Can an irrational agent reason himself to rationality? A Triviality result

Abstract

When an agent that accepts transitivity of preferences as a principle of rationality finds himself expressing intransitive preferences, he has to change some of his expressed preferences so that transitivity will be restored. When such an agent also believes in the existence of some independent betterness relation among the alternatives over which he forms his preferences, it is reasonable to demand that the way he changes his intransitive expressed preferences will be sensitive to his beliefs regarding this betterness relation. It is shown that under two natural conditions for such sensitivity, in case there are infinitely many alternatives, the agent must end up being indifferent between all alternatives except two. Some implications of this result for ethics are discussed.

Introduction

Most of us are prone – under certain conditions – to express, either by actual choice or by reflection, intransitive preferences. At the same time, most of us do accept transitivity of preferences as a condition of rationality. Thus, when we realize 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 transitive, complete and reflexive preferences 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.

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 make sure his preference follow the direction of the betterness relation. In section 2 of this paper I will introduce two such natural conditions and will show that by accepting them one commits oneself to a very disturbing result.

One context in which people usually accept that there exists an objective (that is "objective" in the sense I have introduced) betterness relation among different options is the moral context, and thus the discussion here will be made in this context. However, I believe most - if not all - that will be argued is true also regarding any other context of this kind[1].

The rest of the paper will be organized in the following way. In section 1 I will discuss some background issues and make some preliminary distinctions that

[1] An example for another context of this kind is when one is acting as an agent of another person and tires to follow, as much as possible, the other person's preferences. In such a context the other person's preferences constitute the betterness relation.

will prepare the ground for the result introduced in section 2. In section 3 I will discuss the significance of the result and in section 4 I will discuss some possible routes one might take in order to avoid it and point to one direction that seems most promising to me (without presenting a complete argument).

Betterness and degrees of goodness

There are two kinds of moral judgements, judgements concerning how good or bad a given act or outcome is on its own (let us call them type 1 judgements) and judgements concerning which one of two or more acts or outcomes is morally better than the other(s) (let us call them type 2 judgements).

Intuitively both types of judgements admit of degrees: we can feel more or less strongly about them, we can be more or less confident in making them. It seems to me that these degrees of judgements are best characterized as the degrees of beliefs we have in the propositions that are the objects of the judgements and I will shortly presents some arguments for that. However, I do not take this characterization to be above the need for a justification so in the meantime I need to use some neutral term to describe these degrees and so I will use the term “firmness”. This term, however, is sometimes used in the literature to refer to the degree of evidential support a certain proposition has. It should be clear, thus, that whenever I use this term in this paper I do not

intend the latter. I am only using it in a tentative way to refer to the degrees of judgements that get – under different approaches – different interpretations[2].

These degrees of firmness of moral judgements are different, however, from another kind of degrees we often talk about when we talk about moral judgements, i.e. the degrees of goodness (or badness) and the degrees of betterness of the acts or outcomes themselves. To say that it is worse to kill another person for no reason than to steal from another person for no reason is not the same as to say that the judgement that it is bad to kill another person for no reason is more firm than the judgement that it is bad to steal from another person for no reason. Intuitively, both judgements have the same level of firmness, although the two acts have different levels of badness.

In the same way, to say that of three acts A, B and C; A is better than B and B is better than C but A is better than B more than B is better than C is different from saying that the judgement that A is better than B is firmer than the judgement that B is better than C. Both judgements can be firm to the same degree and still one can feel that A is better than B more than B is better than C.

There is a very natural way in which these two types of moral judgements and the different kinds of degrees associated with them seem to constrain each other, i.e. act (or outcome) A is better than act (or outcome) B iff the level of

[2] Michael Smith (Smith 2002) use the term "Certitude" for the same purpose. However, in the psychological literature, when coming to describe the strength of intuitive moral judgements, the term "Firmness" is used (see Sunstein 2005 and the various replies, for example). I choose the latter because I want to use the adjective form of the concept and since the adjective form of "Certitude" is the same as of "Certainty", choosing "Certitude" seems to miss the whole point of using a neutral term.

goodness of A is higher than the level of goodness of B and in the same way A is better than B to a greater extent than B is better than C iff the difference between the levels of goodness of A and B is greater than the difference between the levels of goodness of B and C.

If we move from moral judgements to moral decisions, a plausible requirement of a rational moral agent[3] who can choose one act out of a set of possible acts available to him is that he will choose (one of the) best acts, i.e. an act which has no other act that is better than it. If the moral betterness relation is transitive there is always such an act and I will assume here that it is[4].

What should an agent do, though, when he realizes that his type 2 moral judgements are intransitive? Then, if he is committed to transitivity, he must change at least some of his judgements so that they will become transitive. How should he do that? It will be helpful to distinguish between two general methods he can use. The first one is to find out, somehow, what the levels of goodness of each one of the acts available to him are and then change his judgements so that the constraint I have introduced will be respected. Another option is to change his judgements directly, somehow.

Let us begin with the first method. On what basis can an agent attach different levels of goodness to different acts? Well, if he has absolute confidence in

[3] And one can argue that being rational in one's moral decisions is a moral requirement. See for example Harsanyi: "…an individual making a moral value judgment must follow, if possible, even higher standards of rationality than an individual merely pursuing his personal interests" (Harsanyi 1978, p. 226).

