...nothing in the system of *The Logic of Decisions* requires us to suppose that the terms of the preference relation are naturalistic propositions. They may equally well be propositions about the agent's preferences" (Jeffrey 1983, p.226).

Thus, it seems natural to use Jeffrey's framework in order to carry on the project. However, as we saw in the previous section, the flexibility that Jeffrey's system allows brings with it several complications. In the context of uncertainty regarding the preference relation itself, the following problem arises. Since the preference relation is defined over all propositions, just like the desirability and probability functions, propositions that describe preference relations between other propositions also stand in preference relation to all other propositions. In the same way, such propositions have a probability value and a desirability value.

In particular, propositions that describe a preference relation between other propositions stand in a preference relation to the propositions to which they refer. Moreover, Jeffrey's desirability axiom makes it necessary that the probability of such propositions constrain the preference relation among other propositions. For example, let us use the notation "AR¬A" to denote the proposition "A is morally preferred to ¬A". Then, Jeffrey's desirability axiom tells

us that (assuming for simplicity that A and ¬A cannot be equally morally valuable):

$$d(A) = p(AR\neg A \mid A)\, d(AR\neg A \cap A) + p(\neg ARA \mid A)\, d(\neg ARA \cap A)$$

Since the preference relation between A and ¬A is determined by their degrees of desirability, it is clear that working within Jeffrey's framework must involve a commitment to some requirements, regarding the way beliefs concerning the preference relation and the preference relation itself constrain each other, that are absent from the Savageian framework.

Thus, carrying on the investigation using Jeffrey's framework might, just as in the case of the DBT, lead to conclusions that can be avoided within Savage's system due to its less flexible structure. Moreover, the idea of assigning degrees of desirability to propositions that describe preference relations between other propositions is not conceptually unproblematic. By saying that, I do not argue that this practice should be rejected. I simply do not want to explore this issue here, but by committing myself to Jeffrey's framework I would have to address it.

In other words, my aim here is to examine whether it is possible to formulate a non-trivial, anti-Humean thesis that makes use of degrees of beliefs about preferences. My aim is not to examine whether higher-order moral judgements are meaningful. However, if I were to choose to work within Jeffrey's framework, I would not be able to separate these two different issues.

The way I choose to deal with this methodological dilemma is by presenting a simple extension to Savage's model that will allow me to discuss uncertainty regarding the preference relation without incorporating it into the description of the states. This will be done by introducing a second probability function that is defined over the set of all possible preference relations among all the possible acts. By doing this, I can avoid the complexities that Jeffrey's framework brings with it, while not being exposed to the problem that the standard Savageian treatment of uncertainty creates in our context. However, as we will see in the next chapter, even within this framework, an important problem arises.

Conclusion

Lewis' argument against anti-Humeanism was based, I have argued, on a valid concern. In any decision theory, desires and beliefs behave differently. The main message of the anti-Humean position is, however, the claim that some desires are constituted by normative beliefs. Thus, when trying to incorporate these beliefs into the decision-theoretical framework we should expect some problems to arise.

Lewis has tried to show that this general consideration does in fact translate into a specific problem for some versions of the anti-Humean position. It is not entirely clear whether he was successful in this attempt, as it is not entirely clear whether the result is driven by the anti-Humean component of the DBT. In any case, the different replies to Lewis, which were discussed in the first section,

show that even if the attempt was successful, it does not necessarily indicate that it can be generalised to version of the anti-Humean position. These two lessons leave us in the following position; the door is still open for introducing an anti-Humean thesis that can be used by an agent who wishes to change some of his inconsistent moral judgements in order to gain consistency. However, we should be aware of Lewis' general worry that threatens any anti-Humean thesis.

In the second section, I argued that the anti-Humean thesis we are seeking should have the following form: it should give a set of conditions that connect one's degrees of beliefs in comparative moral judgements to one's moral preferences. Specifically, I have argued, no reference to degrees of moral value should be made in such a thesis.

We have now reached the point at which such a thesis should be formulated. This will be done in the next chapter using the framework of a simple extension of Savage's model for decision making, which will enable us to both avoid some of the complications Jeffrey's framework brings with it and the limitations that Savage's original model imposes.

The thesis that will be formulated will prove to be consistent with all of Savage's axioms. However, it will also be shown that this happens only in trivial cases. Thus, one might take this result to support Lewis' objection to anti-Humeanism, or even to support a rejection of either moral cognitivism or the possibility of moral uncertainty. In the forth chapter I will, however, argue that such a

conclusion is mistaken and that, in fact, the Humean position faces, in the kind of situations we are dealing with, even more severe problems than the anti-Humean one.

It will also be argued that denying moral cognitivism or the possibility of moral uncertainty will not help us to avoid the problem. Rather, I will argue, the real problem arises from a deeper source which is the tension between the motivational demand and the rationality demand.

Chapter 3: Can an irrational agent reason himself to rationality?

Introduction

As was discussed in chapter 1, most of us are prone, under certain conditions, to express intransitive preferences - either by actual choice, or in reflecting on cases. This is true both in the context of moral decisions and in the context of decisions that involve no ethical aspect. At the same time, most of us do accept transitivity of preferences as a condition of rationality. Thus, when we realise that our expressed preferences are intransitive we usually want to change them so that transitivity will be restored. How can and how should this be done?

In many contexts, one can argue that it does not really matter how this is done. As long as an agent ends up with a preference ordering that obeys the axioms of rationality, there is nothing more that should be said about him from a normative point of view. However, there are some contexts in which it does seem that not all ways of changing one's preferences should be equally acceptable from a normative point of view. These contexts are contexts in which the agent himself believes that there exists some "objective" betterness relation among the options over which he forms his preferences. By "betterness relation" I mean a complete, reflexive, and transitive relation that ranks different options according to how good they are for a certain purpose. By "objective" I mean that this ordering is taken by the agent to be something with respect to which he can form beliefs that are either true or false.

It should be clear from the discussion so far that my assumption here is that in the moral context this condition holds. Our commitment to the rationality demand allows us to treat the "morally superior to" relation as a "betterness" relation (i.e. it is "betterness" from the point of view of morality) and our commitment to moral cognitivism allows us to take it as an "objective" relation[42].

In such contexts, it seems reasonable to demand that when the agent changes his intransitive preferences so that they will become transitive he will be guided by certain conditions that will ensure that his preferences cohere with the betterness relation. In this chapter, I will introduce, within the framework of a simple extension of Savage's model, two such natural conditions. In the next chapter, I will show that by accepting them one commits oneself to a very disturbing result[43].

The reminder of the chapter will be organised as follows**.** In the first section, I will introduce the model and discuss some features of it. In the second and third

[42] In the previous chapter, I argued that moral cognitivism should not be ruled out in the context of an ideal rational agent because of Lewis' result. In the next chapter, I will present some arguments for why moral cognitivism is actually a more attractive approach to take, in this context, than non-cognitivism.

[43] There are other contexts that fall into this category and, thus, the result applies to them too. One such context is a context of an agent who is unsure what his own (non-moral) preferences are. Some might reject that this is even possible. Others, however, insist that it is (for example Richard Jeffrey 1974).

Another such context is the context of a person that acts as an agent of another person. That is, a person that makes choices on behalf of another person and tries to do his best to make these choices according to his beliefs about the other person's preferences. Such an agent may be uncertain about what the preferences are of the person for whom he acts as an agent. This context is, in fact, especially important to us because it is, according to one important consequentialist moral theory, namely revealed preferences utilitarianism, the context in which one's acts constitute the morally right acts to choose (see Harsanyi 1984 for a good discussion) .

While I will limit the discussion here to the moral context, I believe that many of the arguments that I will use can be easily modified to other contexts of the kind I have just characterised and thus, I believe that the result is worrying not only from the point of view of ethics. Here, however, I will only discuss its implications for ethics.

sections, I will introduce the two conditions I have mentioned and discuss their status. I will argue that while the second condition (that will be discussed in section 3) is a genuine principle of rationality, and as such, for somebody that accepts the rationality demand, a moral requirement, the first condition is not.

I will argue that first condition is an explication of the motivational demand. Thus, the discussion regarding this condition will also allow me to present the final, and most exact, formulation of the main problem I address in this thesis. It will also help us to better understand the source of the tension between the motivational demand and the rationality demand, the role the reflective equilibrium method is supposed to play in accommodating this tension, and the relationship of degree of beliefs in comparative moral judgements to all of this. As is often the case in philosophy, having formulated the question in an exact way, distinguished it from related questions and stated the assumptions in an explicit way, finding an answer is quite a straightforward matter. In our case, this will take the form of a simple proof that will be presented in the beginning of the next chapter.

The Model

As mentioned at the end of the previous chapter, the model is a simple extension of Savage's model that allows the agent to have beliefs about the betterness relation between acts without incorporating these into the descriptions of the states. We will also assume, for the sake of simplicity, that unlike in Savage's model, the agent's subjective probability function(s) is given.

Let $\Omega = \{\omega_1 \ldots \omega_n\}$ be a finite set of possible states. Let p be a probability distribution over Ω. Let D = {A, B, C…} be a set of outcomes and let $A = \{a_{1 \ldots} a_k\}$ be a set of acts, where an act is a function from Ω to D. Let ≥ be a regular preference ordering over A (i.e. a complete, reflexive, and transitive relation). The assumption that ≥ is an ordering expresses our commitment to the rationality demand. In the reflective equilibrium we seek, it was argued, one's moral preferences obey the rationality conditions, and thus constitute an ordering. *Thus, it should be clear that I use the model in order to characterise the end state of the agent's deliberation process, not its starting point (in which the agent's moral preferences do not constitute an ordering). The same holds for the other two assumptions I am going to suggest.*

In addition, let >* denote the betterness relation between pairs of acts, i.e. >* is a binary relation over elements of A. For simplicity, we will assume that for any two elements, a_i and a_j, $a_i >^* a_j$ or $a_j >^* a_i$. By assuming this I am ignoring the possibility that the agent gives a positive probability to the possibility that two acts are equally good or desirable, i.e. that neither is better than the other. This assumption will make the discussion simpler and nothing depends on it (the reason for this will become clear as the discussion progresses).

Since we want to allow the agent to have beliefs regarding the betterness relation, we will usually need to refer to the betterness relation as a variable. In these cases, we will just use the notation "**>**". Finally, let q be a probability

distribution over all possible >*s. To be clear, the expression $q(a_i > a_j)$ denotes the sum of the probabilities q gives to all >* such that a_i >* a_j.

It is important to stress that by taking q to be a probability distribution over the set of all possible betterness relations, I do not commit myself, and do not intend to suggest, that either ordinary people or ideal moral and rational agents deduce their beliefs regarding the betterness relations that hold between different pairs of acts from their beliefs over the set of all possible rankings of all the possible acts available to them.

The agents might form their beliefs in such a way (although, as discussed in the previous chapter, I find it psychologically implausible and normatively unappealing), but nothing in the model requires them to do so. This is because I do not assume anything, at this stage, about conditional probabilities, that is, the probability of one act being better than another conditional on other betterness relations holding between other acts. Thus, I do not use any information that one gains from access to a specific probability distribution over the set of all possible rankings of the acts and that one does not have if one only has access to the probability of one act being better than another, for all pairs of acts[44].

[44] Now we can see that the distinction made in the previous chapter between three types of moral uncertainty on a conceptual level, that is the distinction between moral uncertainty that can be reduced to uncertainty about non-normative propositions, moral uncertainty that can be reduced to uncertainty about which moral theory is the correct one and "primitive" moral uncertainty, can be represented formally in a straightforward way. The first kind of moral uncertainty happens when there is no uncertainty regarding the agent's own preferences, the second happens when there is such uncertainty but all the probabilities, including the conditional probabilities, are known to the agent, and the third happens when only non-conditional probabilities are known (or in other words, when the probabilities of conjunctions are not known).

In fact, I am not even committed to accepting that the agent's beliefs are such that only transitive relations get a positive probability. It is consistent with the model that the agent will give a positive probability to an intransitive relation holding between some acts.

As Savage does, we can define each element of D as the constant act (i.e. an act that gives the same outcome in every state) whose value is this element and demand that A include all the possible, constant and not constant, acts. With this, we can treat the agent's beliefs regarding the betterness relation between constant acts as his beliefs about the betterness relation between outcomes and the agent's preferences over constant acts as his preferences over outcomes. For convenience, we will use the notation q(A**>**B) to refer to $q(a_i \mathbf{>} a_j)$ when a_i is the constant act that gives A and a_j is the constant act that gives B.

In the interpretation, p represents the agent's degrees of belief over factual matters in the world, while q represents the agent's degrees of belief over the betterness relation between different acts. Now, in order to present a thesis that connects one's beliefs regarding the betterness relation to one's preferences, we must specify two conditions: a condition that describes the way one's beliefs about betterness relate to one's preferences, and a condition that describes the way one's beliefs about the betterness relations that hold between constant acts relate to one's beliefs about the betterness relations that hold between acts that are not constant.

It is important to stress that even if we are able to specify conditions that ensure that one's preferences obey Savage's axioms, our work will not be done. The second stage in our project will be to formulate rules of reasoning that will make it possible for an agent who finds himself violating these conditions, and thus having inconsistent judgements, to change his beliefs so that the conditions will hold. However, as we will soon see, we will reach a dead end even before getting to this stage.

I will now present two such conditions and explain the motivation behind them. The first condition describes the way beliefs about betterness relate to preferences. I will discuss the motivation behind this condition in length as the discussion will allow me to present the main problem I explore in this thesis in a formal way.

The Likelihood of Betterness Constraint (LBC)

The idea behind the condition is quite simple: for any two acts, prefer the act to which you attach a higher probability of being the better one. Formally:

Likelihood of Betterness Constraint (LBC):

1. $q(a_i \mathbf{>} a_j) > q(a_j \mathbf{>} a_i)$ iff $a_i > a_j$ and

2. $q(a_i \mathbf{>} a_j) = q(a_j \mathbf{>} a_i)$ iff $a_i = a_j$.

In order to avoid possible confusion, it is important to stress that the LBC does not contradict Expected Utility Theory. In fact, I will prove in the next chapter

that for every p, there is always some q such that the LBC is consistent with Expected Utility Theory, even when there are infinitely many outcomes and the set of acts is convex.

Since I assume all of Savage's axioms hold in the model, it is clear that, without the LBC, the agent in the model can be represented as maximising expected utility for some utility function and the distribution 'p'. The LBC condition does not rule this out; but rather it makes use of the two elements that I have added to Savage's model: the betterness relation and the probability distribution q, which is defined over the set of all possible betterness relations. The LBC demands that the agent will always prefer one act to another if he believes it is more likely than not that this act is better than the other, but it does not say anything about the relation between the agent's preferences and the probability distribution p (which is defined over the set of states) and the agent's utility function.

The fact that the LBC is consistent with Expected Utility Theory does not imply, however, that it is justified. I will argue, later on, that in fact, in some contexts, the LBC is unjustified. Firstly, though, I want to explain what can motivate it. The main idea is that the LBC is neither a principle of rationality nor a moral rule, but rather an explication of the motivational demand. As such, for a reasoner who accepts the motivational demand, it becomes a constraint that he must respect. *In contexts where it is unjustified, the motivational demand itself is unjustified.*

The discussion will be developed in four stages. In the first stage, I will indicate what I take to be the source of the initial plausibility of the principle. In the second stage, I will defend it against an immediate objection. This will include some repetitions of points discussed at greater length in the previous chapter. I think doing so will be worthwhile, even at the risk of boring the reader, as it will help us to avoid some possible misunderstandings.

In the third stage, I will point to a new threat to the LBC that arises even for those convinced by my arguments up to this point and inclined to accept the LBC. This problem, I will demonstrate, is just an instance of the Lottery Paradox. This observation will help me to present the final formulation of the main problem I am investigating in this thesis, i.e. the problem of reconciling the motivational demand and the rationality demand, and to explain how exactly it relates to the reflective equilibrium method.

In the fourth stage, I will explain why the LBC ought not to be taken as a principle of rationality or as a moral rule, but will still argue that for an agent who accepts the motivational demand it is inconsistent not to obey the LBC.

The discussion that follows is a delicate matter and should be read accordingly. I am going to argue in favour of accepting the LBC. However, as my conclusion in the next chapter will be that in some cases it is in fact unjustified to follow it, it is crucial that we be as clear as possible regarding the question *what should we accept it as*.

First stage – Initial Motivation

The LBC can be interpreted in three different ways. You can take the agent's beliefs about betterness to constrain his moral preferences according to this rule; you can take the agent's moral preferences to constrain his beliefs about betterness according to it, or finally, you can take the LBC as a constraint on the agent's system of beliefs and treat the word "preferences" as a synonym for "beliefs about betterness", as Broome (2006) suggests we should do in some contexts. On this approaches you will have to interpret the expression "the agent prefers a_i to a_j" as $q(a_i \succ a_j) > q(a_j \succ a_i)$ and the expression "the agent is indifferent between a_i and a_j" as $q(a_i \succ a_j) = q(a_j \succ a_i)$. Any of these three interpretations will do.

Notice that since all three interpretations are possible, one is not committed, by virtue of accepting the LBC, to any specific theory of motivation. One can be an internalist regarding moral motivation, that is one can believe that moral judgements are motivating by themselves, or one can be an externalist regarding moral motivation, that is one can believe that moral judgements are motivating only by virtue of some other attitudes that (always or sometimes) accompany them.

I am not going to discuss here the various positions one can find in the literature regarding the matter[45]. All that I need to assume about motivation is that moral judgements *can* motivate, either on their own, or with the aid of some

[45] For good discussions see Mele (1996), Brink (1997), Rosati (2006).

accompanying attitudes. This assumption is needed in order for me to talk about an ideal moral agent (that is, an agent who is only motivated by moral considerations) as someone who is motivated in some sense, and is already explicit in my commitment to the motivational demand.

Assuming that moral judgements can be motivational, the question is how we can characterise they way they motivate agents to act. Now, for the case of real agents (i.e. agents that are motivated not only by moral considerations), any such characterisation will surely be committed to the following claim. If a moral agent has two incompatible judgements, and thus, must reject one of them, then, after taking into account any possible information he has regarding the causal mechanisms that are responsible for these judgements, or regarding any other matter that he takes to be relevant, he should keep the judgement he feels more strongly about[46].

From a descriptive point of view, this claim can be taken to be definitional: to say that the agent, *all things considered*, chose one judgement over another, just means that he feels more strongly about this judgement. However, the descriptive point of view is not the one that should concern us here. Descriptively, the agent's strength of feeling about his judgements will surely not obey the laws of probability. What we are interested in is the normative point of view. This perspective is the one that initially allowed us to assume that the agent's degrees of beliefs, factual or normative, obey the laws of probability.

[46] Michael Smith (2002) calls this the degrees of "certitude" the judgement has for the agent. Sunstein (2005), and some of the scholars that have replied to him, talk about the "firmness" of moral intuitions. However, in light of our commitment to moral cognitivism we can just use the term "degrees of belief".

Since we have assumed (and defended against one possible objection, but have not yet argued in favour of the fact) that the right way to characterise these "degrees of strength of feeling" is as degrees of belief, and since we are committed to the rationality demand, we can impose on them the demand that they will obey the laws of probability[47].

From a normative point of view, the LBC cannot justifiably be taken as a conceptual truth. I have already mentioned that I do not take it to be either a principle of rationality or a moral principle (soon I will come to discussing the reasons for this). So what can justify it? The answer is that I take it to be an explication of the motivational demand and as such *it is not a demand imposed on moral agents, but rather a demand imposed on philosophers doing ethics*[48]. I do not think that an agent who violates this demand acts either immorally or irrationally. I do not think that an agent who violates this demand is conceptual impossibility, either.

What I do think is that *requiring moral agents to violate this demand* sometimes is something that should be avoided. This is so since if we demand that agents act against what they believe is more likely than not the morally right thing to do, all things considered (e.g. after taking into consideration any information the agents have regarding the degrees of rightness of the acts available and

[47]One might object by arguing that since I want to use degrees of belief in order to determine the agent's preferences, a justification should be given for the claim that beliefs ought to be probabilistic, which does not make use of the agent's preferences. Such a justification is, however, available (see Joyce 1998). In any case, the possibility of relaxing this demand will be discussed briefly at the end of the next chapter.

[48] But, of course, an agent that is involved in a process of achieving a reflective equilibrium is a philosopher doing ethics. He is both a moral agent that tries to decide what he ought to do and a philosopher that tries to formulate conditions that ought to be respected by moral agents. See Rawls (1974), p.7 for a similar point. More will be said about this point later on.

regarding other issues), in most cases they will just ignore this demand, as it amounts to demanding that they act against the judgement they feel more strongly about, the judgement they chose to accept.

Here is another way to put the idea. An ideal moral agent tries the best he can to do the morally right thing. If, after taking into consideration all of the relevant information available to him, he believes it is more likely than not that one act, a, is better than another act, b, then the best he can do is to choose a over b. Choosing b over a will make him vulnerable to the complaint that he could have chosen another act, such that he himself believes it is more likely than not the case that it is morally better than b. To avoid this complaint, the agent must choose a.

Now it is tempting to say, in light of this consideration, that it is not only permissible for the agent to choose a, but also obligatory to do so (either by rationality or by morality), but I do not argue for that (for reasons soon to be presented). My claim here is minimal: an account of moral decision making under conditions of moral uncertainty should not *demand* that agents choose, in some cases, the contrary to what they believe is more likely than not the right thing to do, if it is an account that hopes to actually direct people's choices.

I think that this is quite satisfactory as an initial reason to consider the LBC. However, there are at least two reasons that might make people doubt the LBC. We have already mentioned and discussed one of them in the previous chapter, but it will be helpful to present and discuss it again, as an objection to the LBC.

Second stage: The Maximisation of Expected Moral Value Objection

One immediate objection to the LBC is the following. One might argue that there is one type of case in which it is reasonable to reject this condition: when an agent is uncertain which one of two acts, a and b, is better than the other, but knows that (i) if it is the case that a is better than b then a is much better than b and (ii) if b is better than a, then b is only slightly better than a. In such a case, even if the agent is almost sure that b is better than a, it might be justified for him to prefer a[49].

By now, it should be clear that I do not deny that whenever an agent can use this kind of reasoning, he ought to. The point I have made in the previous chapter is rather that in many cases *one cannot use it*, and these cases are exactly the cases on which I wish to concentrate. Why is it that in some cases an agent cannot use this kind of reasoning? This question brings us back to the discussion in the previous chapter: in order to use this reasoning the agent must be able to attach different levels of goodness to different acts, conditional on some betterness relation holding between them. However, as was explained, if the agent finds himself experiencing intransitive comparative moral judgements in the first place, it must be because he does not have direct access to the levels of goodness of the different acts (because if he had he would just judge

[49] In the literature, the most discussed example for such cases is that of an abortion. It seems that the following description characterises correctly the attitudes of at least some people. These people are almost sure that there is nothing wrong with performing an abortion in the early stages of pregnancy, but also believe that if it is the case that performing the abortion is morally wrong, it is *very* wrong, while if it is true that there is nothing wrong with performing the abortion, not performing it will not be a morally wrong thing to do (or at least not as wrong as performing it in case it is wrong to perform it). See Lockhart (2000) for a discussion.

one act to be better than another iff its level of expected goodness is greater than the other's, which guarantees transitivity).

Now, one might argue that although an agent does not have direct access to the level of goodness of the different acts available to him, he does have direct access to the level of goodness of the different acts conditional on some betterness relation holding between them. In other words, it might be that an agent who is uncertain whether act b is better than act c, but is certain that act a is better than both act b and act c, is also certain, *regarding every possible lottery between a and c*, that if it is the case that act b is better than act c, this lottery is either better or worse than act b. This is what it means, for a rational agent, to have direct access to the level of goodness of the acts conditional on some betterness relation holding between them.

I do not want to argue that such cases never happen. I certainly believe that in many cases, people have partial information regarding degrees of goodness. However, I also believe that in some cases, people do *not* have direct access to a complete goodness function (conditional on some betterness relation holding). In these cases, if they have conditional betterness judgements that respect the axioms of Savage's theory, they can measure the goodness levels they implicitly assign to the different acts. However, if their conditional betterness judgements are not transitive, then there is no goodness function that is consistent with them.

The question I want to explore is how an agent who finds himself in such a position, and who accepts transitivity as a requirement of rationality, should construct his preferences. Arguing that he should do so by always preferring the act with the higher expected goodness level is to beg the question: if he knew, for every act, its expected level of goodness, he would not find himself expressing intransitive comparative moral judgements in the first place.

The best way to look at the model presented at the beginning of this chapter is as a combination of two models, the traditional Savage model for decision making under factual uncertainty and a model of how an agent chooses a preference ordering when he suffers from uncertainty regarding the betterness relation. While the former aims to explain how an agent can construct a utility (or goodness, in the context of moral decision-making) function from his preferences over a rich enough set of acts, the latter aims to capture the reasoning of an agent with intransitive betterness judgements, who still accepts Savage's axioms. Thus, to assume that the agent already has access to the goodness function before he chose his preference ordering is to miss the point.

A more advanced version of the same possible criticism of the LBC would be along the lines of the moral uncertainty literature: although we are sometimes uncertain what the morally right thing to do is, it can be argued that this uncertainty can be reduced to a different kind of uncertainty, i.e. uncertainty regarding which moral theory, or general moral claim, is the correct one. There are two different ways to interpret this position.

The first interpretation is a descriptive one; it takes the argument to be that when we do feel uncertain regarding what is the morally right thing to do this is always because, and only because we are uncertain regarding the validity of some general moral principle. I find this interpretation implausible. People can feel uncertain regarding what is the morally right thing to do even if they do not formulate to themselves any general moral principle.

Moreover, people come to believe in moral theories on the basis of these theories' recommendations in specific cases. When we find out that a general moral principle or moral theory we accept leads to an unintuitive choice recommendation in a specific case, this is a reason for us to reject this moral principle in its conclusive form. However, the position according to which, whenever an agent feels uncertain regarding what is the morally right thing to do in some situation, it is only because he is uncertain regarding which moral theory or general moral claim is the right one, does not allow for such a reasoning process to take place, becasue according to this, our judgements regarding what we ought to do in specific choice situations are derived from our judgements regarding which moral theory is the right one, and not vice versa.

One might, however, deny this claim on a descriptive level, but accept it as a normative principle: whenever one finds oneself uncertain regarding the right thing to do in a specific situation, one must try to reduce this uncertainty to an uncertainty regarding which one of several moral principles or theories is the correct one. This interpretation of the claim is along the lines of my reconstruction of Lockhart's position presented in the previous chapter.

Such a normative requirement is, however destructive for our ethical methodology for the same reason that it is implausible as a descriptive account of the way people do their moral reasoning: it forbids us from using our judgements regarding the right thing to do in specific cases as reasons for accepting or rejecting different moral theories. In other words, it denies not only any version of the reflective equilibrium method, but more generally any form of moral reasoning that aims at moral theories that can have motivational power.

Without being sensitive, in some way or another, to our moral judgements regarding what we ought to do in specific cases, it is hard to see how a moral theory can be motivational. In conclusion, I think the “reducing moral uncertainty to uncertainty about moral theories” claim is implausible at both the descriptive and normative levels, and we must consider cases in which agents suffer from uncertainty regarding the morally right thing to do in a way that cannot be reduced to uncertainty regarding the right moral theory.

For our purposes, it will be best to assume, then, that the agent in our model - after incorporating all of his (partial) beliefs regarding the moral values of the acts available to him conditioned on some further assumptions - still experiences some uncertainty regarding which act is better. This *must* be so, since he has intransitive comparative moral judgements and is still committed to transitivity. In such a case, therefore, commitment to the LBC simply amounts to making use of all the information available and this is, I think, a plausible thing to do.

Here is another way to make the same point. The “maximizing expected moral value objection” seems very simple and very compelling but it is actually somewhat ambiguous. It can be understood in two ways. First, it can be seen as an argument against the claim that the agent should not accept in a RE a judgement that he believes its negation is more probable than it. Second, it can be understood as an objection to the claim that the agent’s moral preferences in a RE ought to be identical to the comparative moral judgements he accepts.

I think that what makes the “maximize expected moral value” objection to the LBC compelling is the second reading. The LBC does seem to be the most plausible criterion for *accepting judgements* in a RE. What “feels” wrong about it is that it seems to go against the demand to maximize expected moral value in *one’s choices*.

However, the LBC does not go against this demand when the “expectation” in question is relative to the non-moral uncertainty the agent suffers from. To see this, recall that the LBC is supposed to apply to an agent in a RE. In a RE, the comparative moral judgments the agent accepts are consistent with the degrees of moral value he attaches to different acts. Thus, by maximizing expected moral value he always chooses according to the comparative moral judgements he accepts. This is just a matter of consistency, so tells us Savage’s representation theorem. In other words, *in order to maximize expected moral value – relative to the factual uncertainty one suffers from - one’s moral preferences in a RE must be identical to one’s comparative moral judgements.*

Now it is, of course, true that by maximizing expected moral value relative to the factual uncertainty one suffers from one might[50] not be maximizing expected moral value relative to the moral uncertainty one suffers from. However, as explained, in the kind of cases I am dealing with one cannot, anyway, maximize expected moral value relative to one's moral uncertainty since one does not have access to the degrees of moral value of the different acts.

So if the LBC is to be rejected on the basis of the "maximization of moral value objection", it is due to the first reading of it, not the second one. The second reading is, however, extremely unintuitive: it implies that moral reasoning *must* lead one to accept judgements that one believes are probably wrong. This leads me to the third stage in my justification of the LBC which has to do with the concept of "acceptance".

Third stage: The Lottery Paradox over Comparative Moral Judgements Objection

It is straightforward to see that the LBC on its own is not enough to ensure that the rationality demand will be satisfied, since if an agent believes that it is more likely than not that act a is morally superior to act b, that it is more likely than not that act b is morally superior to act c, that it is more likely than not that act c is morally superior to act a, and obeys the LBC, then he must violate the transitivity axiom.

[50] One way to look at what I am trying to do in this chapter is as trying to find an answer to the question: under which conditions regarding one's moral uncertainty it is possible for one to maximize expected moral value relative to both moral uncertainty and factual uncertainty.

Notice that such cases are possible even when the agent believes with certainty that the betterness relation is transitive, i.e. when the agent gives a positive probability only to orderings of the alternatives. For example think of a case with a formal structure equivalent to a “Condorcet paradox”, such as the one that is described in the following table.

$^1/_3$	$^1/_3$	$^1/_3$
A	C	B
B	A	C
C	B	A

Table 6

Here, the agent believes with probability $^1/_3$ in each one of three possible ordering of three acts. Although the agent assigns a positive probability only to orderings, he believes with probability $^2/_3$ that act a is better than act b, that act b is better than act c, and that act c is better than act a. Thus, following the LBC leads him to intransitive moral preferences.

This phenomenon can be viewed as an instance of the Lottery Paradox introduced by Henry Kyburg (1961) in which an agent finds himself in the position where he must accept a set of inconsistent judgements. In our case, these are the judgements that one’s moral judgements ought to be transitive

(which we can assume the agent believes with a probability of 1) and the three judgements of the kind presented in the previous paragraph.

The Lottery Paradox is best viewed, I think, as a paradox about acceptance. Indeed, this is the way the paradox is usually presented in the literature. It is worth mentioning, in this respect, that the paradox does not depend on the LBC. Any “threshold account” of acceptance will do (i.e. any account that demands that in order to accept a proposition as true, one should believe it with degree that exceeds some threshold). Indeed, many philosophers take the lesson of the paradox to be that threshold accounts of acceptance are false (see for example Harsanyi 1985, Maher 1993, chapter 6).

In order to truly assess whether such a position is justified, one ought to at least partly characterise what “to accept a proposition” means. Thus, in the literature, much of the discussion regarding the issue is formulated not in terms of the question of what the rational response is for an agent who finds himself in a lottery paradox, but rather, in terms of what the Lottery Paradox teaches us about the concept of acceptance and the rules that govern it (see for example Douven and Williamson 2006, Lance 1995, Jeffery 1956 and 1992).

Most of these discussions take place in the context of the philosophy of science and pertain to the question of theory or hypothesis acceptance. In this context, Bayesians usually adopt a position that either rejects the mere idea of having an attitude of acceptance, or undermines its importance. The reason for this is that, for a Bayesian, a binary relation of accepting/not accepting or believing/not

believing has no role either in decision-making or in reasoning. In order to make decisions, a Bayesian uses his degrees of beliefs together with the utility (or desirability) of the possible outcomes and, when no choices have to be made, the Bayesian takes his degree of belief in a proposition to exhaust all that can be said about his mental attitudes regarding the proposition (see Lance 1995 for a discussion).

It is easy to see that such a reply to the Lottery Paradox is not available to the Bayesian (and to us, by virtue of our commitment to the rationality demand) in our context. In our context, “to accept a judgement” means to be willing to act on it. To accept the judgement that a is morally superior to b means to choose, when acting as moral agents, a over b. This is so since, as was discussed at length, no utility values are available and, thus, we cannot choose an act by maximising expected rightness. What we are after is a way to choose an act, making use only of degrees of beliefs, which will be consistent with maximising expected rightness, according to some rightness function.

This is also the reason why Kyburg’s original suggestion for dealing with the paradox is not available to us here. His solution was to reject the “Conjunction Principle” according to which, if one rationally accepts two propositions one ought to also accept their conjunction. One can accept or reject this principle as a principle of correct reasoning, but in any case, one must make choices when facing moral decision problems and, if one accepts the rationality demand, one’s choices have to be consistent. Thus, I believe that in our particular context, the Lottery Paradox cannot be dealt with on a conceptual level, i.e. it

cannot be taken only as a challenge to different conceptual frameworks designed to deal with epistemological questions. Rather, it constitutes a serious problem for a theory of moral *decision making* under conditions of moral uncertainty, such as the one I am trying to develop here[51].

How should this problem be dealt with? It is clear that choosing the first horn, that of giving up transitivity, amounts to giving up the rationality demand. What about choosing the second horn, i.e. that of violating the LBC? I have argued that this amounts to violating the motivational demand. The argument was that in most cases, a morally motivated agent who is required to violate the LBC, regarding a specific choice, will just violate this requirement. This will be the case at least when the agent has no specific reason not to obey the LBC.

Now, one might argue that when it comes to an agent who finds himself in a Lottery Paradox over his comparative moral judgements, there is such a reason, namely that by obeying the LBC, the agent will necessarily find himself violating another moral judgement he holds, which is the moral judgement that he ought to choose consistently, when making moral choices. He will be violating this judgement by violating transitivity. Does this consideration give us a reason to reject the LBC? Not on its own, I will argue now.

[51] Another way to state the problem is the following. The Lottery Paradox's real message is that when one moves from working with attitudes that come with degrees to working with binary attitudes, something is lost. Thus, the Bayesian response to the paradox is that one should never work with binary attitudes. However, choices, arguably, unlike attitudes, are binary by their nature and thus the Bayesian has no alternative but to make this move. Considering the problem from this point of view reveals, I think, that to the extent that one can give a plausible interpretation to "degrees of choices", one should be able to avoid the problem. One natural interpretation of this sort it to take the chance a specific act gets in a lottery as the degree to which this act is chosen by the agent. This will be discussed in chapter 5 and the discussion will indeed point to a possible solution to the problem.

Implicit in the objection to the LBC presented in the last paragraph is the assumption that the moral judgement that one ought to always choose consistently (in the Bayesian sense) ought to always have priority over any other moral judgement. However, this assumption is dubious. The rationality demand gets its normative force from our belief that rationality is a guide for choices that will best serve the agent's interests (moral interests in our case). The rationality here is instrumental rationality: there is no substantive moral value in obeying its demands. The moral value of obeying its demands comes from the further belief that obeying its demands will best serve other purposes that do have intrinsic moral value.

However, when an agent believes it is more likely that one act is better than the other than that the other is better than it, it is clear that what will best serve the agent's moral interests, in the absence of any information about degrees of rightness, is to choose this act over the other. Demanding that such an agent does otherwise, in the name of transitivity, amounts to putting the cart before the horse. It amounts to demanding that the agent give priority to a moral judgement that gets its moral force from more fundamental moral judgements over one of these more fundamental moral judgements.

One might argue that, from a wider perspective, giving such a priority is justified, since by choosing in an intransitive way, the agent exposes himself to "money pumps", or in the moral context, to "positive moral value pumps". However, this argument misses the point. If the agent has good reasons, *in a particular case,* to suspect that by choosing intransitively, he will be drawn into a

money pump, then this consideration ought to be taken into account through his assessment of the possible consequences of the acts available to him. The agent's degrees of belief in his comparative moral judgements are based, we have assumed, on all the information the agent has that he takes to be relevant to these judgements. However, the mere possibility of being money pumped, *without having any reason to suspect that this possibility will be actually realised*, should not matter much to an agent who must make a specific decision.

Now, as mentioned, I do not want to claim that the LBC is a moral principle. My claim, rather, is that in order for a moral theory to be motivational, it cannot demand that agents (sometimes) violate it in the contexts I have outlined. My argument above should be understood in this spirit. The mere fact that, for an agent who finds himself in a Lottery Paradox, choosing the act to which he assigns a higher probability of being the right act leads to intransitive choices is not a strong enough consideration to make the agent choose another act. This is so since the transitivity requirement is attractive only by virtue of its contribution to promoting the agent's interests, while for an agent in the situation described, it is clear that it does not fulfil this role.

The point, basically, is that people have a reason to respect the rationality postulates in so far as they believe that these postulates help them to promote the ends they wants to promote. When it is clear that respecting these postulates does not do that, but rather prevents them from choosing an act they do believe (it is more likely than not to) promote these ends better, the

postulates lose their ability to motivate people. So, a moral theory that is motivational cannot demand that people do this.

We have to be careful, though, not to overstress this point. It is not my argument that the rationality demand loses its attraction completely when it conflicts with the LBC. Obeying the rationality demand is very attractive for many reasons that are well discussed in the literature. My point is that all of these reasons apply to complete preference rankings, or to whole sets of comparative moral judgements, not to isolated judgements. We do usually want our moral theories to be consistent, but when it comes to a specific decision, the fact that choosing in a way that we judge to be most likely the right way conflicts with other judgements that we accept, does not matter to us much, as we do judge, all things considered, this way to be, most likely, the right way.

This is, I think, what stands at the heart of the tension between the rationality demand and the motivational demand. The latter applies to moral judgements when taken separately, while the former applies to the whole system of one's moral judgements. The latter demands that the moral theory an agent accepts will be consistent with the agent's judgements, while the former requires that it will be internally consistent. We want a moral theory to prescribe only choices that a moral agent will be willing to take, but we also want these recommendations to obey the rationality axioms. However, when an agent finds himself in a Lottery Paradox, the two demands are in conflict with one another.

Thus, it seems that this observation offers us an answer to the question as to whether the rationality demand and the motivational demand can be reconciled. The answer, it seems at a first glance, is “no”, in the general case, as finding oneself in a Lottery Paradox is possible. However, I think this answer is too quick since, for an agent that realises this, it might be possible to avoid Lottery Paradoxes by either choosing degrees of beliefs that do not lead to a Lottery Paradox, or, in the case that he finds himself in a Lottery Paradox, changing his degrees of beliefs in such a way that the paradox disappears. This is, I think, as discussed in chapter 1, the real idea behind the reflective equilibrium method.

The reflective equilibrium method is not merely the demand for a coherent set of beliefs. It cannot be just that since, when dealing only with partial beliefs, as in most cases in reality, this demand is empty. No set of partial beliefs that obey the laws of probability can be incoherent. Incoherence can emerge only between full beliefs or accepted judgements. The reflective equilibrium method tells us that when we find ourselves accepting conflicting judgements we have to change some of them, but it was argued in the first chapter that this demand too is too weak as it leaves the reflective equilibrium method with almost no bite. However, taking the reflective equilibrium method to make use of degrees of belief opens up the possibility of adding a bit of bite to it, by formulating consistency conditions on the way one changes one’s degrees of belief.

In chapter 1 I have quoted some of the scholars that wrote about RE and argued that this idea might be what they had in mind all along. Norman Daniels, in particular, has explicitly argued for it. Daniels, however, did not develop the

idea at all after presenting it. Here, I am trying to actually do what Daniels proposed by suggesting two such constraints. I have argued that the LBC is one of them, and I will present another shortly.

Here, however, a delicate and very important matter arises. I have argued earlier that the LBC is not a rationality condition, but rather, an explication of the motivational demand. However, now I suggest taking it to be a consistency condition that might guide one in one's search for a reflective equilibrium. Isn't being a consistency condition just being a rationality condition?

The answer is: not in the sense that I use the terms here. I will explain this now.

Fourth stage: Why isn't the LBC a condition of rationality?

It was mentioned earlier that when a moral agent is involved in moral reasoning he is both a moral agent and a moral philosopher. He is concerned both with figuring out what he ought to do and with doing this in a way that he finds philosophically defendable. Rawls had the same idea, although he used it for a slightly different purpose. He wrote "... in studying oneself, one must separate one's role as a moral theorist from one's role as someone who has a particular conception" (Rawls 1974, p.17). Now, I will argue that as a moral agent it is not irrational to violate the LBC, but as a moral philosopher who believes that our moral theories should be motivational, it is inconsistent to accept, in a reflective equilibrium, principles that force the agent to violate the LBC.

The second part of this claim seems unproblematic. Since the agent, as a moral philosopher, accepts the motivational demand and since, as I have argued here, the LBC is an explication of this demand, accepting a theory that demands that agents violate the LBC amounts, for such an agent, to being inconsistent.

The first part of the claim is the delicate issue. One might think, at a first glance that it must be the case that the LBC is a rationality principle, since even if one finds oneself in a Lottery Paradox, as long as one is committed to transitivity, that is as long as one is unwilling to give up the moral judgement that one's judgements ought to be transitive, then *taking this commitment into account* one must not be in a Lottery Paradox.

Think of an agent who finds out that he believes it is more likely than not that a is morally superior to b, b to c, and c to a. However, the agent is unwilling to violate transitivity in his choices and thus he concludes that he must act against at least one of these beliefs. Let us assume it is the belief that c is better than a. Now, so the argument goes, since he chose this because he believes in transitivity, it means *that all things considered he believes he ought to choose a over c,* and not, as was the case before he realised that he was in a Lottery Paradox, c over a.

However, this argument is flawed. It assumes that for an agent to accept a proposition (or to "believe", as the word is used in a qualitative way, a proposition) the agent must believe this proposition to be true to a higher

degree than he believes it is false. However, *this is exactly what the argument aims to establish*. The point is that finding oneself in a Lottery Paradox is not something that necessarily indicates a mistake in one's reasoning. Sometimes the evidence is such that it supports having degrees of belief that constitute a Lottery Paradox (as in the original Lottery Paradox, when it is justified to believe with high probability regarding every single lottery ticket that it will not win the lottery, and still to believe with probability 1 that one of the tickets will win).

This point can be put more formally by realising that the condition:

p("a**>**c"│"p(a**>**b) > p(b**>**a)" ∩ "p(b**>**c) > p(c**>**b)") > p("c**>**a"│" p(a**>**b) > p(b**>**a)" ∩ "p(b**>**c) > p(c**>**b)")

i.e. the condition that says that the preferences of an agent who obeys the LBC are transitive does not follow from the condition:

p("a**>**c"│"a**>**b" ∩ "b**>**c") **=** 1

i.e. the condition that the agent believes with probability 1 that his moral preferences ought to be transitive. One can believe with certainty that one ought to choose in a transitive way, but still accept intransitive judgements, by respecting the LBC, and, moreover, one can be justified in doing so.

Thus, I conclude that the LBC is not a rationality principle. This is good news, since if the LBC were a rationality principle, then the result that will be

presented in the next chapter would imply that it is either the case that one of the other postulates of rationality is flawed or that one can be rational only in trivial cases. Neither of these conclusions is an attractive one, for obvious reasons.

The LBC should not be taken as a moral requirement either. This is so because if it is a moral requirement then it is only by virtue of it being a rationality demand and the demand that one ought to be rational when acting as a moral agent. Since the LBC is not a rationality demand, it is not a moral demand either.

At the same time, as I have explained, it *is* a consistency condition that we should aim to obey if we want the account of moral decision making and moral reasoning under conditions of moral uncertainty we formulate to be motivational.

Thus the reason why one should restrict oneself to degrees of beliefs that do not allow for a Lottery Paradox to emerge is not because it is irrational to do so, but rather because when one is in search of a reflective equilibrium, one is looking for an ordered pair of a theory and a probability distribution over all possible comparative moral judgements such that (a) the theory respects the rationality demand and (b) one will be willing to follow the theory's recommendations. In order to find such an ordered pair, the agent must restrict his degrees of belief in the required way. This can be understood as a restriction on one's prior distribution of probability over the possible "morally

superior to" relation, i.e. as the demand to choose a prior that cannot lead one to a Lottery Paradox.