[4] See Broome (1999), Chapter 10, for an explicit justification of this assumption.

some moral theory that assigns such levels he can just use it, but in such a case it is unlikely that he will find himself expressing intransitive type 2 moral judgements in the first place. If he does realize that his judgements are intransitive, it must be, then, because he has no direct access to the levels of goodness of each one of the acts. This can happen either because he is unsure which moral theory is the right one or because he is unsure what levels of goodness a moral theory that he believes in, assigns to each act. Another option is that he just does not base his moral judgements on his beliefs regarding different moral theories (he may be a moral pluralist or he may just not be very impressed by theories in general)[5].

In such cases the agent needs another method to find out the levels of goodness of each act. David Lewis (1988) presented – only in order to reject – a thesis he called "The Desire-as-Belief" Thesis (DBT), a version of which can be taken to be such a method[6]. According to this thesis an agent's desire for a proposition, A, is (or at least should be) 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. Lewis showed that this thesis is consistent with another requirement that he found reasonable, the

[5] These issues are discussed in the literature under the title of "moral uncertainty". See for example Lockhart 2000. The underlying picture used in these discussions is that the source of any moral uncertainty is lack of confidence regarding which moral theory is the right one or what are the implications of a given moral theory to specific cases. The discussion that follows in this section and in the next one, explains, I think, what is missing from this picture. Basically, it is that it ignores the fact that we usually come to believe different moral theories on the basis of their recommendations for specific cases, i.e. we use some version of the reflective equilibrium method, and so there is a need to discuss a more fundamental kind of moral uncertainty: uncertainty regarding which ones of several acts are the best (from a moral point of view) acts to perform that cannot be reduced to uncertainty regarding some more general moral claims.

[6] Although, Lewis himself did not suggest it should be used in such a way.

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.

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.

When applied to the moral domain, notice that in Lewis's formulation the moral beliefs at issue are of the kind I called type 1 moral judgements and thus his triviality result can be interpreted as showing that either type 1 moral judgements are not beliefs or that if they are beliefs, one cannot be uncertain about them. Lewis's own diagnosis was that the problem with the DBT is the first one, i.e. that it is a non-Humean thesis. According to the Humean position, argued Lewis, desires and beliefs are two different mental attitudes that cannot be reduced to each other and the problem with the DBT is that it is trying to do exactly that.

Taking Lewis's diagnosis seriously in the moral domain means adopting a picture in which type 1 moral judgements are not beliefs but rather desires and thus degrees of type 1 moral judgements are not degrees of beliefs but rather degrees of desires.

Notice, however, that what the DBT actually says is that the degree of goodness a morally motivated agent attaches to a proposition is equal to the degree of firmness of the judgement that this proposition is good. This claim should, I think, be rejected by any plausible non-Humean too, as, as was argued, intuitively, one can have two judgements that are equally firm that two propositions are good while still attaching different levels of goodness to the two propositions. I think it will be uncontroversial to argue that it is good both to donate £1000 to charity and £2000 to charity (and that these two judgements are equally firm) but that it is better (ceteris paribus) to donate £2000 than to donate £1000.

According to the DBT it seems that a morally motivated agent that has to choose between donating £1000 and donating £2000 should not prefer the latter as his degrees of confidence in the judgements that these two acts are good are equal. In other words it seems that it is a consequence of the DBT that type 1 moral judgements constrain type 2 moral judgements in the wrong way and thus it should be rejected regardless of Lewis's triviality result.

This leads one to suspect Lewis's diagnosis and indeed when we think of an agent with intransitive moral judgements it seems that the Humean position is not in a better condition than the non-Humean one (in fact, I will shortly argue, it is even in a worse position). This is so since if the agent has intransitive moral judgements there is no desirability (or goodness) function that can represent (or that is consistent with) these judgements.

It seems, thus, that the first method of changing one's intransitive judgements faces some hard problems and thus it might be useful to move to the second one, i.e. to changing one's type 2 judgements directly. This move should not be surprising from a decision theoretic point of view, since in decision theory utilities (or goodness levels, in the moral context) are only representations of preferences. Even those who take utility to be a real mental quantity of some sort, agree that it must be consistent with the agent's preferences among lotteries (that is if he is rational), thus the utility an agent attaches to different propositions can be constructed from his preferences among lotteries involving these propositions. Now, since the utility values a morally motivated agent attaches to different morally significant propositions are just the degrees of goodness he attaches to these propositions (as he his morally motivated), these degrees must be at least consistent with (if not constructed from) his type 2 moral judgements.

To be clear, one can wonder whether the constraint that alternative A should be morally preferred to alternative B iff the level of goodness of A is higher than the level of goodness of B gets its intuitive force from the fact that we do

actually attach different degrees of goodness to different propositions and construct our betterness relation on these degrees according to the constraint (in the same way that we attach different degrees of length to objects and construct our "longer than" relation on these degrees) or whether it is the other way around: we start with our type 2 judgements and then construct the degrees of goodness of propositions from these judgements (in the same way that we can construct degrees of "leftness" relative to us from our judgments regarding which object is to the left of which object relative to us)[7]. However, whichever option you find more plausible, you must admit that in order to find out the levels of goodness of different propositions you can consult your judgments regarding the betterness relation that holds between them, because if you have enough of these judgments you can construct a scale (unique up to affine transformation) that measures their level of goodness given that these judgements respects the axioms of Bayesian decision theory (choose your preferred version).