There is something that might seem a bit artificial about this way of thinking about the issue. When engaged in moral reasoning, people usually do not form a prior over the set of all possible orderings of the alternatives available to them. Moreover, it is common practice in both moral philosophy and everyday moral reasoning to use judgements regarding what the morally right thing to do is in hypothetical situations in order to better evaluate what the morally right thing to do is in an actual situation. Therefore, there is no reason to assume that there is even a fixed algebra over which moral reasoners can form a prior probability distribution (as they constantly enrich the algebra by thinking of more and more hypothetical alternatives).

A more plausible description of moral reasoning would be to take the reasoner to formulate for herself what can be described as restrictions on the set of priors she takes to be acceptable. For example, consider an agent who considers only three possible acts (hypothetical or actual), a, b, and c. After using all of the moral information available to her, the agent might come up with judgements like "it is more likely than not that a is morally superior to b", "it is more likely than not that b is morally superior to c" and "it is more likely than not that c is morally superior to a". These three judgements can be understood as restrictions on the set of possible priors the agent might consider. They are the restrictions that demand that the prior that ought to be adopted by her (over the set of all possible orderings of a, b, and c) will be such that the sum of the

probabilities assigned to the set of all the orderings that rank a over b, will be greater than 0.5, and the same goes for b over c and c over a.

From this point of view, the idea of avoiding Lottery Paradoxes by taking the motivational demand and the rationality demand as restrictions on permissible priors, becomes straightforward: if the agent realises that she holds a set of restrictions that does allow for a Lottery Paradox, she must change them so that she will be sure she will not be drawn into a Lottery Paradox.

In Bayesian statistics, such a practice is very common when coming to choose a prior. One starts with a set of restrictions on the possible priors one believes it is reasonable, on the basis of some background knowledge one has regarding the phenomenon studied, to assign to the possible hypotheses. If, after analysing the mathematical relations between these restrictions, one realises that accepting all of these restrictions leads to a violation of another condition (that one did not take into consideration initially), one has to give up on one of these restrictions.

In our context, the “background knowledge” on the basis of which the agent chooses restrictions on the possible priors includes, possibly among other things, the agent’s commitment to the two demands, and thus, the agent should use the restriction that the set of possible priors does not include any prior that leads to a Lottery Paradox.

Although I find this way of viewing the matter both useful and natural, some Bayesians have rejected the idea of “choosing a prior” based on other considerations, as they hold a position that Richard Jeffrey describes as “radical probabilism”, according to which it is not necessary that “...probabilities be based on certainties...”, but rather “...it can be probabilities all the way down, to the roots” (Jeffrey 1992, p.11). Such Bayesians will prefer to describe the process of choosing a probability distribution that does not lead one to a Lottery Paradox differently.

It is not the case, they would argue, that the agent looks for a prior that respects this demand. The probability distribution that the agent holds at any specific point in time is just the probability distribution that best describes his beliefs. The agent cannot choose a probability distribution, but rather, it is given to him by his actual beliefs.

However, even such Bayesians (and indeed even Jeffrey himself) allow agents to change the probability distribution they hold after gaining new information.

Jeffrey suggested a generalised conditionlsation formula, known by the name of “Jeffrey conditionalisation” (what Jeffrey himself called “probability kinematics”) as a way to rationally update beliefs when the evidence is uncertain. This is exactly what happens in our context. We can think of our agent in the following way; the agent considers two acts, a and c, and uses the information available to him to assess whether it is the case that a is morally superior to c, or vice versa. Now, suppose that after contemplating the issue for some time, the agent

assigns a probability value greater than ½ to c being morally superior to a. However, after doing so he realises that by doing this and by respecting the demand that he ought to always prefer the act to which he gives a higher probability of being the right act, he is drawn to intransitivity, as he remembers (or realizes after thinking about the matter) that he also believes with probability greater than ½ that a is morally superior to a third act, b, and b is morally superior to c.

This agent has to decide how to change his degrees of beliefs about the three propositions so that he will not have to violate transitivity in his choices. Assume that he decides to change his degrees of belief regarding a and c such that he will believe it is more (or equally) likely than not that a is morally superior to c (although he does not have to do this, this is just one possibility). Having decided this, all he has to do is to choose an exact probability value for this proposition and after doing that the probability values for a being morally superior to b, and for b being morally superior to c are fixed by Jeffrey's formula.

Jeffrey conditionalisation, unlike classic Bayesian updating, does not tell the agent exactly how he should choose a new probability distribution, but rather restricts the range of probability distributions available to him, using his old probability distribution. In our context, it will be desirable to find a way to characterise the set of all probability distributions that are allowed by Jeffrey conditionalisation and that do not lead to a Lottery Paradox. Doing so amounts to "solving" the Lottery Paradox in our context, as now the agent has a way to keep transitivity while also keeping the motivational demand.

The natural step one might take at this point is to look for a way to characterise sufficient and necessary conditions for updating one's beliefs using Jeffrey conditionalisation that will ensure that no Lottery Paradox will emerge. In what follows, I will not do that, but instead I will assume there is such a way and will impose another consistency condition that any probability distribution *that results from such a process* must respect. I will then show that the set of probability distributions that respect this condition, as well as the LBC, is trivial. Thus, the task of characterising conditions for belief updating that will make the agent immune to Lottery Paradoxes loses its attraction.

The Expectation of Betterness Constraint (EBC)

The LBC describes the relation between degrees of belief in betterness judgements and preferences. It says nothing about the relation between degrees of belief in betterness judgements regarding constant acts (or outcomes) and degrees of belief in betterness judgements regarding acts with uncertain outcomes. I will now present and justify a condition that describes the latter relation.

First, formally:

Expectation of Betterness Constraint (EBC): For every two acts, a_i and a_j,

$$q(a_i > a_j) = \sum_{wk:ai(wk)\neq aj(wk)} p(w_k)q(a_i(w_k) > a_j(w_k)) \,/\, \sum_{wk:ai(wk)\neq aj(wk)} p(w_k) \ .$$

In words: the agent's degree of belief that act a_i is better than act a_j is equal to the normalised weighted sum of his beliefs that the outcome that a_i gives in any specific state is better than the outcome a_j gives in this state, when the weights are just the probabilities of the different states in which the two acts give different outcomes. States in which the two acts give the same outcome are ignored, according to this rule, by the agent, as in these states he is indifferent between the two acts.

Before I discuss the justification for this condition, it might be useful to demonstrate how it works, using an example. Consider the following table.

	$p(\omega_1) = 0.2$	$p(\omega_2) = 0.3$	$p(\omega_3) = 0.4$	$p(\omega_4) = 0.1$
a_i	A	B	C	B
a_j	B	C	A	B
a_A	A	A	A	A
a_B	B	B	B	B
a_C	C	C	C	C

Table 7

Suppose the agent's degree of belief that outcome A is better than outcome B (that is that act a_A is better than act a_B) is 0.7, that his degree of belief that B is better than C is 0.8 and that his degree of belief that A is better than C is 0.9. What should his degree of belief be that a_i is better than a_j? According to the EBC it should be $(0.2 \times 0.7 + 0.3 \times 0.8 + 0.4 \times 0.1) / 0.9 = 0.4666$.

Here is how the calculation goes: first the agent should check in which states the two acts give the same outcome and ignore these states. In our example this only happens in state ω_4. Next, the agent should give each of the remaining states a weight that is equal to its probability and add up his weighted degrees of belief that act a_i is better than act a_j[52]. Lastly, he should normalise this sum by dividing it by the sum of the probabilities of all the states he did not rule out in the first stage. This last move is necessary in order for the agent's degrees of belief to be probabilistic.

Intuitively, the EBC says that one's degree of belief that one act is better than another should be equal to one's expected degree of belief that this act is better than the other, in case one of the two acts is better than the other. In other words, one should believe that one act is better than the other exactly to the extent that one believes the world is such that this act is better than the other.

This seems very intuitive to me, but we can say a little more about what exactly is intuitive about it. The EBC can be seen as the equivalent condition, in our simple extension of Savage's model, for the conjunction of two independence assumptions Savage made in his original model. These assumptions are the separability assumption (or the Sure-thing Principle or the Independence axiom) and the Option-Independence assumption (the status, in our model, of the third independence assumption Savage made, the State-Independence assumption, will be discussed shortly).

[52] Notice that here I used the assumption that two acts cannot be equally good. It is easy to see that if we will relax this assumption, the EBC will have to be slightly adjusted, but nothing significant will change. See footnote 49 for further discussion of this point.

The separability assumption states that the value the agent attaches to an outcome (i.e. to an act given a state) should be independent of what the act brings about in other states. The Option-Independence assumption says that the probability the agent attaches to a state is independent of the act the agent chooses. Since a state, in Savage's framework, can be seen as an assignment of outcomes to every possible act, the Option-Independence assumption can also be expressed in the following way: the value the agent attaches to an outcome (i.e. to a state given an act) should be independent of what other acts bring in this state.

In our model, the issue is not how the agent values outcomes, but rather how he forms his beliefs regarding the betterness relation between outcomes. Hence, we need analogous assumptions to Savage's original ones. This job is done by the EBC in the following way.

The separability assumption is expressed in the EBC by its commitment to taking the probability $q(a_i(\omega_k)>a_j(\omega_k))$ as the agent's degree of belief that act a_i is better than act a_j in state ω_k. That is, by its not allowing the agent's degree of belief that act a_i is better than act a_j in state ω_k to be influenced by the agent's degrees of beliefs regarding the betterness relation that holds between the outcomes the two acts bring about in other states.

The Option-Independence assumption is expressed in the EBC by its commitment to giving each one of the states in which the two acts bring about different outcomes, a weight which is equal to its (normalised) probability. That

is by its not allowing the agent's degree of belief that act a_i is better than the act a_j in ω_k to be influenced by the agent's degrees of beliefs regarding the betterness relations that hold between other acts in ω_k.

It is easy to see that the EBC, taken together with the LBC, ensures that the separability assumption will be respected and, in the same way, that the Option-Independence assumption will be respected. Both of these assumptions have been criticised in the literature. Here I am not going to go into the details of these debates (although in chapter 5 I will discuss, from an unusual perspective, one issue that is usually discussed, namely the moral value gained by using lotteries). Both of these assumptions are necessary in order to get the representation theorem under Savage's framework and, as I use Savage's framework in this thesis in order to examine the compatibility of the motivational demand and the rationality demand, I am committed to both.

It might be the case that by relaxing these two assumptions, that is by postulating another, maybe more permissive, principle that describes the relation between degrees of belief regarding the betterness relations between constant acts and degrees of belief regarding the betterness relations between acts with uncertain outcomes, instead of the EBC, one will be able to escape the result that will be presented in the next chapter, while still being able to represent an agent that obeys this principle and the LBC as an expected moral value maximiser. However, I doubt that this will be the case, as both the separability and the Option-Independence assumptions are essential to Savage's theorem (I will come back to this point in the next chapter).

However, while the separability assumption is essential for any representation theorem, the Option-Independence assumption is not, and so it might be argued that if I chose to work within a different framework than Savage's, then the result I will present in the next chapter might be avoided. Moreover, while the separability assumption is usually taken to express a principle of rationality, the Option-Independence assumption is not always treated this way and thus, violating it should not necessarily be taken as a violation of the rationality demand.

Thus, there is one possible objection to the EBC, in its role as the Option-Independence assumption, that is not part of the usual debate regarding the separability and Option-Independence assumptions and that arises specifically from the introduction of degrees of belief over betterness relations. In order to avoid misunderstandings, I will address this now.

Specifically, the objection is an objection to the assumption, implicit in the EBC, that the probability that one act is better than another, in the case that a specific state is the actual state, is equal to the product of the probability of the state and the probability of the outcome of the first act in this state being better than the outcome of the second act in this state. It might be argued, that is, that our normative beliefs are not independent from our factual ones in this way. However, the EBC does not allow for such dependency.

I do not find this objection very worrying. First, notice that it does not follow from a mere commitment to a position according to which the truth value of moral claims is constituted by factual matters, that our normative beliefs ought to be dependent on our factual beliefs. Both p(.) and q(.) are subjective and thus, one can accept that ultimately normative claims can be reduced to factual ones, without committing oneself to any specific belief about the exact connection between the two.

What if somebody is committed to some such specific connection? Well, first notice that in order for it to really constitute a problem for the EBC, it must be a very weird connection: it is not enough to have beliefs about some connection between facts and norms, rather it should be a belief about a connection between the factual aspects that the agent takes to be relevant for his moral assessment of the acts available to him, and his normative beliefs. This is so since the states do not incorporate every true fact about the world; only those facts that the agent thinks are relevant to his decision. Thus, in order for a real problem to arise for the EBC, the agent has to have a belief of the form “the mere fact that the outcome of my choice will be A indicates to me that it is more likely than not that A is better than B” (maybe because God loves me or because being the good person I am, it is unlikely that I will choose a bad act).

While I find this kind of belief unreasonable, I do not think it is irrational to hold them. However, even somebody that does hold such beliefs can still accept the EBC, by adding into the description of the states (and thus adding more states to the states set) all the possible connections between factual matters and

normative matters that he takes to be relevant to the decision and constructing his p(.) so that his beliefs regarding these connection will be taken into account.

There might be some cases in which this strategy will lead to a violation of the rectangular field assumption. I could not, however, find any plausible examples. In any case, if it is possible to escape the result that will be presented in the next chapter by relaxing the EBC in its role as the Option-Independence assumption, that will be quite a significant finding, as what it will mean is that mere consistency conditions impose a specific dependency between normative beliefs and factual ones without any sensitivity to the content of these beliefs. I do not know if this should worry us or fill us with hope, but in any case it is not a route I am going to take in this thesis.

A final remark should be made regarding the status of Savage's third independence assumption, the State-Independence assumption, in my extension of his model. The State-Independence assumption states that the value a rational agent attaches to an outcome (i.e. to an act, given a state) should be independent of the state in which this outcomes comes.

In one sense, this assumption is necessarily violated in my model, due to the fact that outcomes are only evaluated in comparisons to other outcomes, and not independently. Thus, except in trivial cases, there will always be cases in which given one state, the agent will evaluate an outcome (i.e. an act given this state) in comparison to the outcome another act brings about in this state in a

different way than he will evaluate this outcome in comparison to what the same other act brings about in a different state.

However, this is only because that in a different state, the other act might bring a different outcome. However, the EBC still requires that for any pair of outcomes, the agent's degree of belief regarding the betterness relation that holds between them will be the same regardless of the state in which the comparison takes place. This is all that the State-Independence assumption aims to ensure.

To conclude, I do not see any reason to take the EBC to be more problematic than the usual separability and Option-Independence assumptions, and I do not think that relaxing the "Option-Independence role" that the EBC plays is a very promising route to take. Thus, I find the EBC a genuine principle of rationality. I will have a little bit more to say about the status of the EBC in the next chapter, though.

Conclusion

I have presented two conditions that describe the way degrees of beliefs in comparative moral judgements constrain moral preferences (or considered judgements) and constrain each other. I have argued that one of these conditions, the EBC, is a genuine principle of rationality and the other, the LBC, is an explication of the motivational demand.

These conditions are supposed to constrain the degrees of beliefs in comparative moral judgements of an *agent who has already gone through the process of changing his degrees of belief using the reflective equilibrium method.* The idea is that taking into account that the end state of the process of reasoning must obey these conditions, as well as Savage's other axioms, it might be possible to formulate principles of reasoning that will lead to such an end state. This, in turn, can be taken as a formalisation of the reflective equilibrium method.

However, in the next chapter I will show that the set of probability distributions that respect these conditions and Savage's axioms is very limited. Moreover, the distributions that are contained in it are not the ones we would hope to find there. What exactly we should learn from this will also be discussed in the next chapter.

Chapter 4: The Triviality result

Introduction

We have now reached the point where we are able to prove the representation theorem.

The theorem will be introduced in the first section, together with a short discussion of its immediate formal implications; the proof is in the appendix. In the second section I will demonstrate, using an example, how worrying the triviality of the representation is. I will then move, in section 3, to discuss some ways to avoid it. I will briefly comment on some “easy” escape routes and will continue with a more serious discussion of the non-cognitivist and Humean options. The triviality, I will argue, cannot be escaped by embracing either one of these alternatives.

I will finish the chapter by pointing to the possibility that the rationality demand and the motivational demand are, in fact, incompatible. As I am not happy with this conclusion, and still hope to find a way to escape the triviality without giving up on either one of the demands, I will not discuss this option at length. I will, however, return to it in the conclusion of the thesis.

A triviality result

Theorem: given that > obeys Savage's axioms, LBC and EBC hold iff for every three outcomes, A, B and C, such that A>B and B>C,

$q(A>C) = q(A>B) + q(B>C) - ½.$

The "if" part of the theorem ensures that the conjunction of LBC and EBC is consistent with Savage's axioms. This is true even if the set of outcomes is infinite and the set of acts is convex. Thus an agent that respects the LBC and the EBC can be represented as an expected moral value maximiser. This is true, given a probability function over the set of comparative moral judgements, q(.), for every possible probability function over the set of states, p(.).

If we were to allow the agent's moral value function to vary with different probability distributions over the set of states, then the theorem would no longer hold. However, as explained in the previous chapter, allowing for this would amount to violating Savage's state-independence assumption. Moreover, it would expose us to the same kind of issues we considered in chapter 2 in the context of the Desire as Belief Thesis; i.e. worries about the way one's moral value function changes when one updates one's factual beliefs. One reason we moved to Savage's framework was to avoid these problems.

In any case, the demand that one's moral value function be independent of one's factual beliefs is highly intuitive and, as was discussed in the previous chapter, violations of it must involve very unusual moral convictions.

The immediate consequence of the "only if" part of the theorem is that in a reflective equilibrium that respects our two conditions and rationality of preferences, the agent can never be equally certain in his judgements about the betterness relations between any three outcomes, A, B and C, such that he prefers A to B and B to C. Specifically, the agent cannot give probability 1 to all such judgements.

Another consequence is that for any three outcomes, A, B and C such that q(A>B)≥ ½ and q(B>C) ≥ ½, q(A>B) and q(B>C) cannot be both greater than ¾, as if they are, then q(A>C) must be greater than 1. In the same way for every four outcomes A, B, C and D, such that q(A>B) ≥ ½, q(B>C) ≥ ½ and q(C>D) ≥ ½, q(A>B), q(B>C) and q(C>D) cannot all be greater than ⅔, *and in* general *for every n outcomes, $A_n…A_1$ such that $q(A_j>A_{j-1}) ≥ ½$ for every $jε\{1…n)$, all the $q(A_j>A_{j-1})$ cannot be, at the same time, greater than n/2(n-1)*[53]. It is easy to see that as n approaches ∞, n/2(n-1) approaches ½, *so at the limit the agent must be indifferent between all acts* (except between the act that is preferred to every other act and the act which is the least preferred of all acts, regarding which he must be certain that the former is preferred to the latter).

This is very disturbing. In the next section I will explain, using an example, exactly how worrying this is.

[53] This is so since the condition q(A>C) = q(A>B) + q(B>C) – ½ must hold for any three outcomes, and so $q(A_n>A_1) = \sum q(A_j>A_{j-1}) - n/2 +1$, and thus when all the $q(A_j>A_{j-1})$ are exactly n/2(n-1), $q(A_n>A_1) = 1$.

What the result means

In order to demonstrate what the result means, let us imagine an agent who has to choose between three acts that can bring - in different states of the world - three possible outcomes: that all the 100 inhabitants of village A will die, that all 200 inhabitants of village B will die or that all 400 inhabitants of village C will die. Assume that the agent is absolutely confident that it is better to save more people than less, thus, q(A**>**C) = q(A**>**B) = q(B**>**C)=1. However, the choice he has to make is not between the sure outcomes but between the following three acts:

	$p(\omega_1) = 4/9$	$p(\omega_2) = 3/9$	$p(\omega_3)=2/9$
a_i	B	B	B
a_j	A	C	C
a_k	B	A	C

Table 8

The agent is inclined, at first, to choose act a_i as this act gives the lowest expected loss of life, but after thinking about the matter for a while and consulting his friends he is not so sure anymore as other considerations start to play a role in his reasoning: the people in village B are younger on average than the people in villages A and C, they also donate more money to charity so that if they die the total amount of money that goes to charity from the three villages will be reduced to the greatest extent. On the other hand, it seems that the

people in village C will be missed by the people in villages A and B more than the people in villages A and B will be missed by the people in villages A and C and B and C, respectively, and so on: the agent thinks of many different considerations that should, so he believes, play a role in his decision[54].

After he finishes this process and he feels he cannot think of any more considerations he should take into account when making his decision, he looks for a way to weigh up all of these considerations. The problem is that he cannot think of any exact method to do this which he finds justified. Thus he goes to consult with his decision-theoretic expert friend. His friend solves his problem as he tells him that he was thinking about the matter in the wrong way; he was trying to build a "moral utility" function so that he could maximise the expected moral utility of his actions, when in fact, his reasoning should have been done the other way around. What he should do is to use his judgements regarding which one of the acts is morally preferred in order to build a utility function that represents these judgements.

The problem is that our agent is not sure which act out of a_i, a_j and a_k is better than which, so he decides to assign probabilities to all the possible betterness judgements and go with the higher probability of betterness, i.e. he decides to respect LBC. He also decides that he should assign these probabilities according to EBC, i.e. he decides that his degree of belief that one act is better than another should be equal to his degree of belief that this act will bring better

[54] Notice that all the possible considerations I have mentioned should apply to the sure outcomes as well. This does not matter for the argument as the result holds for any initial q. If you prefer, just assume that the agent does not find these considerations powerful enough to change his beliefs regarding the sure outcomes, but does feel they have some weight.

results than the other. As he is certain that outcome A is better than B, that B is better than C and that A is better than C, he has all the knowledge that he needs in order to make a decision.

However, since $p(\omega_2)+ p(\omega_3) > p(\omega_1)$ he believes a_i is better than a_j. Since $p(\omega_1) > p(\omega_2)$ he believes that a_j is better than a_k, but since $p(\omega_2) > p(\omega_3)$, he also believes that a_k is better than a_i and thus he has intransitive preferences. What should he do?

Well, one thing the agent can do is to change his degrees of belief in the betterness judgements among the three outcomes. This also makes intuitive sense, as his reasoning has led him to the conclusion that his epistemic system as a whole was inconsistent. Notice that he can do this in such a way that he will still prefer A to B, B to C and A to C and will be certain that for any choice among the acts that can yield only these three outcomes with different probabilities his preferences will be transitive. He simply has to choose degrees of belief that satisfy the condition $q(A>C) = q(A>B) + q(B>C) – ½$. For example, if he wants to keep his degrees of belief in his judgements as close as possible to certainty, he can assign: $q(A>C) = 1$ and $q(A>B) = q(B>C) = ¾$.

Now the agent has transitive preferences over all possible acts involving the three outcomes. In fact, the agent has found his “moral utility” function: as is demonstrated in the Appendix, this function is just the one that gives outcome C utility ½, outcome B utility ¾ and outcome A utility 1(and every affine transformation of it). According to this utility function, the agent prefers act a_k to

act a_i to act a_j. The intransitivity has been resolved since now a_j is not preferred to a_k as q(A**>**C) > q(A**>**B). In other words, the agent was forced to lower his degree of confidence in the judgement that A is better than B as he is committed to transitivity.

Maybe it is better to look at it the other way round: as the agent is committed to transitivity and as he believes it is more likely than not that both A is better than B and B is better than C, then he must be more certain, and the constraint tells us exactly how certain, that A is better than C. However, as there is an upper bound on his degrees of confidence, and as he is absolutely certain that A is better than C, he must lower his confidence in the judgements that A is better than B and that B is better than C[55].

On the face of it, this seems like a positive result. What we have now is a consistent thesis that connects the agent's moral preferences to his moral beliefs, i.e. we have a consistent non-Humean thesis. However, this is not quite accurate. When the agent changes his degrees of belief in the betterness relations among outcomes, he is not only involved in a theoretical exercise aimed at achieving coherence. The way he chooses to change his degrees of belief also has an effect on his preferences among acts regarding other sets of outcomes.

[55] Looking at it that way makes it clear, I think, that even if degrees of confidence in one's moral judgments do not obey the laws of probability, as long as there are upper and lower bounds on them, something like this result will hold. Notice that the equation q(A**>**C) – q(C**>**A) = q(A**>**B) – q(B**>**A) + q(B**>**C) – q(C**>**B) must hold even if q is not a probability measure.

To demonstrate that, assume that our agent has some time before he must make the decision, so he feels that he should reflect more on the issue. His philosopher friend suggests to him that he use a technique common among philosophers: he suggests to him to imagine that there is a fourth possible outcome, D, say that all the 700 inhabitants of the three villages will die, and to check what his moral judgements will be regarding acts that lead, with different probabilities, to any one of the three original outcomes or to the new imaginary outcome (of course, the fourth outcome can be added to the set of outcomes not only as an imaginary exercise but also as a result of some unexpected change in circumstances).

In any case, as the number of outcomes was raised to four, the agent's degrees of beliefs in the betterness relations among the three original outcomes can stay unchanged only if the agent is indifferent between the new outcome D, which he obviously judges to be the worst outcome, and outcome C. Otherwise, the agent will necessarily have intransitive preferences among some acts involving the four outcomes. So if he is determined to prefer C to D he must reduce his degrees of belief in some of his original betterness judgements[56]. As the agent thinks of more and more possible outcomes his degrees of beliefs in his betterness judgements must be reduced more and more, and at the limit he must be morally indifferent between all outcomes, except between the one preferred to every other act and the one dispreferred to every other act, as was explained.

[56] This move is already intuitively unattractive, I think, but its implications for the limiting case are obviously a *reductio ad absurdum* of the method of reflective equilibrium (if one accepts my assumptions).

In other words, if one accepts EBC and LBC and is committed to transitivity in one's moral preferences, a method of reflective equilibrium will push one in the direction of being morally indifferent between any two acts. The method of reflective equilibrium aims at a reflective equilibrium in which all of one's beliefs are consistent with each other. As in many contexts the set of possible outcomes is infinite, i.e. there is a clear method to construct another outcome from any two given outcomes (for instance when the outcomes are possible distributions of goods or money among the members of society over time), it seems that the method of reflective equilibrium aims at moral indifference in these cases.

In chapter 3 I have argued that although lottery paradoxes are possible when one respects the LBC, it might be possible to avoid them by changing one's degrees of belief in comparative moral judgments. I have suggested that this process should be taken as a Bayesian explication for the RE method. Now, however, we see that – in virtue of the EBC – *lottery paradoxes are not only possible in the moral domain but rather escapable only in trivial cases.*

Lottery paradoxes, it was explained, do not indicate a mistake in one's reasoning. They can arise as a rational reaction to misleading evidence. However, when one has all the evidence, no lottery paradox can arise (when one knows which ticket won the lottery, one just believes with probability 1 that this ticket won the lottery).

Thus, the result can be interpreted as implying either one of the following three claims:

1. We can never get access to all the relevant moral evidence, and moreover, the moral evidence that we are exposed to is always misleading.

2. When one is exposed to all the relevant moral information one should be morally indifferent between all acts except two, i.e. morality is such that it only has prescriptive power over the choice between two alternatives, the best possible one and the worse possible one.

3. The real moral superiority relation does not obey the rationality axioms.

I find all of these interpretations extremely unattractive. However, by adopting the first one, one can reasonably argue that the conclusion should be that – since moral evidence is always misleading - we should relax the LBC. By relaxing the LBC, we can keep the assumption that one's moral preferences ought to be identical to one's comparative moral judgements in a RE, only that now some of the judgements the agent accepts in a RE are ones that he believes are probably wrong.

This means, that even in a RE and even after being exposed to all the evidence possible, one *must* sometimes choose against what one believes – all things considered – to be the right thing to do. In other words, any plausible complete moral theory (i.e. a theory that gives prescriptions for every possible choice

problem) cannot be wholly motivational even for ideal moral agents (i.e. agents who are only motivated by moral considerations).

Ways out?

If one is unhappy with the conclusion of the previous section, as I am, one must reject at least one of the assumptions I have used. There are many routes one can take here. I cannot discuss them all, but I will discuss the one that seem to me to have the highest initial plausibility to block the threat of the sceptical interpretation of the result. I will argue, regarding this route, that it is in fact not worthwhile to take it. However, I will point to another possible route that can be taken to constitute a "solution" for the result in a particular context. In the next chapter I will further explore this other direction.

The rest of this chapter will be dedicated mostly to the exploration of these two routes. First, though, I will present a list of other possible directions one might take, and make a few brief comments about them.

In order to avoid the result one can either reject one of the axioms I have used in the model or one of the assumptions, whether implicit or explicit, which I have made when constructing the model. The axioms I have used in the model are Savage's axioms, the LBC and the EBC. What are the assumptions I have made when constructing the model? Well, presenting a full list of these assumptions is, of course, not something that I can hope to establish. However, there are four obvious ones that it will be appropriate to mention here.

First, there are, as discussed in detail in chapter 2, the model's commitments to moral cognitivism and to anti-Humeanism, which I will discuss at length soon. Second, there is the assumption that, given moral cognitivism, moral beliefs come in degrees (i.e. that moral uncertainty is possible), and the assumption that these degrees of belief are probabilistic. Let me begin by commenting firstly on the latter alternative, followed by comments on the other ones.

Rejecting the claim that degrees of belief in comparative moral judgements are probabilistic: conceptually I find this idea quite appealing as it seems, at least on the face of it, that the arguments we have for taking the axioms of the probability calculus to be normative when dealing with factual beliefs cannot be extended to the moral domain in a straightforward way[57]. However, as explained in footnote 44, I doubt (although am not certain) that giving up on this assumption can help one avoid the result (or something like the result), as long as one is committed to the claim that degrees of confidence in comparative moral judgements come in degrees which are bounded from above and below.

Of course, one can reject this last assumption too. This brings us to the next possible route, that of rejecting the mere idea of moral uncertainty.

[57] It is clear why this is so in the case of the Dutch book argument. The Dutch book argument depends on the assumption that at some point in time the outcomes of the bets become certain and the participants gets the prizes they are entitled to. However, one might argue that when betting on the truth of normative claims, this assumption does not hold (this can be justified in several different ways using different meta-ethical positions). I suspect that using similar arguments against other justifications for probabilism (such that of Joyce 1998) will be much harder. As I have not investigated the matter in a serious way, I do not want to take a position regarding the matter. My point is that – from a conceptual point of view – giving up on the assumption that normative beliefs are probabilistic seems to have at least some initial plausibility.

Rejecting the claim that moral uncertainty is possible: in terms of a descriptive interpretation, this claim is obviously false. We do, on many occasions, experience, when considering normative propositions, what some of us choose to describe using the word “uncertainty”: we are less confident regarding some normative judgements and more confident regarding others. Moreover, these different levels of confidence seem to come in degrees.

The question is whether using the term “uncertainty” to refer to this experience is normatively justified. Now, what does it means for it to be *normatively unjustified*? One possibility is that it is unjustified in the sense of demanding that our degrees of belief in normative propositions obey the laws of probability. I have just discussed this possibility. The other possibility is that it is unjustified in the sense of treating these degrees as degrees of belief. It might be, that is, that they are degrees of some other attitude, for instance desires. This is basically the non-cognitivist position. I will discuss it soon.

Other than these two senses, I do not see any other obvious sense in which the claim that moral uncertainty is possible can be normatively unjustified. Of course, one might argue that moral uncertainty is possible but normatively insignificant. That would be the Humean position and I will discuss this position too.

Let me move on, now, to consider the possibility of rejecting one, or more, of the model’s axioms. I will start with Savage’s axioms.

Rejecting Savage's "structural" axioms: this possibility was already discussed in the previous chapter in the context of the justification of the EBC, which is, as explained, the condition that ensures that Savage's independence assumptions will hold in my model. I will make one further remark about the matter when I will discuss the possibility of relaxing the EBC.

Rejecting Transitivity or the Sure-Thing Principle: initially I took these two conditions to be the ones that do not require a justification, as both of them are widely recognised as proper rationality requirements[58]. However, it turns out that by relaxing them, or at least by relaxing transitivity and interpreting the outcomes in such a way that the Sure-Thing Principle cannot be violated, one can get an elegant explanation for the moral value achieved by using a lottery to distribute an indivisible good. This issue will be discussed in the next chapter.

As I will argue toward the end of this chapter, relaxing the two axioms is much less worrying in the context of a *particular moral decision* than in the context of a moral inquiry that aims at a *complete moral theory that gives a prescription for any possible moral decision that agents might face.* Thus, in such a context, I actually believe that relaxing transitivity and/or the Sure-Thing principle is the right move to take.

Relaxing Completeness: I have used two different completeness axioms: completeness of the moral preference ordering and completeness of the moral betterness relation. Relaxing each one of these axioms seems to have the

[58] See Anand, Pattanaik and Puppe (2009), chapters 5 and 6, for good reviews of both the different justifications suggested for each one of these axioms and for their shortcomings.

potential of helping us to avoid the result, as by relaxing it, in just the right way, one might be able to avoid all instances of intransitivity by rejecting the requirement that the agent's moral preferences, or the moral betterness relation itself, be defined over all the pairs of the alternatives over which the intransitivity occurs.

Relaxing completeness can also make sense from a conceptual or meta-ethical point of view, if one endorses the position that some values are incommensurable. However, In our specific context, i.e. in the context of moral *decision-making,* I do not think that relaxing any of the two axioms is a real option: even if one accepts that either the moral preference relation or the betterness relation are undefined over two possible acts, one might still find oneself in a position in which one has to make a choice between these two acts. In such a case, claiming that these relations are undefined over the two acts amounts to claiming that morality does not give a prescription for choosing between these two acts. This in turn, amounts to claiming that, *from a moral point of view, and when it comes to making a decision, one is indifferent between the two acts*[59]*.*

This is not to say, of course, that one cannot have non-moral preferences for one act over another, even if one is morally indifferent between the two acts.

[59] By this I do not intend to suggest that the mental attitude of being morally indifferent between two acts is the same as the mental attitude of being unable to compare the moral value of two acts. I think this claim is false. However, it is still true that when it comes to choice recommendations, the notion of moral indifference and the notion of moral incompleteness are functionally equivalent, that is they carry exactly the same prescriptive information regarding the choice: they do not give any reason to choose any one of the acts *over* the other (this is so even though in the indifference case one has equally strong reasons to choose any of the two act while in the incompleteness case the agent has no reason to choose any of the acts).

Such cases are very common, I think. Here, however, we are interested in the moral preferences relation not the non-moral one, and to claim that this relation is undefined over two acts is just to claim that when one has to make a choice between these two acts, morality does not tell one what to do i.e. *from a moral point of view*, one is indifferent between the acts.

Thus, I believe the completeness assumptions are not what is really at stake here, although it might be important in the context of other moral questions, such as blameworthiness, for example[60].

Relaxing the EBC: I have already defended the EBC in the previous chapter. Here I only want to stress one point regarding the possibility of rejecting it; which is that rejecting it is not enough. One must also supply a replacement for the EBC (possibly formulated using a non-probabilistic measure of certitude of comparative moral judgement) that is *defensible*, *restrictive enough*, and *does not lead to another triviality result*.

[60] It might be worthwhile to add one rather "technical" comment here: For convenience, I have assumed in my formulation of moral decision problems that the betterness relation is such that two acts cannot be equally good (I did not assume that one cannot be morally indifferent between two acts. On the contrary: I have assumed that one is morally indifferent between two acts when one believes to the same degree that each one of them is better than the other). However, if the betterness relation itself is incomplete then, since I have assumed that two acts cannot be equally morally good, it seems that my formulation does not allow for such cases. This assumption, however, was made just for the purposes of convenience. Allowing for the possibility that two acts are equally good (or that none of them is morally superior to the other, or that the moral betterness relation is undefined over them) will force us to make some changes in the EBC condition and will make the necessary and sufficient conditions in the result more complicated (but equally problematic), but nothing conceptually important is involved in it. To see this, interpret the agent's beliefs regarding the betterness relation between two acts as the agent's beliefs regarding the betterness relation between the two acts *conditional on the proposition that says that one of these two acts is morally better than the other*. Since the case when none of these acts is better than the other does not have a bearing on the agent's moral preferences, the agent can just ignore it and concentrate solely on the case when one of these acts is better than the other.

By “defensible” I mean that the restrictions it puts on the way one’s degrees of confidence in comparative moral judgements regarding constant acts are related to one’s degrees of confidence in one’s comparative moral judgements regarding acts that are not constant, do not preclude complete distributions of degrees of confidence in comparative moral judgements over the set of all possible acts, that seem reasonable. My discussion of the EBC in the previous chapter was supposed to show that the EBC *is* defensible in this sense.

By “restrictive enough”, I mean that these restrictions do not allow for complete distributions of degrees of confidence in comparative moral judgements over the set of all possible acts that seem irrational. For example, one possible way to relax the EBC is to deny that degrees of confidence in comparative moral judgements regarding constant acts *restrict in any way* degrees of confidence in comparative moral judgements regarding acts that are not constant. This will allow an agent, for example, to have a very high degree of confidence that one outcome, A, is better than another outcome, B, while also having a very high degree of confidence that the act that brings B with probability 0.99 and A with probability 0.01 is better than the act that brings A with probability 0.99 and B with probability 0.01. If one is willing to accept this, one can escape the triviality result. However, it seems to me that accepting this is very close to rejecting the claim that rationality has anything to say about how to deal with uncertainty.

There might, however, be other (that is, other than the EBC) ways to restrict the set of acceptable distributions of degrees of confidence in comparative moral

judgements that are restrictive enough and defensible. If there is such a way that does not lead to another triviality result, I will be happy to endorse it.

I will now move to a discussion of what seems to me as the most natural route to take, in light of Lewis' discussion of the DBT. Nevertheless, I will argue that it cannot really help us to avoid the implications of the result. I will finish the chapter by considering the possibility of relaxing the LBC.

Non-cognitivism, Humeanism and comparative moral judgements

One of my conclusions from the discussion of Lewis' objection to the DBT in Chapter 2 was that although Lewis' argument fails to refute anti-Humeanism, Lewis' general worry is sound. Now that we have formulated an anti-Humean thesis that blocks all the escape routes from Lewis' result suggested by the different scholars mentioned in that discussion, and still reached a triviality result, it might be argued that this result should be taken to indicate that either anti-Humeanism or moral cognitivism is false.

I will argue against this conclusion now, in two stages. In the first stage, I will argue that adopting a non-cognitivist position does not really solve the problem the result poses to us since it is possible to interpret all the claims that I have made regarding beliefs in comparative moral judgements in a way that most non-sceptical non-cognitivists will have to endorse. Moreover, those who will not endorse it will have to supply a philosophical account that currently does not exist and which there seems to be no natural way to develop. In the second

stage, I will argue that being a non-sceptical, non-cognitivist, Humean exposes one to exactly the same problem the cognitivist, anti-Humean faces in the light of the result. The conclusion will be that the result is not significant to either the cognitivism/non-cognitivism debate or the Humean/anti-Humean debate. What is really at stake, I will further suggest, is the need to choose between the motivational demand and the rationality demand[61].

Let us start with the non-cognitivist option; that is with the claim that the real lesson from the result should be that moral judgements are not beliefs. The first thing to notice is that non-cognitivism indeed seems to be the more natural position for a Humean to adopt[62]. Indeed, when the discussion is restricted to an agent who respects the rationality demand, non-cognitivist and Humean positions seem to be in a better position than cognitivist and Humean ones respectively to explain how moral judgements can motivate, or so it is often claimed[63]. The main idea is that taking moral judgements to be desires rather than beliefs enables one to treat them as motivational without the need to go beyond the demand of expected moral value maximisation in one's account of motivation.

It is the cognitivists, usually, who find themselves in the difficult position of having to tell a story about how moral beliefs can motivate in a way that is

[61] There is one qualification for this conclusion. I will point to one route that still has the potential to help the non-sceptical, non-cognitivist, Humean to avoid the result. However, I do not know how this route can be developed.
[62] See Shafer Landau (2000) and (1998) for discussions. It is worthwhile mentioning, however, that even if this is true regarding the *Humean position*, it does not follow that it is true regarding the *position Hume himself held*. The latter is subject to different interpretations and I do not have any interesting insights to offer regarding the matter.
[63] For example see Rosati (2006).

compatible with the moral value maximisation demand. In order to do this they must reject, as we have seen, the strong Humean position that desires are never constrained by beliefs. However, the picture changes when we start examining, instead of agents who always obey Savage's axioms, ones that sometimes violate them, but still are determined to change their attitudes, when they realise this is the case. In such cases, while the cognitivists have ready-made theories of how such reasoning occurs, namely any theory of belief revision, the non-cognitivists face a problem: they need to present an account of *reasoning with inconsistent attitudes which are not beliefs*, and it is not at all clear what such an account would look like.

To see the point more clearly, consider a non-cognitivist who holds that moral judgements are desires of some kind. Indicating that one accepts "A is right", expresses, for such a non-cognitivist, one's moral desire for A. This desire can come in degrees, of course, but these are not degrees of belief in the proposition "A is right", but rather degrees of desires for A, or for a morally motivated agent, the degrees of rightness he attaches to A. How should such a non-cognitivist treat a judgement that is expressed by sentences like "A is morally superior to B"? He cannot take this judgement to be a belief in the proposition referred to by the sentence, because he is a non-cognitivist. Rather, he would probably take the sentence to express the agent's moral preference for A over B.

Now, if the agent is rational, it follows from this judgement that the agent attaches a higher level of expected moral value to A than to B. However, when

an agent is not rational, for example when he has intransitive comparative moral judgements, there is no desirability function, the expectation of which leads to the agent's comparative judgements. Such an agent has, therefore, inconsistent judgements, but according to the non-cognitivist, this inconsistency cannot be attributed to an error (in the sense of a false belief) the agent has made regarding the degrees of desire he has for different acts. For adopting such a position would just make him a cognitivist who is committed to the claim that degrees of moral desire for an act constitute the degrees of moral value of the act (i.e. he would have to accept that comparative moral judgements are beliefs regarding the ordering that is determined by one's "true" moral desires, which are not directly accessible to one). Rather, a true non-cognitivist must hold that it is an inconsistency between the agent's expected desires, which must lead to transitive comparative moral judgements, and the agent's moral preferences.

How is such inconsistency even possible (that is psychologically)? I am not entirely sure. One possibility is to take comparative moral judgements to be moral desires themselves, only desires of a different kind from the ones expressed in sentences of the form "A is good". John Broome uses the term "comparative desires" (Broome 2006) in this context. The two kinds of desires, the non-cognitivist must claim, constrain each other normatively, but not necessarily descriptively (as if they did, no inconsistency can occur between them).

Another option is to take comparative moral judgements to be something different from both desires and beliefs. Maybe they are, as Allan Gibbard

suggested (Gibbard 2008), plans for future behaviour. For a moral agent to judge that A is morally superior to B is to plan to choose A over B when both acts are available to him. Now, there seems to be nothing conceptually problematic with the idea that an agent has inconsistent plans (but still it is normatively problematic).

Maybe there are other options for the non-cognitivist to take as well. However, the non-cognitivist should be able to explain not only how the inconsistency is possible, but also how it is possible for an agent to reason himself out of it, and this explanation should not be made in terms of beliefs.

Broome (2006) has investigated the problem of reasoning with preferences and presented an outline of a method of doing so. Without going into the details of the suggestion, it is important to note two things about it. Firstly, Broome admits, regarding the claim that reasoning with preferences is possible, that "…much more needs to be done to make the conclusion secure" (Broome 2006, p.15), when the main problem he pointed to is the need to identify "…a criterion for correct reasoning with preferences, as opposed to incorrect reasoning…" (ibid, p.15).

In the absence of such a criterion, and surely we are a long way from formulating a well-agreed criterion like that, most non-cognitivists will accept that when one has to reason with one's preferences, for example when one realises that one holds inconsistent comparative moral judgements, one may be uncertain regarding what is the appropriate way to reason with one's

preferences. So even a non-cognitivist regarding comparative moral judgements can agree that, in face of inconsistency, one may use one's (degrees of) beliefs regarding what is the right way to change one's inconsistent comparative moral judgements, in order to decide how to change them.

This might not make one a cognitivist generally, but for our purposes it will make one cognitivist enough to accept the general strategy suggested here for dealing with inconsistencies. It just means that whenever I treat comparative moral judgements as beliefs regarding the "morally superior to" relation that holds between acts, the non-cognitivist should treat them as beliefs regarding whether it is justified or not (according to some criterion for correct reasoning with preferences) to hold the moral judgement in question.

This seems to be almost a terminological trick and indeed this is, basically, the second issue regarding Broome's attempt that I wish to point out. Broome concluded his paper with a discussion of the relation between preferences and beliefs about betterness. As mentioned before the term "betterness" is usually (that is in his 1991 book) used by Broome in a different way than the expression "morally superior to", but in this paper Broome uses it in a general way according to which it need not be "…betterness from the point of view of the universe. It might be betterness for you, or betterness relative to your point of view, or something else" (Broome 2006, p.15). Therefore, for our purposes, we can just take this to mean "betterness from the point of view of morality".