Thus, it seems that the route to a method of changing one's intransitive moral judgements must go through one's type 2 moral judgements and not through one's type 1 moral judgements, as in order to measure degrees of goodness we must turn to our judgements regarding betterness. The DBT was discussed in length in the philosophical literature[8]. Lewis himself (Lewis 1996) has extended his result to cover other possible formulations of the non-Humean position. Other scholars have proposed various extensions of the

[7] John Broome argues explicitly for the later: "...goodness is actually fully reducible to betterness; there is nothing more to goodness than betterness." (Broome 2000, pp. 163-4).
[8] See for example, Price 1989, Broome 1991, Costa et al. 1995, Hajek and Pettit, 2004, Bradley and List 2009.

result and some attempts to escape it. However, in all of these discussions the concern was still moral judgments of type 1. In what follows I want to explore the other possibility.

Before doing so, however, it is important to be clear what type 2 moral judgements are. Again, we can distinguish between a Humean and a non-Humean answer. According to the non-Humean approach these judgements are just beliefs regarding the betterness relations, and thus degrees of firmness of these judgements are just degrees of beliefs. This picture makes sense also from the point of view of reasoning since if a rational agent finds out that he has inconsistent beliefs (i.e. his belief in the transitivity requirement and his intransitive judgements) he must change some of them.

What will be the Humean treatment of type 2 moral judgements? The immediate answer will be that these judgements express the agent's moral preferences (which are the result of his moral desires). However, if an agent has intransitive type 2 moral judgements this cannot be the case, according to the Humean position, as there is no desirability function that can represent these judgements. The Humean can, of course, claim that in such a case the agent's type 2 moral judgements express his beliefs regarding what his "real" (i.e. the ones that order the different alternatives according to their degrees of goodness) preferences are. This, however, just amounts to adopting the non-Humean position according to which type 2 moral judgements are beliefs regarding betterness while suggesting a specific interpretation of what constitutes the betterness relation.

If the Humean really wants to keep beliefs and desires separate, he must find a way to treat type 2 moral judgements (even when they are intransitive) as an expression of desires, but it is not entirely clear how this can be done. John Broome (2006) suggests that they might be conditional desires. If this is so then one should be able to explain exactly how their degrees relate to degrees of unconditional desires, and even if one finds a good answer to this question, one should still deal with a more troubling challenge which is to give a plausible account of how an agent can change his desires by reasoning. Generally, I find the Humean position regarding type 2 judgements less promising, but even if one can develop this line of thinking, a very similar result to the one that will be presented in the next section will probably hold for it, as long as one takes the degrees of firmness of type 2 moral judgements to be bounded (the reason will be clear after the result is presented).

Adopting the non-Humean position regarding type 2 moral judgements and the second method of changing one's intransitive moral judgements amounts, I believe, to accepting the reflective equilibrium method. In this method we start with a moral theory that we find acceptable (i.e. a theory that tells us what to choose in every situation in which we face a moral decision) and then, in case we judge some of its recommendations for specific cases to be morally wrong, we either adjust the theory or change our judgements concerning the specific cases. The process ends when we find a moral theory, whose recommendations for specific cases (all of them) we judge to

be morally right. In such a process we must reject some of our initial moral judgements and replace them with others after we realize that they conflict with other moral judgements that we have (i.e. the ones prescribed by the theory we hold) or else we must reject some of the moral judgements the theory prescribes to us and replace them with others after realizing that they conflict with some other moral judgements that we have.

Almost nothing was said in the literature about the way we decide which initial moral judgements to keep and which to reject and what (if any) constraints should guide us in such reasoning. In the next section I will suggest two conditions that any plausible method to do so must respect. I will then explore the consequences of these conditions and show that they lead to a problematic result.

Can firmness of moral judgements guide us in our moral reasoning?

It will be helpful at this point to move to a somewhat more formal discussion. This will be done by presenting an extension of Savage's (1972) model for decision making that incorporates beliefs about betterness. We will also assume - for simplicity - that unlike in Savage's model, the agent's subjective probability function is given

Let $\Omega = \{\omega_1 \ldots \omega_n\}$ be a finite set of possible states of the world. Let p be a probability distribution over Ω. Let $D = \{A, B, C \ldots\}$ be a set of outcomes and let $A = \{a_{1 \ldots} a_1\}$ be a set of acts, where an act is a function from Ω to D, and let ≥*

be a regular preference ordering over A (i.e. a complete, reflexive and transitive relation). In addition let >**denote the betterness relation between pairs of acts i.e. >** is a binary relation over elements of A such that for any two elements, a_i and a_j, a_i >** a_j or a_j>** a_i [9]. 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 > (we will, of course, use the same notation to denote the arithmetic "greater than" relation, but it will be easy enough not to get confused between the two different uses[10]). 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.

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 of the world) 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 $q(a_i > a_j)$ when a_i is the constant act that gives A and a_j is the constant act that gives B.

[9] I am ignoring here the possibility that the agent gives a positive probability to the possibility that two acts are equally good or desirable, i.e. that no one of them is better than the other. This assumption will make the discussion simpler and nothing is dependent on it.

[10] Although my choice of notations is somewhat unorthodox, I have found it to be the one that will make the discussion that follows as clearest as possible.

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. We are looking for a "preference rule" that makes use of the agent's q when he has to choose which one of any two elements of A he (morally) prefers to which. The only plausible rule, I will argue, is "prefer the act to which you give a higher probability being the better act":

Likelihood of Betterness Constraint (LBC):

$q(a_i>a_j) > q(a_j>a_i)$ iff $a_i > {}^*a_j$ and $q(a_i>a_j) = q(a_j>a_i)$ iff $a_i = {}^*a_j$.