Broome claimed that “When you use a sentence like ‘Rather walk than drive’ you may well be expressing a belief about betterness, and not a preference…” (ibid, p.15) and admits that “…it is hard to distinguish the functional roles of a preference and a belief about goodness” (ibid, p.16). He also noted that this fact “…explains why many non-cognitivists about value think that a belief about betterness is indeed nothing other than a preference” (ibid, p.16), but concluded that “In so far as the two converge, I am inclined in the opposite direction: a preference may be nothing other than a belief about goodness. It may turn out that reasoning with preferences is nothing other than reasoning with beliefs” (ibid, p.16).

Now, I might add, even if reasoning with preferences is something distinct from reasoning with beliefs, then from a normative point of view, reasoning with preferences ought to be consistent with reasoning with beliefs regarding what is the right way to reason with preferences, and since this is so, one can, instead of reasoning with preferences, *reason with beliefs regarding what are the preferences one ought to have, if one were to correctly reason with one’s preferences.*

A similar story can, of course, be told regarding other interpretations of comparative moral judgements. If they are, as Gibbard purpose, plans, then one has to give an account of reasoning with plans that does not make use of beliefs, and instead of doing that one can just use reasoning with beliefs regarding the right way to change one’s plans, and so on. All that is needed in order for this strategy to be available is to accept the claim that *there is some*

criterion for correct reasoning with comparative moral judgements. Denying this amounts, as was already claimed, to denying the whole idea of moral reasoning: if there is no criterion of correct reasoning, then one can just choose which act is the morally superior one arbitrarily. This amounts to moral scepticism.

The lesson, thus, is the following. The moral cognitivism/non-cognitivism debate is not really an issue in our context, as long as: 1. one is willing to accept that there is a justified way to reason with comparative moral judgements; and 2. one can form beliefs regarding which way this is, reason with them, and *one is willing to accept that the set of beliefs one accepts when the process of reasoning is over ought to be consistent with the set of attitudes one would accept if one were reasoning directly with one's comparative moral judgements* . Denying 1 amounts to scepticism (and in any case, when one's judgements are inconsistent and one is still committed to the two demands, one must accept 1). Denying 2 is a possibility, I admit. However, avoiding the triviality result by denying 2 and holding a non-cognitivist position is not enough. One must also present an account of normatively correct reasoning with attitudes which are not beliefs. To my knowledge, such an account is lacking.

Still, it might be that the real issue is the Humean/anti-Humean debate, which I shall focus on now, not the cognitivism/non-cognitivism one.

A Humean can be either a non-cognitivist or a cognitivist. As mentioned, a more natural position for the Humean is non-cognitivism, but nothing really prevents

him from being a cognitivist. It is just that the cognitivism to which he is committed is somewhat trivial: it accepts that people may have moral beliefs but does not allow these to play any motivating role for their behaviour. What motivate moral agents, the Humean argues, are their moral desires.

How should a Humean address the problem of an agent who expresses inconsistent moral preferences? Well, the Humean has three options. Firstly, he can be what Broome (1999 chapter 5) called a "non-moderate" Humean, that is, he can bite the bullet and insist that, even if the agent explicitly endorses rationality, he cannot choose consistently. This is so because the agent's desires are inconsistent and only desires can motivate. In the terminology that I have introduced here, this amounts to rejecting the rationality demand.

Secondly, the Humean can be what Broome called a "moderate Humean"; that is, a Humean who accepts that when an agent realises that his preferences are inconsistent he ought to change them so that they will become consistent[64]. Now, Broome has convincingly, in my opinion, argued that the moderate position cannot be really separated from the non-moderate one, but even if one rejects Broome's argument, the question of how the agent ought to change his preferences is still open. Now, no matter what answer the moderate Humean gives to this question, it will involve reasoning. The agent is supposed to use his commitment to some consistency conditions, and some of his attitudes in order to change other attitudes.

[64] This is not exactly the way Broome characterizes the moderate Humean position, but in our context this is the right way to characterize the alternative to the non-moderate position.

This brings us back to the conclusion of the discussion regarding the non-cognitivist option: in order to be a moderate and a Humean, one must provide an account of correct reasoning with preferences and deny that one can avoid reasoning with preferences by reasoning with beliefs about what would one's preferences be if one were to reason correctly with them. As I have already mentioned, I do not deny that this is a possibility that might be worth exploring. It is just that I do not know how to do so.

There is a third option for the Humean to take. The Humean might argue that in cases in which he has inconsistent preferences, then even though he should not (or cannot) change these preferences, he can still choose in a consistent way, contrary to some of his preferences. This, however, amounts to relaxing the motivational demand

Thus, I conclude that although blaming the result on either the anti-Humean or cognitivist commitments of the model is still a possibility, it is not a very promising route to take. The last possibility I want to consider now is that of relaxing the LBC.

Are the motivational demand and the rationality demand compatible?

I have avoided questioning, until now, the LBC. This is so since, as I have argued, the LBC is an explication of the motivational demand and so in order to examine whether the rationality demand and the motivational demand are compatible I had to assume that the LBC does hold.

Now it is time, however, to consider the possibility that the motivational demand is in fact incompatible with the rationality demand and that this is exactly what the triviality result reflects. As I have stressed, I do not like this conclusion and I will be happy to reject it in case a plausible solution can be found to the problem which the result poses.

It would be helpful, at this point, to remind ourselves what has led us to where we are now. Our starting point was the tension between the rationality demand and the moral intuitions we have. I have demonstrated in the first chapter that both self reflection and psychological evidence suggest that we will sometimes have inconsistent moral intuitions. This observation has led us to the conclusion that it is not our moral intuitions that motivate us, when we act as moral agents, but rather our considered judgements. Moral intuitions can serve as evidence when we construct our considered judgements, but there are other types of evidence that should be taken into account as well.

The next step in my argument was to identify considered moral judgements with beliefs and to formulate consistency conditions on the way the degrees of these beliefs constrain the comparative moral judgements we end up accepting (that is, acting on) as moral agents. However, it turned out that it follows from these conditions that the comparative moral judgements we end up accepting must be either inconsistent or trivial.

We can now turn the reasoning that has led us to this point on its head: instead of starting with the phenomenon of us having inconsistent moral intuitions and reaching the conclusion that we cannot avoid the inconsistency, we can start with the phenomenon of people who have beliefs (that come in degrees) about what ought they to do from a moral point of view, and reach the conclusion that whatever the degrees of their beliefs are, they must be inconsistent (or trivial). In other words, it is not only true that realising that we hold inconsistent judgement should lead us to be uncertain regarding these judgements, but also that being uncertain regarding our judgements must lead us to accept some inconsistent judgements.

From this perspective, it is not some contingent fact about our psychological structure that is responsible for the tension between moral motivation and rationality; rather the problem lies at the conceptual level. I have already suggested one way to make sense of this claim: motivational demands are external to one's behaviour. They require that one's behaviour be consistent with some set of motivational factors, whatever these are. Rationality demands, on the other hand, are internal to one's behaviour. They require that different choices made by the same agent will be consistent with each other in specific senses.

So it might be that the fact that we sometimes hold inconsistent moral intuitions is not due to some unfortunate causal chain that made us this way. No matter how evolution works, no matter how our intuitions are formed, we will end up having inconsistent judgements.

To make things clearer, imagine an agent all of whose moral intuitions are consistent. It might be argued that such an agent has no reason to form beliefs regarding the question as to what ought he to do in a specific situation. He can just obey his intuitions, and these will make him always choose consistently. It is true that if such an agent starts, for some reason, to form moral beliefs of this sort, and chooses to choose according to them in the way the LBC requires, he must end up choosing inconsistently, but why should he form such beliefs in the first place?

Here is one way to look at this. In a sense, we are all such agents. As many have recognised[65], it is always possible to rationalise a set of choices by individuating the alternatives in a finer way. Strictly speaking, no two choices are choices between the same alternatives. The alternatives are always different in some way: they are located at different points in time; they are available from different menus of alternatives, and so on. Thus, using the finest individuation possible, the rationality axioms can never be violated, since these axioms always describe a relation between different choices among the same alternatives. Using the finest individuation possible, all of us, always, have consistent comparative moral intuitions.

In order to avoid this conclusion, in order to add at least some bite to the rationality axioms, we must admit that there are such things as what John Broome (1991) calls “rational requirements of indifference”. That is, some

[65] See Broome (1991) chapter 5 for a discussion.

conditions that require us not to differentiate between some alternatives. In the context of moral choices, it is better to call these requirements "moral requirements of indifference" as it is not inconceivable that there are cases in which it is rational to differentiate between two alternatives but, nevertheless, morally wrong to do so (or at least morally unjustified)[66]. From this point of view, the rationality conditions are in fact conditional requirements. They forbid some patterns of choices on the condition that the alternatives over which these choices are defined are not to be distinguished according to some requirement of indifference.

In order for us to know whether our choices are rational or not we must, then, make some judgements regarding which features of the world are, and which are not "difference makers" (or "justifiers", as Broome calls them) regarding specific choices.

Consider the case of an agent who *experiences uncertainty in his judgements regarding which features of the world are rational or moral requirements of indifference*[67]. Such an agent necessarily also experiences uncertainty regarding the truth of some comparative moral judgements. This is so, since if there is no uncertainty regarding the betterness relation, i.e. if the betterness relation between *all possible propositions* is known, it is also known, for every

[66] For example, intuitively it does seem rational to differentiate between killing an innocent person and getting nothing in return to killing an innocent person and getting a pound in return, but morally, it is unjustified to do so.

[67] Of course, we must allow for such cases, if we want our account to have any bearing on real life decisions. As Broome writes "I would not expect everyone to agree readily about what a particular requirement of indifference rationality imposes. That is likely to be discovered only by debate" (Broome 1991, p.105). However, until this debate will be settled, we must allow for uncertainty. The case is even stronger when considering *moral* requirements of indifference, as will be soon demonstrated.

three propositions, whether each one of them is a justifier with regard to the other two[68].

Thus, an agent who experiences no uncertainty regarding comparative moral judgements must also experience no uncertainty regarding judgements about which features of the world are rational or moral requirements of indifference. By *modus tollens,* if we allow for uncertainty regarding the latter we must allow for uncertainty regarding the former and then we are back to the triviality result.

A possible response is to deny that one can be uncertain regarding the truth of judgements about which features of the world are moral or rational requirements of indifference. By arguing this, however, one rejects the idea that our moral judgements should be sensitive, in some way or another, to what we take to be relevant moral considerations. That is, one rejects the idea that being exposed to arguments *that one takes to be relevant from the point of view of a moral agent* can make one change one's degrees of confidence in one's moral judgements. And again, in other words; we should not always expect a rational reasoner to reason in a way that is sensitive to what *he himself takes to be legitimate reasons* to have judgements of some sort or another. This seems to me very close to a *reductio ad absurdum* of the mere possibility of moral reasoning, i.e. of the use of reasons in order to reach a conclusion in a moral context.

[68] To see that, notice that a proposition, A, can be a justifier regarding two other propositions, B and C, only if there are cases in which the following holds: Although raising the probability of B to be true is better than raising (to the same extent) the probability of C to be true, raising the probability of "A and B" to be true is not better than raising the probability of C to be true (or vice versa).

The tension between the motivational demand and the rationality demand has reappeared, thus, in a different form: either we have to accept that the rationality demand is empty, i.e. to argue that there are no such things as rational requirements of indifference, or we have to deny this, while accepting that sometimes our choices should not be affected by considerations that we judge to be relevant to the choices we face (or ought to be affected by considerations we judge to be irrelevant to the choices we face).

This conclusion rings a bell; it brings us back to Peter Singer's famous argument that was discussed in the first chapter. Consider the following four propositions:

A: "I will save the life of a child, x, for the cost of a pair of shoes".

¬A: "I will not save the life of a child, x, for the cost of a pair of shoes".

B: "the child, x, is drowning in a pond next to me".

C: "the distance of the child, x, from me does not make a difference from a moral point of view".

Now, as discussed in the first chapter, Singer's argument relied on the assumption that the following conditional is true: if C is true then if it is morally

obligatory to make “A and B” true over making ¬A true, it is morally obligatory to make A true over making ¬A true.

We have seen, however, that if we allow for uncertainty regarding C, we must allow for uncertainty regarding some comparative moral judgement between propositions like A, “A and B”, and ¬A. This, in turn, leads us to the triviality result. In order to avoid this conclusion, Singer must deny that we can be uncertain regarding propositions like C.

Indeed, it is possible to interpret Singer’s position in exactly this way, i.e. as the position according to which we have to assign probability 1 to C even if we have strong intuitions, or other reasons to believe, it might be false[69]. The problem with this position, I have argued, is that you can indeed hold it, state it, preach it to others, but you cannot reasonably expect people to follow it. As such, one may ask, what is it good for?

I do not have a good answer to this question, but I do have a little bit more to say about the matter. I will return to this issue in the Conclusion. In any case, it seems, in light of the previous discussion, that it might indeed be the case that there is an inherent tension between the demand that our moral judgements be internally consistent and the demand that they be consistent with a set of some

[69] See Kamm (1999) for a discussion. In the same way, if instead of assigning probability 1 to C (and by that ignoring any reason one might have to believe it is false) one assigns to it probability 0 (and by that one ignores any reason one might have to believe it is true), one is free to assign any degrees of belief one wishes to the judgements concerning A, “A and B”, and ¬A.

external factors; reasons, motivating states, judgements about what can rationally make a difference, and so on.

If this is indeed the case, then we must choose, in situations in which the two demands are in fact in conflict with one another, which one of them we are willing to give up. I want to suggest that the answer to this question might be context-dependent. Specifically I want to consider two contexts. The first context is that of a moral inquiry that aims at a complete moral theory, i.e. a theory that gives a prescription for every possible choice. I do not have a good answer to the question in this context. I find it unacceptable to give up on the rationality demand, but I also find it hard to give up the motivational demand. In the conclusion for the thesis I will raise some (rather speculative) thoughts I have on the matter.

The second context is that of a moral inquiry that aims at a recommendation for a specific moral decision. In this context, I find giving up on the motivational demand completely unacceptable. If the moral inquiry really aims to direct people's behaviour, it cannot prescribe a choice that ideal moral agents will not be motivated to make.

The only option that remains in such a context is, thus, to relax the rationality demand. This is, however, much less troubling when thinking of the matter in the context of specific moral decisions people might face. Notice that all that the result presented at the beginning of this chapter shows is that when an agent respects the LBC and the EBC he must have, except in trivial cases, intransitive

moral preferences over *some possible acts*. This does not mean, however, that the intransitivity will arise over the acts *that are actually available to the agent*.

If the intransitivity does not arise over the actual acts that are available to the agent, it seems to me, in light of the discussion so far, that the agent should just choose one of the acts that are ranked at the top of his moral preference ordering over the acts that are available to him. He should not be worried, that is, by the fact that his preferences are intransitive regarding some possible acts, maybe ones that he has not even thought of. He cannot do anything sensible about this intransitivity in any case, so the result shows us. Thus, he should just thank his good fortune and choose rationally.

What if, however, one has the bad fortune to find oneself in a choice situation in which one has intransitive moral preferences over the set of acts available to one? Firstly, I think, one should try to avoid the intransitivity by revaluating the moral information available to one and by collecting more evidence (maybe evidence regarding degrees of moral value). What if after doing that, one still finds oneself having degrees of belief such that one's moral preferences are intransitive? I have a suggestion which I will discuss in the next chapter.

Conclusion

I have explored in the last four chapters the route that I find to be the most promising one to take in order to investigate the possibility of arriving at a rational and motivational moral theory. In the first two chapters, I have mainly

defended the decision to take this route from possible and actual objections and explained why other possible routes are unlikely to lead us to a desirable destination. In the third chapter, I have followed this route to the point at which I was able to construct a formal representation of the main question I explore in this thesis. In this chapter, however, it became clear that this route, as promising as it seems to be, leads to a dead end.

Having reached this conclusion, I have argued, firstly, that the failure of my attempt is not due to the two central assumptions I used, namely that value judgements are beliefs and that these beliefs should constrain our moral preferences. On the contrary, by relaxing these two assumptions, in the way non-cognitivists and Humeans do, one would reach the dead end I have reached, much faster. Secondly, I have briefly explored some possible routes to avoid the result by relaxing some of the assumptions I have used.

Without giving up on the hope of still finding another route that will not reach a dead end, I moved on to considering the possibility that all possible routes lead to the same, unfortunate, destination at which I have arrived.

In a sense my inquiry should have ended here. However, I have noticed that, as negative as the conclusion I have reached seems to be, it does have some positive implications. I have tentatively mentioned one of them in the last section: the result can be taken to *explain* why we sometimes have inconsistent moral intuitions and this explanation is conceptual, not evolutionary or psychological. I will mention some other possible positive implications in the

conclusion of the thesis. In the next chapter, though, I am going to discuss in more depth one positive implication.

Appendix

Theorem: given that > obeys Savage's axioms, LBC and EBC hold iff for every three outcomes, A, B and C, such that A>B and B>C, q(A>C) = q(A>B) + q(B>C) – ½ .

1 EBC: for every two acts, a_i and a_j,

$q(a_i > a_j) = \sum_{wk:ai(wk)\neq aj(wk)} p(w_k)q(a_i(w_k) > a_j(w_k)) / \sum_{wk:ai(wk)\neq aj(wk)} p(w_k)$.

2. LBC: For every two acts a_i, a_j, $a_i \geq a_j$ iff $q(a_i > a_j) \geq q(a_j > a_i)$.

Proof:

If:

Since from the conjunction of LBC and EBC we know that for every two acts a_i and a_j , $a_i \geq a$ iff

$$\sum p(w_k)q(a_i(w_k) > a_j(w_k)) \geq \sum p(w_k)q(a_j(w_k) > a_i(w_k)) ,$$

it is enough, since we assume all of Savage's axioms, for us to show that if the constraint that for every three outcomes, A, B and C, such that $q(A > B) \geq ½$ and

$q(B>C) \geq ½$, $q(A>C) = q(A>B) + q(B>C) – ½$, is satisfied then there is a utility function that gives a value to each outcome such that:

$\sum p(w_k)u(a_i(w_k)) \geq \sum p(w_k)(u(a_j(w_k)))$ iff

$\sum p(w_k)q(a_i(w_k)>a_j(w_k)) \geq \sum p(w_k)q(a_j(w_k)>a_i(w_k))$

Or:

$\sum p(w_k)(u(a_i(w_k)) – u(a_j(w_k))) > 0$ iff

$\sum p(w_k)(2q(a_i(w_k)>a_j(w_k)) – 1) > 0$

In order to do this, let us define the utility function in the following way. Let A_1 denote the outcome that is the least preferred of all options (as we assumed that the agent's beliefs regarding the betterness relations among outcomes are transitive there must be one option like that), A_2 the outcome that is dispreferred to all outcomes except A_1 and so on until A_n:

$U(A_1) = ½$

And for every $j≠1$

$U(A_j) = q(A_j>A_1)$

Thus, for every two outcomes A_k and A_m, such that k, m ≠ 1,

$u(A_k) – u(A_m) = q(A_k > A_1) – q(A_m > A_1)$

When either k or m equals 1 we just replace the relevant expression with ½.

Assume, WLOG, k>m, then if m≠1:

$q(A_k > A_1) = q(A_k > A_m) + q(A_m > A_1) – ½$[70] , or:

$q(A_k > A_m) = q(A_k > A_1) - q(A_m > A_1) + ½ = u(A_k) – u(A_m) + ½$

So we get:

$u(A_k) – u(A_m) = q(A_k > A_m) – ½$

and it is straightforward to verify that the last expression holds also when m=1.

So now we know that:

$\sum p(w_k)u(a_i(w_k) – u(a_j(w_k) > 0$ iff

$\sum p(w_k)(q(a_i(w_k) > a_j(w_k)) – ½) > 0$

[70] This is so because the condition q(A>C) = q(A>B) + q(B>C) – ½ must hold for *every* three outcomes, A, B and C, such that A>B and B>C.

And so we arrived at the desirable conclusion that:

$\sum p(w_k)u(a_i(w_k) - u(a_j(w_k) > 0$ iff $\sum p(w_k)(2q(a_i(w_k) > a_j(w_k)) - 1) > 0$.

Only if:

First, notice that since we assumed that the agent satisfies all of Savage's axioms, there is a (unique up to affine transformation) utility function, u, such that when the agent maximises his expected utility relative to this function and to p he gets his preferences (we have just constructed this function).

Consider now the following two acts:

	ω_1	ω_2	ω_3
a_i	A	B	C
a_j	B	C	A

Table 9

When $p(\omega_1)=p(\omega_2)=p(\omega_3)$ the agent must be indifferent between a_i and a_j since $Eu(a_i) = Eu(a_j)$ for all utility functions. Thus, it must be true also that:

$q(A>B) + q(B>C) + q(C>A) = q(B>A) + q(C>B) + q(A>C)$

or (since we assumed that for every two acts, a_i, a_j – and so for every two outcomes, A, B – either A>B or B>A) :

$$2q(A>B) -1 + 2q(B>C) -1 = 2q(A>C) -1$$

And we get:

$$q(A>C) = q(A>B) + q(B>C) – ½$$ [71]

This equation must hold for every three outcomes A, B and C such that q(A>B) ≥½, q(B>C) ≥ ½ (and thus q(A>C) ≥½) and for every q that always (i.e. for every p) yields transitive preferences.

[71] And it is easy to see that this means that q(A>B) + q(B>C) ≤ 1.5, i.e. there is an upper bound on how certain the agent can be regarding these two judgements.

Chapter 5: Doing the best one can and the rightness of lotteries

Introduction

Many people share the intuition that, in some choice situations, using a lottery among (some of) the alternative courses of action open to an agent is the morally right thing to do. In the philosophical literature, several justifications for this intuition are presented. John Broome's well-known justification for this intuition[72] is based on the idea that what makes a lottery the morally right thing to do (when it is the morally right thing to do) is that it is fairer than any of the definite choices available to the agent. Thus, Broome's explanation of what makes a lottery right has two parts: first, he presents an account for the fairness of lotteries and second, he argues that in some situations the fairness consideration is strong enough to make the fair act the right act.

In this chapter, I will present a new justification for the rightness of lotteries, which is based on relaxing the transitivity axioms in the framework of the model presented in chapter 3. According to my justification, a lottery is justified in some (but not all) situations when an agent suffers from moral uncertainty. I will argue that in these situations, using a lottery is the best one can do, given one's moral uncertainty. I will characterise the set of situations in which a lottery is justified according to my account and present an explication for the term "the best one can do".

[72] See Broome (1990), (1991), (1994), for example. Other discussions of the questions include Hooker (2005), Sher (1980), Saunders (2009), Rescher (1969), Glover (1977).

However, unlike Broome, I will not argue that using a lottery, when it is the right thing to do according to my account, is also the fair thing to do. One could take a further step and try to argue that what makes a lottery right according to my account is also what makes it fair. Hence, one can argue that being fair is just doing the best one can to do the right thing, but one does not have to take this further step. I think there might be good reasons to take this further step[73], but I will not argue for it here. Here I only present a justification for the use of lotteries, not an account of fairness.

Is my account a rival to Broome's account? Not necessarily. One can hold the position that some lotteries are morally right for the reasons Broome presents and some are right for the reasons I present (and some may be right for other reasons). However, in section 4 I will compare the recommendations that my accounts gives in some cases to the recommendations that Broome's account gives in those cases and consider their relative strengths and weaknesses. I will argue that by accepting my account, one is able to avoid some of the problematic implications of Broome's account without having to give up on its positive ones.

First, however, I will present my account and contrast it with Broome's. This will be done in the following way: in Section 1, I will present Broome's account as well as some background issues that will be of later use. In Section 2, I will

[73] Hooker (2005) acknowledges (and refers to others who acknowledge) that "…fair is often used with a very broad meaning. A 'fair decision', in this very broad sense of 'fair', means a decision that appropriately accommodates all applicable moral distinctions and reasons". (Hooker 2005, p.331). This is in line with the "being fair as doing the best one can" thesis, only that under the explication presented here for "doing the best one can" such an understanding of fairness can also explain why lotteries are sometimes fair.

critically discuss an assumption that Broome implicitly uses in his account, namely, that it is possible to compare the strength of the moral claims of different people and will show how this relates to the idea of moral uncertainty. In Section 3, I will present my account for the rightness of lotteries.

The Fairness of Lotteries

Broome's starting point is the intuition that "Sometimes a lottery is the fairest way of distributing a good…" (Broome 1990, p.87). Broome also holds that because of this fact "…there will certainly be some circumstances where it is better to hold a lottery than to choose the best candidate deliberately" (Broome 1990, p.99).

This claim, by itself, poses a problem for Broome that he has to deal with even before presenting his justification for the intuition he started with: it seems that any moral preference ordering that ranks a lottery between two actions above both of these actions must violate the Sure Thing Principle (STP). The STP demands that when an agent is uncertain what the consequences of some of the actions available to him will be, then, when he evaluates these actions, he can disregard any state of the world in which all of them bring the same outcome. Consider, for example, the following table:

	ω_1	ω_2
l	A	B
a	A	A
b	B	B

Table 10

The SP demands that if the agent prefers act a to act b, then he should prefer act a to act l and act l to act b. Thus, it is easy to see that a lottery between two alternatives should never be preferred to both of them.

One way to deal with this problem is to reject the STP in moral contexts[74]. However, this is not the strategy Broome adopts and he, as well as others, has presented very convincing arguments against it[75]. Broome's suggestion for dealing with the problem is different. He suggests that in cases in which a lottery seems to be morally preferable to any of the alternatives over which it is defined, we have to include the fairness achieved by using the lottery in the description of the outcomes[76]. By following this suggestion, the STP is not violated because it does not apply. This is demonstrated in the following table:

[74] This is the position adopted, for example, by Diamond (1967), who first introduced this problem.
[75] Broome (1984), Section 2.
[76] See Karni (1996) for a similar suggestion.

	ω_1	ω_2
l	A achieved by a lottery	B achieved by a lottery
a	A achieved by a definite choice	A achieved by a definite choice
b	B achieved by a definite choice	B achieved by a definite choice

Table 11

Since, under the new interpretation of the situation, the two possible outcomes that act l might bring are different from the outcomes acts a and b bring, the STP does not apply to the decision-problem in question and so is not violated.

Notice that, by using this argument, Broome could have consistently argued that there are no cases in which the fair act is not the right act (a claim that he explicitly denies). Act l can be ranked at the top of an agent's moral preference ordering and no principle of rationality will be violated. However, there is a price for such a move. By including in the description of the outcomes some properties of the acts that (may) bring them about, one violates "the rectangular field assumption". As explained, this assumption requires that an agent's preference ordering be defined over the set of all possible acts that can be constructed by assigning any one of the possible outcomes to any one of the different states of the world. This assumption must be violated in our example if we follow the "redescribing the outcomes" method, as it is obvious that under

the new description there is no possible act that brings the outcome “A achieved by a lottery” in every state of the world.

Now, as Broome stresses, it is hard to treat this latter assumption as a genuine principle of rationality. Its role is rather to make the framework rich enough to allow for the representation theorem to hold. So Broome argues that he prefers to sacrifice this assumption in order to save the STP. I agree, but that does not change the fact that if we violate this assumption we cannot get (at least under Savage’s framework) a representation theorem, which means that if one does accept that sometimes a lottery between two acts should be preferred to both of them, one must hold that an agent with such preferences cannot be described as maximising the expectation of any quantity – call it goodness, positive moral value, moral utility or any other name you like – based on his preferences.

Although Broome does not explicitly claim this, it seems that this is his reason for allowing the right act to differ from the best act (i.e. the act that brings the most good), when fairness considerations are involved. Broome holds that “…goodness is actually fully reducible to betterness; there is nothing more to goodness than betterness”. (Broome 1999, p.164). If the right act is always the best act, which is sometimes a lottery, then in the framework of Savage which is the one Broome adopts, one cannot construct a goodness function out of the betterness ordering[77].

[77] While still being optimistic regarding future developments, Broome believes that “…none of the other frameworks suggested in the literature …quite solves the problem…” (Broome 1999, p.117). However, see Bradley (2007) for a representation theorem that does. In any case it is worth mentioning that the method of “redescribing the outcomes” has other unattractive features beside the violation of “the rectangular field assumption” to which it leads. For a discussion see Steele (2006).

I have lingered on the discussion above as it will serve me later in order to support the account of the fairness of lotteries that I will put forward. However, nothing in the considerations mentioned explains why lotteries are fair. The discussion only concerned the question of whether choosing a lottery over all available definite acts must violate some principle of rationality. As explained, Broome holds that this is not the case, but he also suggests an account of fairness that explains what can make a lottery fairer than definite choices. Here it is.

Broome takes fairness to be a proportional satisfaction of claims of different people. The satisfaction should be proportional to the strength of the claims in the sense that "…equal claims require equal satisfaction, that stronger claims require more satisfaction than weaker ones, and also – very importantly – that weaker claims require some satisfaction" (Broome 1999, p. 117). Claims (for some good), according to Broome, are reasons of a special kind to give the good to a specific person: they are reasons that constitutes"…*duties owed to the candidate herself…"* (Broome 1999, p.115; Broome's italics). I will not discuss this definition here, but rather will take it as given.

It is important, though, to see how, by using this definition of fairness, Broome is able to justify the use of lotteries. When one has to distribute some indivisible good among a group of people who have claims to this good and when there is not enough of the good to satisfy all claims, no possible distribution will be

completely fair. However, instead of choosing the distribution that brings the most good, but that might be extremely unfair, one can choose a fairer distribution by giving each one of the individuals some chance to get the good. This is, claims Broome, "…not perfect fairness, but it meets the requirement of fairness to some extent" (Broome 1999, p.119).

Two points must be emphasised with regards to this account. The first is that it will not always be justified to prefer the fairer option to the option that brings the most good. Sometimes the goodness consideration will override the fairness consideration and sometimes the opposite will hold. Thus, as was claimed before, sometimes the fair act is not the right act and sometimes the right act is not the act that brings the most good.

The second point is that in just the same way that claims should be satisfied in proportion to their strength when the goods are divisible, in the case of an indivisible good, the chance each person gets in the lottery should be proportional to the strength of his claim to the good. Thus, Broome's account allows for lotteries in non-trivial cases too, i.e. not only when one is indifferent with regard to who should get the good.

There are many unclear points in Broome's argument: what exactly makes a reason to give the good to a person into a claim by this person? Why is it that a chance for a good can substituted for the good itself? Why is it that claims require proportional satisfaction in the first place? However, I am not going to attack any of these points. Instead, I am going to focus my attention on an

implicit assumption that Broome makes, namely the assumption that we can, somehow, compare the strength of the claims of different individuals. This assumption is implicit in the requirement for a proportional satisfaction of claims, but Broome does not discuss it. I will do so in the next section.

Interpersonal comparisons of strength of claims and moral uncertainty

It will be useful to remind ourselves, firstly, of a different and more famous problem of interpersonal comparisons; interpersonal comparisons of utility. Here is the usual description of the problem: if an agent's preferences respect the axioms of Bayesian decision theory, then it is possible to represent his choices as a maximisation of expected utility for a unique probability function and a utility function which is unique only up to affine transformation (i.e. if, for some probability function, a utility function u represents the agent's preferences, then this is true for any utility function of the form v = au +b). This means that if we want to measure an agent's utility from different outcomes, using only (but all) information regarding his preferences over uncertain (as well as certain) prospects, we are not allowed to attach any significance to the zero point and to the unit size of the scale we will obtain (in the same way that we are not allowed to attach such a meaning to the zero point or to the size of units when we measure temperature).

It is easy to see that by choosing different "b"s in the formula "v=au+b", we change the zero point and that by choosing different "a"s, we change the unit's size. Since, given a set of rational preferences, we can use any "a" and any "b"

we like and represent these preferences as a maximisation of expected utility, we should not attach – based only on our information regarding the preference ordering - any significance to the “a” and to the “b” we choose.

A straightforward consequence of this observation is that using only the utility functions we construct on the basis of rational preferences over uncertain prospects, we cannot justifiably compare the utility levels of different agents on the same scale. Now, those who argue that interpersonal comparisons of utilities are meaningless base their argument exactly on this observation; since we have no method of making interpersonal comparisons of utilities, they are meaningless[78].

The validity of this argument should not concern us here. What is important for our purposes is the structure of it. This is so because it seems that, in the case of interpersonal comparison of claims, the structure is exactly the opposite. As I will demonstrate, it seems that in the case of claims it is tempting to argue that because sometimes we *do know* how to compare the claims of different people, in other cases *we know we cannot*.

Here is an example. Consider three candidates, A, B, and C, claiming one kidney. Assume that the moral evaluator is certain that A is better suited to get the kidney than B (e.g. he is younger, has greater chances for a successful operation, and is superior to B in every other respect that the evaluator takes to

[78] One obvious reply to this argument is that we might be able to use other kinds of information than the agents’ preferences in order to make these comparisons (e.g. psychological or biological information), and in fact this is the line taken by Harsanyi (for example see his 1955 paper). Originally, however, the argument was made by Robbins (1934), who held an extreme positivist approach that was immune to such a manoeuvre.

be relevant). The same holds for B when compared to C, and so also for A when compared to C. However, the only reason that C is in a need of the kidney is because he donated one of his own kidneys to A, a couple of months before the moral evaluator faced the decision.

If you like, you can also imagine that when C donated his kidney to A, they signed a contract that said that whenever there arises a case in which a kidney can be given to one of them, C should get it. Moreover, you can imagine that when the evaluator asks A what he thinks the decision should be, A admits that he believes that in a choice between him and C, C should get the kidney, but, in a choice between him and B, he should get the kidney, and in the same way B claims that A should get the kidney rather than him, but he should get the kidney rather than C, and C agrees that he should get the kidney rather than A, but B should get the kidney rather than him. What should the evaluator do?

Before I suggest an answer to this question, it is important to try and analyse it using Broome's framework. Intuitively, we know how to compare the strength of the claims of each of two individuals[79]. We know that A's claim to the kidney is stronger than B's, B's is stronger than C's, and C's is stronger than A's. However, *because* we know that, we also know that we cannot compare all these claims on a single scale. If we could, we would not form intransitive judgements regarding which candidate has a stronger claim for the kidney.

[79] It is worth mentioning that Broome himself takes kidney transplant cases as paradigmatic cases in which claims arise. See Broome (1990), p.99.

Now, Broome believes that "…fairness is concerned only with how well each person's claim is satisfied *compared with* how well other people's are satisfied. It is concerned only with relative satisfaction, not absolute satisfaction" (Broom 1999, p.117, Broome's italics). However, in order to make such comparisons, we must assume that we can measure the satisfaction of claims on the same scale, which, in turn, commits us to the possibility of comparing the strength of different people's claims on one scale. As Broome argues, "Take a case where all the candidates have claims of equal strength. Then fairness requires equality in satisfaction" (Broome 1999, p.117), but in the example above, it is tempting to claim that there is no sense in which we can compare the strength of the claims of the three candidates on the same scale and so, there is no sense in which we can argue that their claims are equal (or not).

Of course, the moral evaluator can always deny this and claim that there is such a way, but this will commit him to accepting that at least one of the three initial intuitive judgements he formed (which are shared also by the candidates themselves) is wrong – but which one?

It is important to stress that the point of this example is not to suggest that sometimes the relation "morally ought to be chosen rather than"[80] is intransitive. I believe it is transitive. The point, rather, is to suggest that the assumption that interpersonal comparisons of the strength of claims are always possible is

[80] I do not use the term "betterness relation" in order to be consistent with Broome's terminology. As explained earlier, Broome takes the betterness relation to represent only some moral considerations – not including fairness considerations – but he does believe (as I do) that both the betterness relation and one's overall moral judgements regarding what one ought to choose (what I called the "morally ought to be chosen rather than" relation) must be transitive (but not necessarily identical). I will claim that there is no need for two different relations and will show that, even with this assumption, sometimes the use of lotteries can be justified.

implausible, and hence, the justification for the fairness of lotteries that is based on it is incomplete. However, even if one is committed to the possibility of interpersonal comparisons of the strength of claims, nothing in what Broome says tells him how to make these comparisons. I think that the natural reaction to the dilemma in the example is to be uncertain regarding what is the morally right thing to do, and that this is true both if you deny the possibility of interpersonal comparisons of the strength of claims and if you accept it.

The idea of being uncertain regarding the question of what is the morally right thing to do is the key element that will help me to develop my alternative account of the rightness of lotteries. Before doing so, it is important to stress again what led us to this idea: this was either the impossibility of making interpersonal comparisons of strength of claims, or the difficulty of finding out how this should be done. More generally, as was discussed at length in Chapter 1, in many cases in which we have to compare the relative strength of different kinds of moral considerations, we are drawn to feeling uncertain regarding what is the morally right thing to do in a situation. This can be either because we are not sure of the appropriate relative weight we should give to each one of the considerations involved, or because we do not believe it is even possible to compare the relative importance of each, but still believe we must make a choice.

This last observation suggests that the cases in which Broome's account will justify a lottery are roughly the same as those in which my account will justify one. My account will justify a lottery only when one is uncertain regarding what

is the morally right thing to do and this happens, roughly, when one has to compare the relative importance of different moral considerations. Comparing the strength of the claims of different people usually falls under this characterisation.

This is not always the case. There are cases in which my account would recommend a lottery and Broome's would not, and vice versa. Sometimes both accounts will recommend a lottery, but different kinds of lottery. I will discuss some of these cases in section 4.

Moral uncertainty and lotteries

There are many different ways to interpret the result presented in the beginning of the previous chapter. Here I will concentrate on one of them. It can be argued that the result shows that when an agent is uncertain as to what is the morally right thing to do, and when he has no direct access to degrees of moral value, then, except in trivial cases, he must have intransitive moral preferences. What should the agent do in such cases?

Here is one possible answer: if we allow the agent to use mixed strategies, i.e. if we demand that the set of acts available to the agent is convex, that is it must include all the mixed strategies over this set, then there always exists an act that the agent believes is more likely or equally likely better than any other act available to him. It seems reasonable to demand that the agent choose such an

act[81]. To see this, let us start with the case of only 3 acts with regards to which the agent has intransitive preferences. We can do this by using the example from the previous chapter.

An agent has to choose between three acts that can bring about, in different states of the world, three possible outcomes: that all the 100 inhabitants of village A will die, that all 200 inhabitants of village B will die, or that all 400 inhabitants of village C will die. Assume the agent is absolutely confident that it is better to save more people than fewer people, thus, q(A**>**C) = q(A**>**B) = q(B**>**C)=1. However, the choice he has to make is not between sure outcomes, but between the following three acts:

	$p(\omega_1) = 4/9$	$p(\omega_2) = 3/9$	$p(\omega_3)=2/9$
a_i	B	B	B
a_j	A	C	C
a_k	B	A	C

Table 12

The agent is following the two conditions mentioned, i.e.

[81] This demand can be seen as a generalisation of the demand not to choose an act to which another act is preferred, which is usually used to justify the transitivity axiom (for example see Davidson, McKinsey and Suppes 1955). When it is impossible for the agent to have transitive preferences, and when there is no act that is preferred to all the acts over which the intransitivity occurs, this demand cannot be respected. However, I will now show that the generalization of this demand, i.e. the demand not to choose an act such that there is another act that one believes it is more likely than not better than the first act, can always be satisfied if the set of acts is convex.

1. EBC: for every two acts, a_i and a_j,

$$q(a_i > a_j) = \sum_{wk:ai(wk)\neq aj(wk)} p(w_k)q(a_i(w_k) > a_j(w_k)) / \sum_{wk:ai(wk)\neq aj(wk)} p(w_k) .$$

2. LBC: for every two acts a_i, a_j, $a_i \geq^* a_j$ iff $q(a_i > a_j) \geq q(a_j > a_i)$.

Now, since $p(\omega_2)+ p(\omega_3) > p(\omega_1)$, he believes a_i is better than a_j to degree 5/9. Since $p(\omega_1) > p(\omega_2)$, he believes that a_j is better than a_k to degree 4/7, but since $p(\omega_2) > p(\omega_3)$, he also believes that a_k is better than a_i to degree 3/5 and, thus, he has intransitive preferences.

Of course, realising this, the agent may choose to revise some of his degrees of belief so that his preferences will become transitive. If he succeeds in doing this, then this agent can be described as an expected moral value maximiser, i.e. by changing his degrees of belief in such a way that his preferences will become transitive and the two axioms will be satisfied, the agent implicitly assigns degrees of moral value to the three possible outcomes. However, the triviality result shows that, as he considers more and more possible outcomes (hypothetical or real), using such a strategy must lead him, at the limit, to be morally indifferent among all acts, except two. If we want to avoid this conclusion, we must accept that in some situations the agent does have intransitive moral preferences. So, for convenience, we can assume that the agent in our example has already made all the revisions of his degrees of beliefs that he is willing to make.

We are looking now for a mixed strategy, M, over the three acts such that the agent will believe that M is better than or equal to each of them. We can look at this in the following way: when the agent is using a mixed strategy, he adds some uncertainty to the uncertainty from which he already suffers: he transforms any world ω_i to which he gives a positive probability into *three* worlds, the probability of each one of these being the multiplication of the probability of the original world by the probability that the mixed strategy the agent uses gives to one of the original acts. Here is how this is done in our example:

	$p(\omega_1)^*$ $M(a_i)$	$p(\omega_1)^*$ $M(a_j)$	$p(\omega_1)^*$ $M(a_k)$	$p(\omega_2)^*$ $M(a_i)$	$p(\omega_2)^*$ $M(a_j)$	$p(\omega_2)^*$ $M(a_k)$	$p(\omega_3)^*$ $M(a_i)$	$p(\omega_3)^*$ $M(a_j)$	$p(\omega_3)^*$ $M(a_k)$
M	B	A	B	B	C	A	B	C	C
a_i	B	B	B	B	B	B	B	B	B
a_j	A	A	A	C	C	C	C	C	C
a_k	B	B	B	A	A	A	C	C	C

Table 13

Now, M is preferred or equal to a_i only when the agent believes it is more likely or equally likely that M is better than a_i, i.e, when the sum of the degrees of beliefs that the outcomes that M brings in every possible world in which M and a_i bring different outcomes, weighted by the probabilities of these worlds, is higher than this sum for a_i. Since we assumed that the agent's degrees of belief

regarding the betterness relations among pure outcomes are all equal to 1, this happens when:

$p(\omega_1)*M(a_j) + p(\omega_2)*M(a_k) \geq p(\omega_2)*M(a_j) + p(\omega_3)*M(a_j) + p(\omega_3)*M(a_k)$

We can do the same for M in relation to a_j and a_k, and we get three inequalities with three variables. Every inequality in this system can be derived from the other two, but we also know that $M(a_i) + M(a_j) + M(a_k) = 1$. It is easy to see that there is a unique solution to this system in which the equality relation holds for all inequalities. For the values in this example, this solution is $M(a_i) = M(a_j) = M(a_k) = 1/3$, and in the general case:

$M(a_i) = (2q(a_j > a_k) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1) + (2q(a_k > a_i) - 1))$

$M(a_j) = (2q(a_k > a_i) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1)) + (2q(a_k > a_i) - 1))$

$M(a_k) = (2q(a_i > a_j) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1) + (2q(a_k > a_i) - 1))$

These values also have an intuitive interpretation, which will be discussed in section 4.

The story, however, does not end here, as it is easy to see that for every mixed strategy, such as M, there exist two other acts such that the agent has intransitive preferences over M and these two acts. In our example, for instance, this can be done in the following way:

	$p(\omega_1)$* $M(a_i)$	$p(\omega_1)$* $M(a_j)$	$p(\omega_1)$* $M(a_k)$	$p(\omega_2)$* $M(a_i)$	$p(\omega_2)$* $M(a_j)$	$p(\omega_2)$* $M(a_k)$	$p(\omega_3$* $M(a_i)$	$p(\omega_3$* $M(a_j)$	$p(\omega_3)$* $M(a_k)$
M	B	A	B	B	C	A	B	C	C
N	A	A	B	C	C	A	C	C	C
L	B	A	B	A	C	A	C	C	C

Table 14

The reasons are identical to the reasons for the intransitivity in the original example.

However, notice that N and L are not mixed strategies over the three original acts. Given the set of the original acts and every mixed strategy over them, there is a unique mixed strategy that respects the condition that the agent should never choose a strategy when there exists another strategy available to him which he believes is more likely than not to be better than the strategy he actually chose. It seems, then, that in this kind of case, the only rational choice for the agent is this mixed strategy.

What happens, though, when the set of available strategies contains more acts? For example, what happens if this set contains the three acts from our example, acts N and L, and every mixed strategy over these 5 acts? Is it still true that there exists a unique mixed strategy, M, over this set, such that there

is no strategy in this set that the agent believes it is more likely than not that it is better than M?