First, in order to avoid a possible confusion, it is important to stress that the LBC does not contradict Expected Utility Theory (and in fact, the proof in the Appendix shows that 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 to hold in the model it is clear that the agent in the model can be represented as maximizing expected utility for some utility function and the distribution "p". Now, 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" that 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 that 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.

Maybe it will be helpful to compare my model to Savage’s original model. In Savage’s model one cannot be uncertain regarding what one’s preferences are. This is so because the probability function p is defined over the set of states and there is no further probability distribution that is defined over a set of possible preferences ordering. On the contrary – given the utility function and p the preferences are determined so we cannot express in Savage’s model uncertainty regarding the preferences themselves. A similar situation holds in Richard Jeffery’s theory (Jeffrey 1965), where p is defined over the set of propositions but still given p and the desirability function the preferences are determined (Jeffery did suggest in a later paper[11], though, that we should try and incorporate uncertainty regarding the preferences relation into the model).

I do not want to take a position regarding the question whether being uncertain regarding one’s own preferences is possible or not. Those who accept that this is possible can think of the betterness relation just as the agent’s true preferences. Those who do not accept that this is possible can think of it as a separate ordering. In any case, however, the demand that the agent’s beliefs regarding this ordering – whatever it is - should constrain the preferences he adopts seems unavoidable.

[11] Jeffrey 1974.

Actually, the LBC can be interpreted in three different ways. Either you can take the agent's beliefs about betterness to constrain his moral preferences according to this rule or you can take the agent's moral preferences to constrain his beliefs about betterness according to it. Finally, you can take this rule 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) and then you will have to interpret the expression " the agent prefers a_i to a_j" as $q(a_i>a_j) > q(a_j>a_i)$ and the expression "the agent is indifferent between a_i and a_j" as $q(a_i>a_j) = q(a_j>a_i)$. Any of these interpretations will do.

The LBC seems to me highly intuitive but there is an immediate objection to it that must be dealt with. One might argue that there is one type of cases 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 if it is the case that A is better than B than A is much better than B and if B is better than A than B is only slightly better than A then even if the agent is almost sure that B is better than A it might be justified for him to prefer A.

It should be clear that I do not deny that whenever an agent can use this kind of reasoning, he ought to. My point is rather that in many cases one cannot use it, and these cases are exactly the cases I wish to concentrate on. Why is it that sometimes an agent cannot use this kind of reasoning? This question brings us back to the discussion in the previous section: in order to use this reasoning the agent must be able 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 type 2 moral judgements in the first place, it must be because he does not have a direct access to the levels of goodness of the different acts (as if he had he would just judge one act to be better than another iff it's level of goodness is greater than the other's, which guarantee transitivity).

Now it might be that although an agent does not have a direct access to the level of goodness of the different acts available to him, he does have a 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 a 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 most cases people do not have a direct access to a complete goodness function (conditional of 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 level they implicitly assign to the different acts, but if their conditional betterness

judgements are not transitive, than there is no goodness function that is consistent with them.

The question I want to explore in this paper is how should an agent that finds himself in such a position – but still accept transitivity as a requirement of rationality – construct his preferences. Arguing that he should do that 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 type 2 moral judgements in the first place.

The best way to look at the model presented in the beginning of this section is as a combination of two models, the traditional Savage's model for decision making under factual uncertainty and a model of how an agent chooses a preferences ordering when he suffers from uncertainty regarding the betterness relation. While the former aim 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 made by an agent with intransitive betterness judgements, who still accepts Savage's axioms. Thus, assuming that the agent already has an access to the goodness function before he chose his preferences ordering misses the point.

A more advance version of the same possible criticism on the LBC will be the following. Although we are sometimes uncertain regarding what is the morally

right thing to do, 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. This is the line of thinking adopted by most philosophers in relation to what is now known as "moral uncertainty"[12]. There are two different ways to interpret this position.

The first interpretation is a descriptive one, i.e. 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 their choice recommendations for 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, as according to this our judgements regarding what we ought to do in specific choice situations are derived from

[12] See Lockhart 2000, for example.

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 and reduce this uncertainty to an uncertainty regarding which one of several moral principles or theories is the correct one.

Such a normative requirement is, however, both vague and destructive for our ethical methodology. It is vague because it is not clear in what sense a moral principle (or a moral theory) is more than the sum of all of its choice recommendations for specific cases. To accept a moral principle, one might argue, is simply to accept all the moral choices it leads to.

It is destructive to 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.

So I think the “reducing moral uncertainty to uncertainty about moral theories” claim is implausible at both 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 purpose, 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 is must be so since he has intransitive type 2 moral judgments and still is committed to transitivity. In such a case, thus, commitment to the LBC just amounts to making use of all the information available and this is, I think, the only plausible thing to do.

However, notice that following the LBC does not necessarily lead the agent to have transitive preferences, even in the case of constant acts. For example when for three constant acts a_i, a_j and a_k, $q(a_i > a_j) > ½$, $q(a_j > a_k) > ½$ and $q(a_k > a_i) > ½$, the agent’s preferences will be intransitive. However, we can assume that if the agent realizes that his judgments regarding the bettereness relations between different pairs of constant acts are intransitive and yet he is committed to transitivity he will change his judgments – in some way or

another – so that they will be transitive, i.e. we can assume that that if $q(A>B)> ½$ and $q(B>C)> ½$ then $q(A>C)> ½$[13].