The answer to the existence question is ‘yes’ (I will get back to the uniqueness question soon). To see that, we can think of the agent as playing a game against himself in which the payoffs for every combination of strategies are the agent's degrees of belief that one of these strategies is better than the other. The intuition is that when the agent has to make a choice, my demand from him is that, given what he chooses, there is no other strategy he could have chosen that he believes will be better. So we can think of it in the following way: the agent considers the strategies available to him and asks himself - for each one of them – “given that I choose this strategy, will there be a better strategy for me to choose?” If the answer is ‘yes’, he should not choose this strategy. It is easy to see that this condition holds for the two players in the game only when they play Nash equilibrium strategies.

Now, since the agent plays against himself, the game is symmetric: the strategies and the payoffs for each combination of strategies for the two players are identical. In the same way, since the two players represent the same agent, the equilibrium must be a symmetric one, since the agent can choose only one strategy.

So what we have is a 2-player symmetric game and every symmetric game has a symmetric Nash equilibrium (See Nash’s original [1951] paper).

To see things more clearly, let us construct such a game, using our original example. Each player has three pure strategies, a_i, a_j, and a_k. The payoff every player gets from choosing an act a, while the other agent chooses act b, is just his degree of belief that a is better than b. Since we assume that the agent ignores worlds in which the two acts give the same outcome, we can assign a payoff of ½ to every result in which the two players choose the same pure strategy. So here is the game:

	a_i	a_j	a_k
a_i	½ , ½	$q(a_i > a_j)$, $q(a_j > a_i)$	$q(a_i > a_k)$, $q(a_k > a_i)$
a_j	$q(a_j > a_i)$, $q(a_i > a_j)$	½ , ½	$q(a_j > a_k)$, $q(a_k > a_j)$
a_k	$q(a_k > a_i)$, $q(a_i > a_k)$	$q(a_k > a_j)$, $q(a_j > a_k)$	½ , ½

Table 15

Notice that if the agent has transitive preferences, i.e. if $q(a_i > a_j) \geq ½$, $q(a_j > a_k) \geq ½$ and $q(a_i > a_k) \geq ½$, the only Nash equilibrium is that both players play the pure strategy a_i . However, when the agent has intransitive preferences - which is the case we are interested in, i.e. when $q(a_i > a_j) \geq ½$, $q(a_j > a_k) \geq ½$ but $q(a_k > a_i) \geq ½$ - there is no pure strategies Nash equilibrium. However, there is a mixed strategies equilibrium and, in this case, it is unique.

Now, if it is the case that either 1) for any number of pure strategies in a symmetric 2-players game, the mixed strategy symmetric Nash equilibrium is unique, or 2) if it is not unique, there is always a strategy that belongs to this set

such that there is no other strategy that belongs to the set which is preferred to it, then we have a choice rule that respects the demand that the agent should never choose a strategy such that he believes there exists another strategy available to him which is better. Moreover, this choice rule sometimes recommends (i.e. whenever the agent has intransitive preferences) the use of a mixed strategy.

It turns out that although sometimes the mixed strategy symmetric Nash equilibrium is unique (for example whenever there are only three pure strategies available), in the general case, 1) is false. However, 2) is always true. Specifically, the agent must believe, regarding every two mixed strategies that are played in a symmetric equilibrium, that it is equally likely that either is better than the other, and so, must be indifferent between them. The reason is simple. Since any mixed strategy in an equilibrium is a best response to any other strategy, in particular a mixed strategy in a symmetric equilibrium is a best response to a mixed strategy in another symmetric equilibrium (to see that there might be more than one symmetric equilibrium, think of a cycle of 4 alternatives with equal strength of beliefs in each of the betterness relations: both a mixed strategy that gives to two pure strategies probability ½ and a mixed strategy that gives each pure strategy ¼ will satisfy the condition)[82].

[82] Note that this result does not depend on the two conditions presented. Many other decision theories that allow for intransitive preferences can serve. For example, if instead of using the degrees of belief in the betterness relations as the payoffs of the game, we use expected regret levels, the situation will be the same. More generally, Peter Fishburn (1984) has proved that whenever intransitive preferences can be represented by an SSB utility function, this will be the case.

To conclude, what we have shown is that if an agent respects the two conditions presented above then, although – except in trivial cases – he must have intransitive moral preferences, if the agent is allowed to use lotteries, there always exists a lottery regarding which he believes no other definite act or lottery is more likely than not better than it. Thus, for such an agent, it seems that the only rational choice will be to choose one of these lotteries.

Recall that the main challenge for any account – like Broome's - that tries to justify the use of lotteries on grounds of fairness is to deal with the apparent violation of the Sure Thing Principle. Recall also that the only way to reconcile the STP and the claim that it is sometimes morally better to choose a lottery over a definite act was to re-describe the outcomes in such a way that the fairness of the procedure would be incorporated into these. Choosing this strategy, however, makes it impossible – at least in Savage's framework - to describe an agent who prefers a lottery to a definite act as maximising the expectation of any quantity: goodness, moral utility, or what have you.

By following the account presented here, we can see that the agent ought to choose a lottery exactly in those cases when he cannot anyway maximise any quantity, i.e. when his preferences are intransitive. To be more precise, what I am arguing is that whenever an agent has transitive moral preferences, he should simply choose the best strategy available to him. However, when the agent suffers from some moral uncertainty, if he obeys the two conditions presented above, he must have intransitive preferences over some acts. This does not mean that he believes the moral betterness relation is intransitive. I

have assumed that the agent believes it is transitive. However, since all he can rely on are his beliefs about this relation, if he respects the two conditions, he has no way to avoid intransitivity. Thus, in the cases when the intransitivity arises, it seems that *the only rational thing for him to do is to choose a lottery.*

In sum, on my account, choosing a lottery is not only not an irrational thing to do, but rather – whenever it is justified to choose a lottery – the *only* rational thing to do. It is clear that on this account there is no need to claim that sometimes the right thing to do is not to choose the best act when possible: one ought always to choose the best act, but, when one is uncertain which act this is, the only rational thing to do is to use a lottery. Is it also the best thing to do? Well, yes and no. No, in the sense that by choosing a lottery the agent knows for sure that there is another act available to him that brings a higher amount of expected goodness (but he does not know which act this is). Yes, in the sense that, given his uncertainty, this is the only rational thing for him to do and since one ought to be rational in one's moral choices, then choosing the lottery is the only morally justified act[83].

It turns out that this account also has some nice predictions regarding the kinds of lotteries we ought to use. Some of these will be discussed in the next section.

[83] And note that by choosing a lottery the agent still respects the motivational demand since there is no act such that he believes it is more probable than not that it is morally superior to the lottery.

Which lotteries are justified?

In this section I will consider some of the predictions of my account regarding when, when not, and which lotteries are justified. This is not my main argument for my account. That has already been presented. However, I am aware of the fact that some of the steps I took in presenting my account could be rejected. Now, what I want to do is to give you a reason to think twice before doing that. The reason is that by accepting my account we gain an explanation for some judgements that, I think, are intuitive.

Of course, different people have different intuitions and this is particularly true with regards to the case of the moral value of lotteries where our intuitions may not be very strong. Thus, my discussion in this section will have to rely heavily on 'intuition pumps'. I will try to 'pump' to you the intuition that the recommendations of my account are correct in the cases I will present. I hope, however, that these will not be misleading intuition pumps, but rather constructive ones[84], i.e. they will push forward intuitions that we have an independent reason to accept and not ones which will lead us to more trouble.

It was argued in section 2 that the account presented here for the rightness of lotteries will give, in some cases, similar recommendations to those of Broome's account. The reason for that, it was argued, is that in my account the use of lotteries will be justified only when moral uncertainty arises and, roughly speaking, cases of moral uncertainty arise when the need to compare the

[84] See Dennett (1984) and (1994) for a discussion of the difference between the two.

relative strength of different moral considerations arises. Thus, since interpersonal comparisons of the strength of claims fit these criteria, both my account and Broome's may recommend a lottery in cases that fall into this category.

Now we have the tools to demonstrate this claim in a more precise way. Consider a case of three individuals, i, j, and k, all in need of a kidney. There is only one kidney available and the moral evaluator is uncertain regarding who should get the kidney. His degrees of beliefs are such, though, that he believes it is more likely than not that i should get the kidney rather than j, it is more likely than not that j should get the kidney rather than k and it is more likely than not that k should get the kidney rather than i. As was shown in the previous section, in such a case my account will recommend the following lottery among i, j and k:

$$M(a_i) = (2q(a_j > a_k) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1) + (2q(a_k > a_i) - 1))$$

$$M(a_j) = (2q(a_k > a_i) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1)) + (2q(a_k > a_i) - 1))$$

$$M(a_k) = (2q(a_i > a_j) - 1) / ((2q(a_j > a_k) - 1) + (2q(a_i > a_j) - 1) + (2q(a_k > a_i) - 1))$$

In other words, the weight individual i gets in the lottery, that is the chance that he will get the kidney, should be proportional to the moral evaluator's degree of belief that giving the kidney to j is better than giving it to k. This is simply a result of the model and the assumptions presented in the previous section. However,

here is one way to make this demand intuitive. The moral evaluator believes that if k does not get the kidney, i should get it (since he believes that giving the kidney to i is, more likely than not, better than giving it to j). The only reason the evaluator thinks i should not get the kidney is that he believes it is more likely than not that it is better to give it to k than to i. Thus, to the extent that the evaluator believes the kidney should not go to k, he should give it to i. The extent that the evaluator believes the kidney should not go to k is his degree of belief that it is better to give the kidney to j than to give it to k. Thus, it makes sense that the evaluator should give the kidney to i with a probability that is proportional to his degree of belief that k should not get it, i.e. his degree of belief that it is better to give the kidney to j than to k.

What would Broome's account recommend in this case? According to Broome's account, each person should get a chance which is proportional to the strength of his claim for the kidney[85]. What is the strength of the claim of each one of the individuals for the kidney? Nothing in Broome's account tells us how to calculate this, but, given the evaluator's beliefs regarding the betterness relation that holds between the three definite acts open to him, we can argue that the evaluator takes i to have a claim for the kidney only on the grounds that the evaluator himself believes that it is more likely than not that it is better to give the kidney to j rather than k. This is so, since i should get the kidney only to the extent that k should not get it (since the evaluator believes it is more likely than not that k should get it rather than i). Thus, the strength of i's claim for the

[85] And note that Broome takes kidney transplant cases to be the most likely candidates for the use of lotteries: "Consider, for instance, life-saving medical treatment such as kidney replacement. It seems plausible that, in these matters of life and death, fairness is particularly important. And it seems plausible that everyone has a claim to life, even if on other grounds some are much better candidates than others" (Broome 1990, p.99).

kidney – from the point of view of the evaluator - should be proportional to the evaluator's degree of belief that it is better to give the kidney to j rather than to k. The same argument holds, of course, for j and k.

Broome does not supply us with a clear criterion for when a reason to give an indivisible good to a person constitutes a claim by this person. However, by accepting my account, the following criterion (which Broome explicitly denies) arises. In the absence of any reason *not* to give the indivisible good to the person, any reason to give the good to the person constitutes a claim by this person. When the moral evaluator has transitive moral judgements, the only person who has a claim for the good is, thus, the one that the agent judges to be the best candidate. However, when the moral evaluator has intransitive judgements, each of the individuals has a claim to the kidney and so, according to Broome's own account; each one should get a chance to get the kidney which is proportional to the strength of his claim.

Notice that, if my account is adopted, the claims discourse is not needed. The only thing that we have to assume in order to support the lottery is that the moral evaluator is rational in his moral choices (that is, rational in the sense presented and defended in the previous section). However, the predictions of my account are that the lottery chosen will be the one that Broome's account (under a specific interpretation) recommends. This is, I think, evidence for my account.

Let us move now to examine some cases in which the two accounts give different recommendations. Let us start with the simplest case, which is also the one that is the most discussed in the literature. This is the case in which there is no moral uncertainty and the moral evaluator is morally indifferent between two possible acts[86].

For example, consider a kidney case in which there is one available kidney and two candidates, identical in every respect that the evaluator takes to be morally relevant. In this case, although my account allows the use of a lottery, it does not make it strictly morally superior to any one of the two definite acts (i.e. each one of the acts of giving the kidney to one of the candidates). Broome's account, on the other hand, does make the lottery that gives equal chances to the two candidates, morally superior to both any one of the two possible acts or any other lottery.

On the face of it, this looks like a weakness of my account, as intuitively the lottery that Broome's account recommends in this case *is* strictly morally superior to any other possible act. However, there is a price that Broome must pay here. If the fairness consideration adds some moral value to the lottery in case there is no moral uncertainty and the evaluator is morally indifferent between the two candidates, it should do so also in the case where there is no moral uncertainty, but the evaluator is not morally indifferent between the two candidates.

[86] See Diamond (1967) for example.

For example, consider yet another kidney case involving only two candidates, but this time the candidates are identical in everything, apart of the fact that one has a slightly higher chance of a successful operation. According to Broome's account, there must be some cases in which a lottery between the two definite acts will be morally superior to the act of giving the kidney to the candidate with the slightly higher chances of a successful operation.

In order to generate a lottery under Broome's account you can reduce the difference in the chances of a successful operation between the two candidates as much as you want. At some point – if Broome's account is not empty – you will reach a difference in chances such that choosing a lottery between the two candidates will become morally preferred to simply giving the kidney to the one with the (slightly) higher chances of success.

However, if you are consistent in your choices, you will always make the same choice. Thus, if you face a similar choice over and over again you will always prefer the lottery to the option of simply giving the kidney to the candidate with the slightly higher chances. But no matter how small the difference is between the two candidates' chances for a successful operation, after making this decision enough times this will result in preferring a policy that generates more loss of life to one that generates less. Now, ask yourself; do you still find it intuitive that there is any amount of fairness that can be gained by using a lottery which is sufficient to compensate for the loss of life resulting from choosing the policy that recommends using a lottery over the one that does not?

The trade-off has now become clear: if one is willing to accept that in the indifference case, a lottery is not strictly morally superior to the definite acts, one can deny that the fairness consideration is strong enough to lead to morally preferring a policy that generates more loss of life to one that generate less. If, on the other hand, one is willing to accept that sometimes a policy that generates more loss of life is morally superior to a policy that generate less, one can argue that in the indifference case, the lottery is strictly morally superior to any other act.

Another option is to retain both the judgement that in the indifference case the lottery is strictly morally superior to any other act and the judgement that when the evaluator is not morally indifferent and there is no moral uncertainty involved, a lottery is never justified. One can do this by limiting (in a somewhat artificial way, and contrary to what Broome argues) Broome's account to cases of indifference, or by making use of another account of fairness that kicks in only in cases of indifference. This move is unattractive for obvious reasons, but these reasons are theoretical, not ethical. Out of the three options mentioned, I find this last option the least unattractive.

Notice, however, that by accepting my account, one can accept that the fairness consideration kicks in only in cases of moral indifference, but still accept that in some cases in which the evaluator is not morally indifferent between the acts available to him, choosing a lottery is the morally right thing to do. In these cases, choosing a lottery is the morally right thing to do not because the lottery

is fairer than any other act, but rather because moral uncertainty is involved which allows for the application of my account.

I think that this feature of my account - that it can recommend a lottery even in non-trivial cases, but only when moral uncertainty is involved - is very attractive. Here is a case that can demonstrate this. Consider again a single kidney case, but this time there are ten people, i, j, k, and l1...l7, waiting for the kidney. Assume that the evaluator, after thinking about the decision for a while and gathering relevant information, summarises his judgements using the following table:

Age	Chances of success	Any other relevant consideration
i	K	j
j	I	k
k	J	i
l1…l7	l1…l7	l1…l7

Table 16

In other words, the evaluator believes that, from the point of view of the age of the candidates, i is more suited to get the kidney than j, j is more suited than k, and k is more suited than any of l1…l7. However, from the point of view of the chances for a successful operation, k is ranked above i who is ranked above j,

who is ranked above l1..l7. Finally, when the evaluator thinks of any other relevant moral consideration he ranks j above k, k above i and i above l1…l7.

What should the moral evaluator do? Well, one thing he can do is to try to give a relative weight to each one of the categories and, using these weights, to derive a combined ordering. If he manages to do this and get a transitive ordering, I believe he should simply give the kidney to the person ranked at the top, which will be, of course, either i, j, or k.

The problem, though, is that this kind of case is exactly the kind in which the agent might become uncertain regarding which act is the best choice and so it might happen that (using the assumptions of the previous section) he will find that he has intransitive preferences among i, j, and k. In such a case, my account will suggest a lottery, but this lottery will give a positive chance only to i, j and k and no chance at all to l1…l7. To see why this is the case, recall the analogy with a game that I used in the previous section to show why there always exists a lottery that is weakly preferred to any other act. It was demonstrated that, when the agent chooses such a lottery, his choice must constitute Nash equilibrium in the game he plays against himself.

Now, it is well known that a mixed strategies Nash equilibrium must give a positive chance only to rationalisable strategies, i.e. to strategies that can survive the process of iterated elimination of dominated strategies. It is clear that giving the kidney to each one of l1…l7 is not a rationalisable strategy because it is dominated by giving the kidney to either i, j, or k. Thus, according

to my account, if the agent should use a lottery (which might or might not be the case depending on the agent's beliefs) this lottery must give a positive chance only to i, j, and k.

What will be the recommendation of Broome's account? Well, first, it might be that the goodness considerations in this case will override the fairness considerations and thus, no lottery will be recommended. However, if this is not the case and some lottery will be recommended, then this lottery must give a positive chance to each one of the ten candidates since each one of them has, according to Broome, a claim for the kidney. This seems to me extremely unintuitive. Giving a positive chance to each one of the candidates reduces the chances of i, j, and k, and this is so even though the evaluator is sure that it would be wrong to give the kidney to anybody but i, j, or k among whom he is uncertain who should get the kidney.

Broome would argue that this might be justified because although, in terms of goodness, giving the kidney to one of l1…l7 would be a suboptimal choice, l1…l7 have claims to the kidney and these claims must get partial satisfaction. However, when I try to make intuitive sense of this claim, I find myself imagining a potential conversation between the moral evaluator and each one of the candidates. If the evaluator chose i, j or k, he would have a good response to any complaint made by l1…l7 for not choosing them, i.e. that he believes it is better to give the kidney to someone else. Thus, intuitively, l1…l7 do not have a justified claim for the kidney. This is, however, not the case regarding i, j and k, since, if the evaluator chooses i, for example, then k can justifiably complain

that he should have been chosen because the evaluator himself believes it is more likely than not that k is more suited than i. Thus, it sounds reasonable, on the face of it, that only i, j and k have a claim for the kidney.

The above was simply in order to show that it is unintuitive that l1…l7 have a claim for the kidney. However, I do not believe that even i, j, and k have a claim for the kidney, because for any complaint made by one of them, the evaluator has a good response; namely, that if he had given the kidney to the complainer, then somebody else would have a justified complaint. The right way to analyse the situation, I believe, is to give up the claims discourse altogether, and instead demand that the moral evaluator does what he believes to be the morally best act. If he does not believe for any act that it is more likely than not the best act, he must choose a lottery that he believes no other act (or lottery) is more likely than not better than it. As was explained, there always exists such a lottery.

Conclusion

I have presented an account of why choosing a lottery over a definite act is sometimes the right thing to do. According to this account, one ought always to choose the best act available when one can. When one cannot, one should use a lottery, and this is because using a lottery is the only rational thing to do in such a situation. So my account succeeds in satisfying both the demand that moral preferences be rational and the demand that one ought always to choose the best act available. Moreover, I have argued that the lotteries suggested by my account are the right ones. In some cases the same lotteries will be

recommended by Broome's account, but sometimes the recommendations will differ. In these cases, I have argued, the recommendations of my account are more intuitive than those of Broome's.

One can accept the account presented here for the rightness of lotteries and reject Broome's, but one can also accept both these accounts as different valid justifications for the use of lotteries. One can also take the account presented here not only as an account of the rightness of lotteries, but also as an account of the fairness of lotteries, but one does not have to do so. If one does, then one can think of being fair as doing the best one can. If one does not, than this is ok too, as long as one believes one *ought* to do the best one can.

Conclusion

This thesis is a documentation of an inquiry I have been pursuing. In this inquiry I have *done the best I could* to find an account of what it means to do the best one can and of what it means do try to do better, when it comes to making moral decisions. By “the best I could” I mean two things: 1. Whenever a decision had to be made regarding which one of several theoretical paths ought to be taken, I chose the one that seemed to me to have the highest plausibility to be the right one. 2. After making each one of these decisions, I did not just ignore all of the reasons (both those that are discussed in the literature and those that are not) which I had for *not making this decision*. Rather, I tried to accommodate them into the account I was building.

Although I did the best I could, I was only partly successful in my attempt; I think I was able to find a plausible account of what it means to do the best one can to do the morally right thing. However, I was not able to find such an account for what it means to try to do better.

My account of what it means to do the best one can to do the morally right thing can be summarised in the following algorithm. When one has to make a moral decision one should take the following steps.

1. One should first judge, regarding each pair of the acts available to one, which of the acts is morally superior to the other.
2. If one experiences no uncertainty regarding all of these judgements, one should check whether these judgements lead to a moral preference relation among the acts that obeys Savage's axioms. If one does experience uncertainty regarding some of the judgements, one should skip to step 4.
3. If they do, one should choose according to one's judgements. If they do not one *will*[87] become uncertain regarding some of one's judgements and then one ought to move to step 4.
4. If one is able to reduce the uncertainty one experiences to uncertainty regarding which one of several competing moral claims is correct, to assign degrees of moral value to each one of the acts available to one conditional on each one of the claims being true, and to compare these degrees across states, one ought to choose the act that maximises expected moral value.
5. If one is unable to do that, one should follow the rule *"do not choose an act such that you believe it is more likely than not that another act that is available to you is morally superior to it"*. If one can use lotteries, this rule is a decision rule, i.e. it points to a non-empty set of permissible acts in every choice situation. In most cases, my hypothesis is, obeying this rule will lead to a transitive preference relation. In such cases one should just choose the act that is ranked at the top of this preference relation. In

[87] This is not a normative requirement. It is a partly descriptive and partly conceptual claim that was defended in Chapters 1, 2 and 4.

some cases, obeying this rule will lead to an intransitive preference relation. In such cases one should move to step 6.

6. One should do one's best, given the time and information constraints one is subject to, to re-evaluate the moral evidence available to one and to change one's degrees of belief in the different judgements accordingly. At the end of this process, one should obey the rule. It might be that the re-evaluation leads to a transitive preference relation (in light of the rule) and then one should choose the act that is ranked at the top of this relation. It might be that the re-evaluation does not lead to a transitive preference relation (in light of the rule) and then obeying the rule is possible only if one chooses a lottery.

Following this algorithm seems to me a plausible account of what it means to do the best one can to do the morally right thing. However, doing the best one can to do the morally right thing is not the same as actually doing the right thing. This is why there is a need for another account: an account of how one should try to do better, i.e. an account of moral reasoning that aims at finding a complete moral theory that gives prescriptions for every possible moral choice.

Although I did the best I could to find such an account too, I was not able to find a satisfactory one. My inquiry has led me to a conclusion that I find hard to accept: a complete moral theory that is achieved by following the most plausible reasoning procedure I could think of must be either trivial, or one that violates at least some rationality principles, or one that gives recommendations that an ideal moral agent will not be willing to follow.

There is a “lottery paradox flavour” to this conclusion. Although I judge each of the assumptions I have used in my inquiry to be probably (as the term is used in a non-technical sense) justified, I judge what necessary follows from them to be probably unjustified. Thus, just like the moral agent who realises that his degrees of belief in comparative moral judgements constitute a lottery paradox, what I have to do now is to try to do better: I have to reason myself out of the paradox by thinking harder, re-evaluating my evidence and questioning my assumptions.

I plan to do that. For now, however, all I can offer to the reader are some thoughts about possible directions for this further investigation. They are rather speculative, though. Thus, the next few pages should be read with caution. What I am offering to the reader here are not definitive claims I can defend, but rather ideas for some general directions I think it might be worthwhile to try and develop in a more serious way.

Let us go back, then, to the discussion at the end of Chapter 4. There I claimed that an agent, whose moral intuitions are such that by choosing according to them, he will always obey the rationality axioms, might never be in need to form beliefs regarding which possible act is morally superior to which. However, I have also claimed, following Broome and others, that without having beliefs regarding which aspects of the world are morally significant, an agent cannot take the rationality axioms to be restrictive in any way.

I have shown that by allowing for uncertainty regarding which aspects of the world are morally significant, one commits oneself to allowing uncertainty regarding comparative moral judgements, which, in turn, exposes one to the triviality result. The only way to escape the result, therefore, is to deny uncertainty regarding which aspects of the world are morally significant. In order to do that, however, a reasoner must reason in a way that is insensitive to what *he takes to be legitimate reasons* to have judgements of some sort or another. This last consideration has led me to the conclusion that the real tension is the one between the rationality demand and the idea of moral *reasoning* (in the sense of using reasons in order to reach a conclusion) which aims at a complete moral theory.

Now, in Chapter 1, I quoted Singer in what seems to be a commitment to the view that whether we should reject or accept the method of reflective equilibrium should be determined by whether the conclusions which the method leads to are in line with Singer's view. Singer was assuming in that context that, by allowing for a very wide interpretation of the reflective equilibrium method, the moral views he endorses cannot be ruled out due to their unintuitiveness. We have seen that this is true, but the opposite claim also holds: by adopting a wide interpretation of the reflective equilibrium method, Singer's views cannot be justified either.

So maybe what Singer wants to claim, or in any case what can be claimed, is the following: there is indeed no place for reasoning in the moral context. Maybe, when it comes to moral questions, we should not aim to discover some

truth of the matter. Rather, what we really ought to do is to try to bring ourselves to accept the moral judgements we wish - and by that I mean either "believe we better" or "desire" - to act upon (regardless of the position we hold about whether moral claims have truth values or not).

There is nothing inconsistent, both for a cognitivist and for a non-cognitivist, in holding one moral judgement while desiring not to hold it, or believing it is better not to hold it. This is obvious, I think, in the case of Singer's example. Even those who hold a position, according to which we are under no obligation to help starving children around the world, can (and probably do) accept that it would be better if we would hold the opposite position. After all, holding the opposite position will, if we are ideal moral agents (and perhaps even if we are not), lead to our actually helping starving children around the world, and this is surely a good thing to do.

From this point of view, Singer's argumentation can be seen not as an attempt to justify his position, but rather as an attempt to move the audience to act in some way. The "child in the pond" story can be seen not as a device used to expose some inconsistency in our moral intuitions, but rather as an "intuition pump"[88], used to promote in us specific intuitions.

So here is one possible answer to the question as to what stating, preaching and arguing for moral theories that cannot motivate us is good for: it is good for changing our motivational states. Moral philosophy should not aim, that is, at

[88] See Dennett (1984) and (1995) for discussions of this concept.

discovery. It should aim at construction: the construction of us as better people. From this perspective, the problem with all of the attempts to block Singer's inference from the "child in the pond" story to the conclusion that we ought to dedicate most of our time, and donate most of our money in order to save dying children around the world is not that they are wrong but rather that they promote the wrong intuitions. Why spend so much energy and time arguing for a conclusion we wish people will not accept?

I do not think this argument actually solves the problem that the result poses to the possibility of moral reasoning, but it does show that even if no good solution can be found, moral debate can still be valuable. In any case, it is important to stress again, the kind of moral reasoning described here is special. It is moral reasoning that aims at a complete moral theory, i.e. a theory that gives a prescription for every possible choice. Some philosophers have already questioned, for different reasons from mine, the possibility of arriving at such a theory using reasoning.

One of these philosophers was G.E. Moore, who wrote "Ethics, therefore, is quite unable to give us a list of duties: but there still remains a humbler task which may be possible for Practical Ethics. Although we cannot hope to discover which in a given situation, is the best of all possible alternative actions, there may be some possibility of shewing which among the alternatives, *likely to occur to any one,* will produce the greatest sum of good. The second task is

certainly all that Ethics can ever have accomplished…" (Moore 1965, p.149, Moore's italics)[89].

Moore, so it seems, was not very troubled by this conclusion. I am, but I also think that Moore's observation (which is in line with the algorithm presented above) that the conclusion does not pose a problem for moral reasoning in the context of practical ethics is very encouraging.

One thing that I find especially encouraging about it is the implications it has for the question as to how the process of making public choices should be made. According to the "division of labour" picture I have presented in the introduction, we first have to decide which ends we should care about from a moral point of view (and for that we have to consult ethicists) and then, having reached a conclusion regarding this matter, we have to decide what is the best way to promote these ends (and for that we have to consult economists or other professionals). If one accept Moore's conclusion then, in light of my discussion so far and contrary to the use that is sometimes made of Moore treatment of moral intuitions, one must take this picture to be flawed.

If rational moral reasoning can take place only in the context of particular decisions, then ethicists cannot establish the task assigned to them by this picture, in any case. Moreover, ethicists and economists that do not avoid being engaged in moral reasoning, can and should play a role when it comes to giving

[89] Surely, Moore's conceptual framework is very different from the one developed here, but the point expressed in the quote is at least very close to my conclusion: we cannot hope to arrive through moral reasoning at a moral theory that gives a prescription for every possible moral choice.

recommendations for acting in the context of specific decisions. In order to do this, ethicists must engage themselves with the non-moral aspects of the decisions under considerations. In the same way, economists and other professionals, must engage with the moral aspects of these decisions and avoid leaving this job to the ethicists alone.

I have discussed this matter in the Introduction in the context of the climate change debate. Now, I hope, I have supplied, as a by-product of my inquiry, a more explicit argument against the “division of labour” picture.

This is nice, I think, but the truth is that it does not really make me feel better about the result. If you feel like me, then although I do not have any further argument to offer you, you might find some comfort in realising that others have reached similar conclusions before, and still were able to find some joy in their life.

One of them was David Hume, who wrote in the conclusion to the first book of his Treatise of Human Nature:

“But what have I here said, that reflections very refined and metaphysical have little or no influence upon us? This opinion I can scarce forbear retracting, and condemning from my present feeling and experience. The intense view of these manifold contradictions and imperfections in human reason has so wrought upon me, and heated my brain, that I am ready to reject all belief and reasoning, and can look upon no opinion even as more probable or likely than

another. Where am I, or what? From what causes do I derive my existence, and to what condition shall I return? Whose favour shall I court, and whose anger must I dread? What beings surround me? and on whom have, I any influence, or who have any influence on me? I am confounded with all these questions, and begin to fancy myself in the most deplorable condition imaginable, invironed with the deepest darkness, and utterly deprived of the use of every member and faculty.

Most fortunately it happens, that since reason is incapable of dispelling these clouds, nature herself suffices to that purpose, and cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of backgammon, I converse, and am merry with my friends; and when after three or four hours' amusement, I would return to these speculations, they appear so cold, and strained, and ridiculous, that I cannot find in my heart to enter into them any farther".

I am off to play backgammon, then.

References

Anand, P. Pattanaik, K. P. & Puppe, C. (2009), *The Handbook of Rational and Social Choice,* Oxford University Press.

Baron, J. (1995). A psychological view of moral intuition. *Harvard Review of Philosophy, 5*, 36-40.

Baron, J, (2000), Can we use Human judgements to determine the discount rate?, *Risk Analysis,* 20, 861-68.

Binmore, K. (2009), *Rational Decisions,* Princeton University Press.

Binmore, K. & Voorhoeve, A. (2008), *Similarity-Based Decision-Making in Moral Decisions: An Experiment,* Unpublished manuscript.

Bradley, R. (1999), More Triviality, *Journal of Philosophical Logic,* 26.

Bradley, R. (2000), A Preservation Condition for Conditionals, *Analysis,* 60.

Bradley, R. (2007), A Unified Bayesian Decision Theory, *Theory and Decision,* 63:3. pp.233-263.

Bradley, R. List, C. (2009), Desire as Belief Revisited, *Analysis,* 69(1), 31-37.

Brink, O. D. (1997), Moral Motivation, *Ethics,* 108, 1, 4-32.

Broome, J. (1984), Uncertainty and Fairness, *Economic Journal, 94*, pp. 624-632.

Broome, J. (1990), Fairness, *Proceedings of the Aristotelian Society,* 91, pp. 87-102.

Broome, J. (1991), *Weighing Goods,* Oxford: Blackwell Publisher. (1)

Broome, J. (1991), Desire, Belief and Expectation, *mind,* 398, 265-257. (2)

Broome, J. (1994), Fairness versus Doing the Most Good, *Hasting Center Report 24*, pp.36-39*.*

Broome, J. (1999), *Ethics out of Economics,* Cambridge University Press.

Broome, J. (2006), Reasoning with Preferences?, in Olsaretti, S. (ed.), *Preferences and Well-being,* Cambridge University Press.

Broome, J. (2008), The ethics of climate change, *Scientific American.*

Byrne, A. Hajek, A. (1997), David Hume, David Lewis and Decision Theory, *Mind,* 106, 423.

Bucciarelli, M. Johnson-Laird, P. N. (2008), The Psychology of Moral Reasoning, *Judgment and Decision Making,* 3, 2, 121-139.

Colander, D. (2007), Retrospectives: Edgeworth's hedonimeter and the quest to measure utility", *Journal of Economic Literature*, 43, 9-64.

Costa, H. A. Collins, J. Levi, I. (1995), Desire as Belief Implies Opinionation or Indifference, *Analysis,* 55(1), 2-5.

Damasio, A. (1994), *Descartes' Error: Emotion, Reason, and The Human Brain,* Putnam, New York.

Daniels N. (1979), Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy*, LXXVI, 5, 256-82.

Daniels N. (1996), *Justice and Justification: Reflective Equilibrium in Theory and Practice*. New York: Cambridge U Press.

Dasgupta, P. (2007), Commentary: The Stern Review's Economics of Climate Change", *National Institute Economic Review, 199, 4-7.*

Davidson, D. McKinsey, J.C.C. Suppes, P. (1955), Outlines of A Formal Theory of Value I, *Philosophy of Science,* 22,2, 140-60.

Daskal, S. (2010), Absolute value as Belief, *Philosophical Studies,* 148, 2.

Dennett, D. (1995). Intuition pumps, *The third culture*, New York: Simon & Schuster.

Dennett, D. (1984). Elbow room: The varieties of free will worth wanting. Cambridge, MA: Bradford Books/MIT Press and Oxford University Press.

Diamond, A. P. (1967), Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment, *The Journal of Political Economics,* 75:5, pp. 765-766.

Douven, I. & Williamson, T. (2006), "Generalizing the Lottery Paradox", *The British Journal for the Philosophy of Science*, 57, 4, pp. 755-79.

Fishburn, C.P. (1984), SSB Utility Theory: An Economic Perspective, *Mathematical Social Science,* 8:1, pp.63-94.

Frankena, W. (1973), *Ethics*, Prentice-Hall, Inc.

Gibbard, A. (1981), Two Recent Theories of Conditionals, W.L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs: Conditionals, Beliefs, Decision, Chance, and Time,* Dordrecht, Holland: D. Reidel, 1981, 211–47.

Gibbard, A (1990). *Wise Choices, Apt Feelings*, Clarendon Press, Oxford.

Gibbard, A. (2008), *Reconciling Our Aims,* Oxford University Press.

Glover, J. (1977), *Causing Death and Saving Lives,* Penguin.

Goodman, N. (1965). Fact, Fiction and Forecast, Bobbs-Merrill, Indianapolis.

Greene, J. D. (2002). The Terrible, Horrible, No Good, Very Bad Truth About Morality and What To Do About It. Department of Philosophy, Princeton University. (advised by David Lewis and Gilbert Harman).

Greene, J. D. (2007). The secret joke of Kant's soul, in *Moral Psychology, Vol. 3: The Neuroscience of Morality*: *Emotion, Disease, and Development*, W. Sinnott-Armstrong, Ed., MIT Press, Cambridge, MA.

Hahn, U. Frost, J. M. Maio, G. (2005), What's in a heuristic?, *Behavioral and Brain Science,,* 28:4, 551-2.

Haidt, J. (2000), The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment, *Psychological Review,* 108, 814-834.

Hajek, A. Pettit, P. (2004), Desire Beyond Belief, *Lewisian Themes,* Oxford University press.

Harsanyi, C. J. (1955), Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility, *The Journal of Political Economics,* 63:4.

Harsanyi, C. J. (1978), Bayesian Decision Theory and Utilitarian Ethic, *The American Economic Review*, 68, 2, 223-228.

Harsanyi, C. J. (1985), Acceptance of empirical statements: A Bayesian theory without cognitive utilities, *Theory and Decision,* 18,1, 1-30.

Hanna, R (2004). Kant's Theory of Judgment, *Stanford Encyclopedia of Philosophy.*

Hooker, B. (2005), Fairness, *Ethical Theory and Moral Practice,* 8, pp. 329-352.

Hume, D. (2006), *A Treatise of Human Nature,* BiblioBazaar.

Jackson, F. and Smith, M. (2006), Absolutist Moral Theories and Uncertainty, *Journal of Philosophy,* 103, 267-283.

Jeffrey, C.R. (1956), Valuation and Acceptance of Scientific Hypotheses – Discussion, *Philosophy of Science*, (23(3):237-46.

Jeffrey, C.R. (1965), *The Logic of Decision*, The University of Chicago Press.

Jeffrey, C.R. (1992), *Probability and the Art of Judgement*, Cambridge Studies in Probability, Induction, and Decision Theory, Cambridge & New York: Cambridge University Press.

Joyce, M. J. (1998), A Nonpragmatic Vindication of Probabilism, *Philosophy of Science,* 65, 4, 575-603.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 49–81), New York: Cambridge University Press.

Kahneman, D. (2003), Maps of Bounded Rationality: Psychology for Behavioral Economics, 93, 5, 1449-1475.

Kamm, F.M. (1999), Does Distance Matter Morally to the Duty to Rescue? *Law and Philosophy,* 19, 6, pp. 655-681.

Karni, E. (1996), Social Welfare Functions and Fairness, *Social Choice and Welfare,* 13:4, pp.487-496.

Kolodny, N. (2005), Why be rational? *Mind,* 114, 509-63.

Kyburg, H.E. (1961) *Probability and the Logic of Rational Belief*, Middletown, CT: Wesleyan University Press.

Lance, N. M. (1995), Subjective probability and acceptance, *Philosophical Studies,* 77, 1, 147-79.

Lewis, D. (1976), Probabilities of Conditionals and Conditional Probabilities, *The Philosophical Review,* LXXXV, 3.

Lewis, D. (1980), A Subjectivist's guide to Objective Chance, in Jeffery, R. C. Ed. *Studies in Inductive Logic and Probabilities,* Vol. II, Berkeley: University of California Press.

Lewis, D(1988). Desire as Belief, *Mind,* 97, 323-32.

Lewis, D (1996). Desire as Belief II, *Mind, New Series,* 105, 418, 303-13.

Lockhart, T. (2000), *Moral Uncertainty and its Consequences,* Oxford University Press.

Loomes, G., Orr, S.W., & Sugden, R. (2009), Taste uncertainty and status quo effects in consumer choice, *Journal of Risk and Uncertainty*, 39, 113-35.

Maher, P. (1993), *Betting on Theories*, Cambridge: Cambridge University Press.

Mele, A. (1996), Internalist Moral Cognitivism and Listlessness, *Ethic,* 106, 727-53.

Mikhail, J. (2007), Universal moral grammar: theory, evidence and the future, *Trends in Cognitive Science,* 114, 143-152.

Monin, B. Pizzaro, A. D, Beer, S. J (2007). Reason and Emotion in Moral Judgment: Different Prototypes Leas to Different Theories, *Do Emotions Help or Hurt Decision Making,* Russell Sage Foundation, New York. (a)

Monin, B. Pizzaro, A. D, Beer, S. J (2007). Deciding vs. Reacting: Conceptions of Moral Judgment and the Reason-Affect Debate, *Review of General Psychology,* 11(2), 99-111. (b)

Moore, G. E. (1903), *Principia Ethica,* Cambridge University Press.

Nash, J. (1951), Non-Cooperative Games, *Annals of Mathematics,* 54:2, pp.286-295.

Nussbaum, M (2001). *Upheavals of Thought: The Intelligence of Emotions*, Cambridge University Press.

Oddie, G. (1994), Harmony, Purity, Truth", *Mind*, 103: 452-72.

Piller, C (2000), Doing What is Best, *The Philosophical Quarterly,* 50, 199.

Price, H. (1989), Defending Desire as Belief, *Mind,* XCVIII, 389, 119-127.

Rawls, J. (1971), A Theory of Justice, Harvard University Press, Cambridge, Massachusetts.

Rawls, J. (1974). The Independence of Moral Theory, Proceedings and Addresses of the American Philosophical Association 48, 4-22.

Rescher, N. (1969), The Allocation of Exotic Life Saving Therapy, *Ethics,* 79. pp.173-186.

Robbins, L. (1932), *An Essay on the Nature and Significance of Economic Science*, London: Macmillian.

Roberts, C. R (1999). Emotions as Judgments, *Philosophy and Phenomenological Research,* LIX, 3, 793-798

Rosati, S. C. (2006), Moral motivation, *Stanford Encyclopaedia of Philosophy.*

Saunders, B. (2009), A Defence of Weighted Lotteries in Life Saving Cases, *Ethic Theory and Moral Practice,* 12, pp. 279-290.

Savage L, J. (1972), *The Foundations of Statistic*, Dover Publications.

Scanlon, T.M. (2003), "Rawls on Justification", in Samuel Freeman, ed., *The Cambridge Companion to Rawls*, Cambridge: Cambridge University Press, 20, 149-60.

Sepielli, A. (2009), What to do when you do not know what to do?, *Oxford studies in Metaethics,* 4, Oxford University Press.

Shafer-Landau, R. (1998), Moral Motivation and Moral Judgment, *Philosophical Quarterly,* 48, 353–8

Shafer-Landau, R. (2000), A Defence of Motivational Externalism, *Philosophical Studies,* 97, 267–91

Sher, G. (1980), What makes a Lottery Fair?, *Nous,* 14:2, pp. 203-216.

Singer, P. (1972), Famine, Affluence, and Morality, *Philosophy and Public Affairs,* 1:3, 229-43.

Singer, P. (2005), Intuitions, heuristics, and utilitarianism, *Behavioral and Brain Sciences*, 28, pp 560-1.

Slovic, P. Finucane, M. Peters, E. MacGregor, G. D (2007), The Affect Heuristic, *European Journal of Operational Research,* 177, 3, 1333-1352.

Slovic, P. Lichtenstein, S. (1983), Preferences Reversal: A Broader Perspective, *American Economic Review,* 73, 596-605.

Smith, M. H. (1988), Making Moral Decisions, *Nous,* 22, 1, 89-108.

Smith, M. (1987), The Humean Theory of Motivation", *Mind*, 96,36-61.

Solomon, R (1976). *The Passions: Emotions and the Meaning of Life*, Garden City, N.Y.

Smith, M. (2002). Evaluation, Uncertainty and Motivation, *Ethical Theory and Moral Practice,* 5, 305-320.

Steele, K. (2006), What Can We Rationally Value? A Lecture given at the *Third Annual Austin-Berkeley-CMU Formal Epistemology Workshop.*

Stern, N. (2007), *The Economic of Climate Change: the Stern Review,* Cambridge University Press.

Stich, S. (1988). Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity, *Synthesise*, 74, 391-413.

Sunstein, R. C. (2005), Moral Heuristics, *Behavioural and Brain Sciences,* 28, 531- 542.

Timmons, Mark. (2007), Toward a sentimentalist deontology.In Moral psychology, *volume 3, the neuroscience of morality: emotion, brain disorders*, *and development*, ed. Walter Sinnott-Armstrong, 93–104. Cambridge: MIT.

Tversky, A. (1969), Intransitivity of Preferences, *Psychological Review*, 84, 327–52.

Van Roojen, M. (2008), Moral Cognitivism vs. Non-Cognitivism, *The Stanford Encyclopedia of Philosophy.*

Voorhoeve, A. (2008), Heuristics and Biases in a purported counter-example to the acyclicity of “better then”, *Politics, Philosophy and Economics,* 7, 3, 285-299.

Voorhoeve, A. (2009), *Conversations on Ethics,* Oxford University Press.

Walzer, M. (1983), *Spheres of Justice,* Basic Books.

Weintraub, R. (2007), Desire as Belief, Lewis notwithstanding, *Analysis,* 67, 2.

Weinberg, J. M. Nichols, S. Stich, S. (2001), Normativity and Epistemic Intuitions, *Philosophical Topics*, 29, 1&2, 429-460.

Weitzman, L. M. (2007), A Review of The Stern Review on the Economics of Climate Change, *Journal of Economic Literature,* XLV, 703-24.