We do, however, want the agent's beliefs regarding the betterness relations between constant acts to impose certain restrictions on his beliefs regarding the betterness relations between non-constant acts. A natural restriction will be that agent's degree of belief that one act with uncertain outcomes is better than another such act should be equal to his expected degree of belief that this act is better than the other. In other words, the agent's degree of belief that one act is better than another should be equal to his degree of belief that the world is such that choosing the first act is better than choosing the other. Formally:

Expectation of Betterness Constraint (EBC):

For every two acts, a_i and a_j,

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

In words: the agent's degree of belief that act a_i is better than act a_k is equal to the normalized weighted sum of his beliefs that the outcome that a_i gives in any specific world is better than the outcome a_j gives in this world, when the weights are just the probabilities of the different worlds in which the two acts give different outcomes. Worlds in which the two acts give the same outcome

[13] I do not try to suggest that the question of how exactly the agent should change his moral judgements regarding the betterness relations between constant acts in case they are intransitive is a trivial one, but as in the following I will show that even with the assumption that the agent finds a satisfactory way to do that we face a harsh problem, it will be easier to use it.

are ignored – according to this rule – by the agent as in these worlds he is indifferent between the two acts.

This demand can be justified in the following way. If in a certain world, w_i, a_i gives one outcome, A, while a_j gives another outcome, B, the agent can split this world into two worlds, identical in everything except that in one world A is better than B (let us call this world w_{iab}) and in the other B is better than A (let us call it w_{iba}). If $p(w_i)$ and $q(A>B)$ (i.e. $q(a_i(w_i)>a_j(w_i))$) are probabilistically independent then the probability the agent gives to w_{iab} is just $p(w_i)q(A>B)$ and the same holds for w_{iba}. In such a case, what the constraint demands is that the probability the agent gives to a_i being better than a_j is just the probability he gives for the world being such that what he will get from choosing a_i is better than what he will get from choosing a_j, which seems the only reasonable thing to do.

One may reject the assumption that q(.) and p(.) are probabilistically independent on the ground that our moral judgments must depend on our factual beliefs about the world. While I believe this to be true, I also believe that in most cases we do not know how exactly this dependency work and so we usually have no reason to assume dependency between p(.) and q(.) (Remember that both p(.) and q(.) are subjective). In cases we do have some information regarding this dependency; it is, of course, always possible to add more worlds to the states set in a way that will make p(.) and q(.) probabilistically independent.

Also, it is important to notice that allowing dependence here demands more than the mere assumption that our moral judgments depend on some of our factual beliefs. It demands that they are dependent on our factual beliefs regarding the consequences of our choices, as the states of the world, as usually understood in decision theoretic models, do not incorporate every true fact about the world but only those facts that the agent thinks can influence the outcomes of his choices. While one might still argue that one's moral judgements depend on this kind of factual beliefs (e.g. he holds beliefs such as "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"), we certainly want to allow agents not to be committed to such dependence and so we must consider the case where p(.) and q(.) are probabilistically independent.

So we are looking at an agent that when facing a decision between two acts always prefers the act he believes to a greater degree to be better than the other and that believes that one act is better than another just to the degree that he believes that the outcome that he will get from choosing this act will be better than the outcome he will get from choosing the other act[14]. The question that has to be addressed now is for which qs these two conditions are consistent (for every possible p).

[14] It might seem, at a first glance, as if the first principle is a non-consequentialist one while the second is a consequentialist one, but I do not think this is what is going on here. The first principle is not necessarily a non-consequentialist principle as even consequentialists will accept that when an agent believes it is more likely than not that one outcome is better than another, and has no uncertainty about the consequences of his acts he should prefer the act that brings out the more-likely-to-be-better consequence (and remember that beliefs regarding constant acts are just beliefs regarding outcomes). In a similar way, the second principle is not necessarily a consequentialist principle as even non-consequentialists will agree that the agent's degree of belief that a certain act is better than another is just his expectation for this act to be better than the other.

Notice first that out of any three elements of D, {A, B, C} there must be two pairs such that q(A>B)≥ ½ and q(B>C) ≥ ½[15] and since we assumed that the agent's preferences over constant acts are transitive we know that q(A>C) ≥ ½ as well.

Consider now the set of all qs such that for at least three outcomes (or constant acts) A, B and C, such that q(A>B) ≥1/2, q(B>C) ≥ ½ and q(A>C) ≥1/2, q(A>C) < max {q(A>B), q(B>C)}. It is easy to see that for such qs the agent's preferences must be intransitive for some p. For example look at the following matrix:

	ω_1	ω_2
a_i	A	C
a_j	C	A
a_k	A	B

If $p(\omega_1) > p(\omega_2)$ the agent believes that ai is better than aj. Since q(A>C) < max {q(A>B), q(B>C)}, when $p(\omega_1)$ and $p(\omega_2)$ are close enough to each other the agent also believes that a_j is better than a_k. However, for every probability distribution, the agent believes that a_k is better then a_i and hence his preferences will be intransitive.

[15] As either q(A>B) ≥ ½ or q(B>A) ≥½. Assume the first one holds then if q(B>C) ≥½ we meet the requirement and if q(C>B) ≥½ than either q(C>A) ≥½ and we meet the requirement since q(C>A) ≥½ and q(A>B) ≥ ½ or q(A>C) ≥½ and we meet the requirement since or q(A>C) ≥½ and q(C>B) ≥½.

So a necessary condition for LBC and EBC to be consistent (assuming the preference relation is an ordering) is that $q(A>C) \geq \max \{q(A>B), q(B>C)\}$. However, this is not a sufficient condition. To see why, remember that we assumed that the agent's factual beliefs respect the laws of probability and that his preferences respect the Sure-thing principle as a consequence of our second condition (in which we ignore states of the world that give the same outcome for both acts). Thus there is a (unique up to affine transformation) utility function, u, such that when the agent maximizes his expected utility relative to this function and to p he gets his preferences.

Consider now the following two acts:

	ω_1	ω_2	ω_3
a_i	A	B	C
a_j	B	C	A

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:

$q(A>C) = q(A>B) + q(B>C) – ½$ [16]

This equation must hold for every three outcomes A, B and C such that $q(A>B) \geq ½$, $q(B>C) \geq ½$ (and thus $q(A>C) \geq ½$) and for every q that always (i.e. for every p) yields transitive preferences.

It turns out that this condition is also a sufficient one (the proof is in the Appendix).

An immediate consequence of this condition 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 that agent cannot give probability 1 to all such judgements.

Another consequence is that for any three outcomes, A, B and C such that $q(A>B) \geq ½$ and $q(B>C) \geq ½$, q(A>B) and q(B>C) cannot be both greater than ¾, as if they are q(A>C) must be greater than 1. In the same way for every four outcomes A, B, C and D, such that $q(A>B) \geq ½$, $q(B>C) \geq ½$ and $q(C>D) \geq ½$, q(A>B), q(B>C) and q(C>D) cannot all be greater than 2/3, and in

[16] And it is easy to see that this means that $q(A>B) + q(B>C) \leq 1.5$, i.e. there is an upper bound on how certain the agent can be regarding these two judgements.

general for every n outcomes, $A_n \ldots A_1$ such that $q(A_j > A_{j-1}) \geq ½$ for every $j \varepsilon \{1 \ldots n)$, all the $q(A_j > A_{j-1})$ cannot be, at the same time, greater than $n/2(n-1)$[17]. 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 preferred by all acts, regarding which he must be certain that the former is preferred to the latter).

I find this result very disturbing. Exactly how disturbing, will be discussed in the next section.

Can an irrational moral agent reason himself to rationality?

The result established in the last section is that given the following two conditions, an agent with beliefs regarding the betterness relations among different outcomes that do not obey the constraint that for every three outcomes, A, B and C, such that $q(A>B) \geq ½$ and $q(B>C) \geq ½$, $q(A>C) = q(A>B) + q(B>C) – ½$, must have intransitive preferences between some acts for some ps:

1. **EBC**: For every two acts, a_i and a_j,

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

.2. **LBC:** For every two acts a_i, a_j, $a_i \geq^* a_j$ iff $q(a_i > a_j) \geq q(a_j > a_i)$.

[17] 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 $n/2(n-1)$, $q(A_n > A_1) = 1$.

As intuitively it is not irrational to have degrees of beliefs that do violate the restriction it seems that we face a problem. I have already discussed in the last section why I think these two conditions should be accepted as requirements of rationality, but it might be helpful to stress again that this is true even from a Humean perspective. Here is why: on the face of it the result can be taken as an argument for a radical Humean thesis according to which preferences - even moral preferences - must be independent of beliefs of any kind (even beliefs about betterness). Thus, the Humean response to the result might be to reject LBC.

However, as was discussed in the first section, this response is too quick. If an agent has transitive preferences then this agent might not need to consult his judgments about betterness, but as was argued, most of us do sometimes find ourselves expressing (either by actual choices or by reflection) intransitive preferences. When we find ourselves in such a position - while we are unwilling to give up our commitment to transitivity - we must reverse some of our preferences among pairs of acts in order to retain transitivity. The question is how should we do that? The Humean answer to this question is that we should change our preferences in such a way that we will always prefer the act with the higher expected desirability. However, since our initial preferences were intransitive, there is no desirability function that can represent them, so we cannot consult them in order to get the desirability values of every outcome. The Desire-as-Belief Thesis can be seen as another method to get the desirability values of the outcomes – i.e. the agent can get them by setting degrees of beliefs on propositions of the form "proposition A is

desirable" - but as was argued, it leads to very unintuitive results, and in any case, it is a non-Humean thesis. So it seems that the Humean too must turn to his degrees of moral beliefs in order to decide which ones of his initial moral judgments he should reject. However, if his moral beliefs do not obey the restriction that for every three outcomes, A, B and C, $q(A>C) = q(A>B) + q(B>C) – ½$, he must have intransitive preferences if his beliefs respect EBS, and so he is not in a better position than the non-Humean.

How bad is this position? In order to examine that, let us imagine an agent that 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

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 with his friends he is not so sure anymore. He thinks that in

choosing act a_i he is being unfair to the people in village B as they get no chance to be saved. Furthermore, 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[18].

After he is done with 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 that, which he finds justified. Thus he goes to consult with his decision-theoretic expert friend. His friend solves his problem: 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 maximize 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.

[18] Notice that besides the fairness consideration, all the other possible considerations I have mentioned should apply to the sure outcomes as well. This does not change anything in the argument as the result holds for any initial q. If you prefer just ignore the other considerations or assume that the agent does not find them powerful enough to change his beliefs regarding the sure outcomes, but does feel they have some weight.

The problem is that our agent is not sure which acts out of a_i, a_j and a_k is better than which, so he decides to assign probabilities for all the possible betterness judgements and go with the higher probability (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 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[19]. Notice that he can do that 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 acts that can yield only these three outcomes with different probabilities his preferences will be transitive. He just has to choose degrees

[19] Another thing he can do is to give up EBC, but then he has two options: either to give up the whole idea that his beliefs regarding the betterness relations among outcomes constrain his beliefs regarding the betterness relations among acts or use another constraint that connects the two kinds of beliefs. I find the first option highly unintuitive, and I doubt that there is any constraint that can replace EBC, guarantee transitivity and allow greater freedom in the beliefs regarding the betterness relations among outcomes. I could not find one and in any case I find the justification that I gave in the last section to the EBC quite convincing.

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 judgment 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[20].

On the face of it, this seems like a positive result as what we have now is a consistent thesis that connects the agent's moral preferences to his moral

[20] Looking at it that way, makes it clear, I think, why 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.

beliefs, i.e. we have a consistent non-Humean thesis. However, this is not quite right. When the agent changes his degrees of beliefs 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.

To demonstrate that, assume that our agent has some free time before he has to make the decision, so he figures he should reflect a little bit 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 will be his moral judgements regarding acts that can bring with different probabilities any one of the three original outcomes or 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 the 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

judgments. As the agent thinks of more and more possible outcomes, his degrees of beliefs in his betterness judgments must be reduced more and more, and at the limit he must be morally indifferent between all outcomes (except between the one preferred to all and the one preferred by all, as was explained).

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 him 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 (e.g. 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.

Ways out?

If one finds – as I do – moral indifference to be an unacceptable option, he must give up something. Some will be inclined to give up transitivity of our moral preferences, but as I accept Harsanyi's demand that when one acts as a moral agent he should respect conditions of rationality at least as firm as those he respects when acting out of his own interest, I find this option very unattractive. Another option is to give up completeness of the moral

preferences ordering. This also makes intuitive sense as - as many have argued - many times we do feel that two alternatives are incomparable from a moral point of view. However, when one has to make a decision he must choose one of the acts available to him and if he really believes they are incomparable from a moral point of view then - when it comes to his making a decision - he is just morally indifferent between the acts. So in a sense rejecting completeness amounts to accepting moral indifference.

Another possible route one might take is to reject the assumption that degrees of firmness of type 2 moral judgements respect the laws of probability. This can be done in two ways. One can either develop a Humean account of type 2 moral judgements (which must include some solutions to the problems I have mentioned in the first section, particularly to the problem of changing one's desires by reasoning) or argue that type 2 moral judgements are beliefs but not probabilistic beliefs. As I already mentioned, though, it seems to me that adopting any one of these options will only help one to avoid the result (or something much like it) if one's account will allow the degrees of firmness of type 2 moral judgements to be unbounded, and this is , I think, very unintuitive.

Yet a fourth option is to give up any one of my two conditions. I have already discussed why this is unattractive too, but it is important to stress again that it is not enough just to reject them: one must suggest a plausible alternative to them that does guarantee transitivity. If one wants to reject LBC one should offer us another way to determine preferences when our intuitive judgements

are intransitive, and if one wants to reject EBC one should offer us another way to describe the way our judgements regarding the betterness relations among outcomes (or constant acts) constrain (or not) our judgments regarding the betterness relations among acts.

I doubt that there is a plausible way to do that regarding EBC that does guarantee transitivity. However, as already mentioned, there is one natural way to do that regarding LBC, i.e. to set our moral preferences according to the level of goodness or badness of the acts. The problem was that given that we do sometimes have intransitive preferences, we cannot consult them in order to construct a moral desirability function. However, if we had a moral desirability function, that is not constructed out of our moral preferences, available to us somehow, we could just use the values it gives us and ignore LBC.

Such a move amounts, I believe, to a rejection of the method of reflective equilibrium, as what an agent can do is just to adopt some moral theory that does assign exact goodness levels to different acts and if these lead him to unintuitive type 2 moral judgements he can just ignore his intuitions. Thus, my discussion can be seen as an argument against the method of reflective equilibrium, or more generally against any meta-ethical approach that treats our moral judgements regarding what we should morally do in specific cases as restrictions on our moral theories (Stich, Nichols and Weinberg, 2001 call them "intuition driven romanticism" approaches).

In the literature, rejection of the reflective equilibrium approach usually goes hand in hand with strong support for some version of utilitarianism. However, it is important to note that - in light of the discussion here - any moral theory that does offer us moral desirability values but only ones that are themselves supposed to be constructed out of people's preferences over acts will face exactly the same problem, and so preference-based utilitarianism, for example, is in no better position than most deontological theories in regard to the results presented here.

In any case, giving up the idea that our moral intuitions should restrict the moral theories we accept seems to me a too radical move to make, as even if this helps us to avoid the result presented here it does so only at the price of giving up any hope that our moral reasoning will affect our moral behaviour. William Frankena wrote "Morality is made for man not man for morality" (Frankena 1973, p. 116). The way this quote is usually understood is as stating that our moral theories should not be too demanding so that they will not become irrelevant, but a moral theory can be demanding not only in the sense that it requires from people to sacrifice too much in terms of their persona interests, but also in the sense that it requires the wrong things, that its recommendations are very unintuitive.

So I think some role must be left for intuitive moral judgements, but it might be that this role is more complicated than what the reflective equilibrium method suggests. Although the reflective equilibrium approach is - so I believe - the most developed and systematic treatment of the role of intuitions in our

reasoning in the philosophical literature, there are some other alternatives[21]. The key feature that might make any such alternative a possible candidate to solve our problem is treating our intuitive judgements not only as restrictions on or inputs into our moral reasoning but also as outputs of it. This, I think, can make violations of LBC more plausible, as maybe we should accept only the first part of Frankena's quote. Maybe morality is made for man **and** man is made for morality. In other words maybe we can use our moral reasoning not only in order to find answers to moral questions but also in order to change ourselves so that we will be more inclined to accept moral recommendations that we find unintuitive[22]. A more systematic treatment of this idea is, however, outside of the scope of this paper.

Conclusion

It was demonstrated that degrees of confidence in moral judgements are not suited - under plausible assumptions - to replace degrees of goodness. Since the degrees of goodness we attach to different acts must be such that our moral judgements are transitive, when our moral judgements are not transitive we cannot attach degrees of goodness to the set of acts available to us. It was argued that any plausible method of changing our moral judgements in order to gain transitivity must lead - at the limit - to moral indifference between all acts.

[21] Daniel Dennett's use of the concept of "intuition pumps" for example. See Dennett 1995 for a discussion.

[22] See also Haidt's (2001) social intuitionist model of moral judgements for a similar idea from a descriptive point of view.

In order to avoid this result at least one of the assumptions used must be rejected (and replaced by a plausible alternative), but it was argued that there seems to be no straightforward way to do it. It was also suggested that one possibility is to look for another source from which we can get the degrees of goodness of different acts. Such a source must be (almost by definition) a moral theory, but such a moral theory must be justified to us not (only) in virtue of the plausibility of its recommendations to specific cases, but rather on other grounds.

The discussion here was made in the context of moral judgements, but as was mentioned in the introduction, the same problem holds for any context in which a rational agent believes in the existence of some ”objective” betterness relation his preferences should mimic. In such contexts, whenever the agent realizes his judgements about this betterness relation are intransitive there seems to be no rational way for him to reason himself to rationality.

Appendix: A proof for the result

Theorem: Given that > is an ordering, 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,

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

2. LBC: For every two acts a_i, a_j, $a_i \geq^* a_j$ iff $q(a_i>a_j) \geq q(a_j>a_i)$.

If:

Since from the conjunction of LBC and EBC we know that for every two acts a_i and a_j , $a_i \succeq^* a$ 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 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:

$$\sum p(w_k)u(a_i(w_k)) \geq \sum p(w_k)(u(a_j(w_k))) \text{ iff } \sum p(w_k)q(a_i(w_k)>a_j(w_k)) \geq \sum p(w_k)q(a_j(w_k)>a_i(w_k))$$

Or:

$$\sum p(w_k)(u(a_i(w_k)) – u(a_j(w_k))) > 0 \text{ iff } \sum p(w_k)(2q(a_i(w_k)>a_j(w_k)) – 1) > 0$$

In order to do that let us define the utility function in the following way. Let A_1 denote the outcome that is preferred by all other 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 preferred by all outcomes except A_1 and so on until A_n:

$U(A_1) = ½$

$U(A_j) = q(A_j>A_1)$

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 w.l.g. k>m, then if m≠1:

$q(A_k>A_1) = q(A_k>A_m) + q(A_m>A_1) – ½$, 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 $\sum p(w_k)(q(a_i(w_k)>a_j(w_k)) – ½) > 0$

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:

A proof for this direction was already given in section 2.

References

Bradley, R, List, C (2009), Desire as Belief Revisited, *Analysis,* 69(1), 31-37.

Broome, J. (2006), Reasoning with Preferences?, in Olsaretti, S. (ed.), *Preferences and Well-being,* Cambridge University Press.

Broome, J. (1999), *Ethic out of Economics*, Cambridge University Press.

Broome, J. (1991), Desire, Belief and Expectation, *mind,* 398, 265-257.

Costa, H. A. Collins, J. Levi, I. (1995), Desire as Belief Implies Opinionation or Indifference, *Analysis,* 55(1), 2-5.

Dennett, D. (1995). Intuition pumps, *The third culture*, New York: Simon & Schuster.

Frankena, W. (1973), *Ethics*, Prentice-Hall, Inc.

Haidt, J. (2001), The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment, *Psychological Review,* 108:4, 814-834.

Hajek, A. Pettit, P. (2004), Desire Beyond Belief, *Lewisian Themes,* Oxford University press.

Harsanyi, J. (1978), Bayesian Decision Theory and Utilitarian Ethic, *The American Economic Review*, 68, 2, 223-228.

Jackson, F. and Smith, M. (2006), Absolutist Moral Theories and Uncertainty, *Journal of Philosophy,* 103, 267-283.

Jeffrey, C.R. (1965), *The Logic of Decision*, The University of Chicago Press.

Jeffrey, C. R. (1974), Preferences among Preferences, *Journal of Philosophy,* 71, 377-91.

Lewis, D (1996). Desire as Belief II, *Mind, New Series,* 105, 418, 303-13.

Lewis, D(1988). Desire as Belief, *Mind,* 97, 323-32.

Lockhart, T. (2000), *Moral Uncertainty and its Consequences,* Oxford University Press.

Price, H. (1989), Defending Desire as Belief, *Mind,* XCVIII, 389, 119-127.

Savage L, J. (1972), *The Foundations of Statistic*, Dover Publications.

Sepielli, A. (2009), What to do when you do not know what to do?, *Oxford studies in Metaethics,* 4, Oxford University Press.

Smith, M. (2002). Evaluation, Uncertainty and Motivation, *Ethical Theory and Moral Practice,* 5, 305-320.

Sunstein, R. C. (2005), Moral Heuristics, *Behavioural and Brain Sciences,* 28, 531- 542.

Weinberg, J. M. ,Nichols, S. , Stich, S. (2001), Normativity and Epistemic Intuitions, *Philosophical Topics*, 29, 1&2, 429-460.