Published in *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*, ed. Toni Rønnow-Rasmussen, Björn Petersson, Jonas Josefsson & Dan Egonsson, 2007.

Fresh Air: The Reliability of Moral Intuitions

Folke Tersman
Stockholm University & University of Auckland

1. Introduction

Imagine a closed room where the oxygen is running low. The people in the room are dozing off, and some are even on the verge of drifting into unconsciousness. When fresh air is finally let in, some are enlivened and energized, whereas others, who are in a sadder state, are less easily revived.

This image may come to mind when one ponders examples of when new empirical findings (from psychology, perhaps, or sociology or neuroscience) find their way into philosophical debates. Some get excited and write papers in which they indicate that the results have far-reaching implications. Others are more cautious, and tend almost automatically to suspect that research of this kind will leave philosophy as it is.

It seems to me that, boringly enough, it is often the sceptical attitude that wins out in the end. In the early papers, the arguments that are supposed to bring out the revolutionary implications are merely sketched. Later, when the details are to be filled in, things turn out to be more complicated than expected. The relevance

of the new data is questioned, and the objections are often so compelling that the findings are eventually forgotten. The problem is that so many moves are open to a clever philosopher, that people will soon figure out ways to accommodate the new data within just about any philosophical theory.

One recent example of when new empirical data have stirred enthusiasm is the research about moral intuitions by the philosopher Joshua Greene, in collaboration with some psychologists and neuroscientists at Princeton University (see [Greene et al 2001]). Greene and his colleagues used modern brain imaging techniques to explore what went on in people's brains when they were contemplating certain practical dilemmas. More specifically, they focused on Philippa Foot's well known 'trolley case' (see [Foot 1967], but also [Thomson 1967]). In the original version of the case, a runaway trolley will kill five people if it is allowed to proceed on its present course. The only way to stop it is to flip a switch that will turn it onto another set of tracks where it will kill one person instead of five. Should you flip the switch? Most people say 'yes'. At the same time, most people deny that it would be legitimate to stop the trolley by instead pushing a stranger from a footbridge above the tracks (we assume that we cannot stop the trolley by jumping ourselves, as we weigh too little). This may seem surprising. In both cases, we save five by sacrificing one, and consistency might seem to require that we judge them similarly.

What Greene found, however, was that, when people were contemplating these cases, different areas in their brains were engaged. When the subjects considered the footbridge case, certain brain areas associated with emotions were

activated. By contrast, reflection upon the original trolley case triggered areas associated with reasoning and cognition. This and similar results led the researchers to a general conclusion, namely that reflection upon 'personal' cases–cases that would involve a personal violation–engages people's emotions in a way that 'impersonal' cases don't. This is a very rough account of the results. But the details, for example regarding how to distinguish 'personal' cases from 'impersonal' ones, are not pertinent to the rest of my discussion.[1]

One of the philosophers who are impressed by these results is Peter Singer (see in particular [Singer 2005]). Singer thinks that they undermine a certain way of arguing in normative ethics, namely the strategy of criticizing a moral principle on the ground that it conflicts with common moral 'intuitions'. For example, according to utilitarianism, we should always act so as to make the outcome best (where this could mean, for example, that there is no alternative action that would bring about about a greater net sum of well-being). So utilitarianism entails that it could be right to kill one person in order to benefit others. Indeed, utilitarianism entails that this could be right even if the benefited people are already quite well off, and even if they gain very little (provided that they are sufficiently many). This appears highly counter-intuitive to many people, who therefore reject utilitarianism.

[1] For a fuller account of the original study, see [Greene 2002].

However, Singer thinks that Greene's results suggest that this strategy is not viable.[2] It is based on the assumption that intuitions should be treated as some kind of 'data' or 'evidence'. And, according to Singer, Greene's results undermine that assumption. This also means, he thinks, that they cast doubt over the method of reflective equilibrium.[3] For that method does indeed conceive of intuitions as a kind of evidence, at least in so far as they constitute 'considered moral judgments'. Roughly, the idea of reflective equilibrium tells us to start by articulating a number of plausible normative principles and then explore which of these that best squares with our considered judgments. If the principle that stands out as the most promising after this scrutiny still conflicts with some of the judgments, we should modify it (or discard the recalcitrant judgment, at least if there is an independent reason for doing so), until we reach 'equilibrium'. Having reached that state, the resulting principle is justified (for us). Singer has been sceptical toward the method of reflective equilibrium for a long time (see, e.g., [Singer 1974]), and thinks that Greene's results provide further support for that scepticism.

The purpose of this paper is to examine these contentions. I start by offering a definition of the notion of a moral intuition. I then distinguish between two ways in which Greene's findings lend support for a sceptical attitude towards intuitions.

[2] There is also recent work in social psychology that Singer takes to support his skepticism, such as the research by the psychologist Jonathan Haidt. See [Haidt 2001] and [Greene and Haidt 2002].

[3] We owe the notion of reflective equilibrium to John Rawls. See in particular [Rawls 1971]. For later developments, see, e.g., [Daniels 1979], [Brink 1989: Chapter 6] and [Tersman 1993].

I shall argue that, given the first version of the challenge, the method of reflective equilibrium can easily accommodate the findings and (*pace* what Singer thinks) remain a distinctive theory about the justification of moral claims. As for the second version of the challenge, I shall argue that it does not so much pose a threat specifically to the method of reflective equilibrium but to the idea that moral claims can be justified through rational argumentation in general.

2. Moral Intuitions

By a moral intuition, I mean a moral judgment that is accepted by someone not merely on the ground that he realizes that it follows from some moral theory or principle that he also accepts.[4] Given this definition, one can believe that there are intuitions without being committed to any of the metaphysical and epistemological claims associated with the intuitionism of philosophers like G.E. Moore, H.A. Pritchard and W.D. Ross. Thus, one is not committed to Moore's view that moral terms such as 'good' and 'right' stand for simple, unanalyzable non-natural properties (see [Moore 1903]), nor to the idea that we have some special cognitive faculty or organ by which we can grasp moral truths.

There are other conceptions of a moral intuition than the one I have chosen. Some reserve the phrase 'moral intuitions' for judgments that we make spontaneously without having given the evaluated case any serious thought at all.

[4] The view that intuitions are non-inferential in this sense is congenial with the views Henry Sidgwick expresses when he elaborates his 'philosophical intuitionism'. See [Sidgwick 1907: Chapter XIII]. For Sidgwick's views on this matter, see also [Audi 1996: 109, 131f].

Some even reserve it for those not yet verbalized 'gut-feelings' that precede the formation of a verbalized judgment. Singer sometimes seems to have such a notion in mind, but it is clearly too narrow in the present context. I want to explore the implications of Greene's results for the method of reflective equilibrium. And the judgments we are to test different principles against according to this method (our 'considered moral judgments') clearly include judgments about cases that we have reflected upon, while they exclude mere 'gut-feelings' that do not as yet constitute judgments (see, e.g., [Rawls 1971: 20-21, 47-48]).

What Singer is critical about is that intuitions are taken as *evidence*; i.e., as having an analogous role to that of observations when scientists are testing empirical theories. The obvious reason why observations have this role is that, normally, and unless we have some particular reason to suspect that an observational belief is formed under the influence of sub-optimal perceptual conditions, we have reason to believe that it is true. Singer takes Greene's results to show that it is implausible to assign a similar role to intuitions.

Let us call the status that is assigned to observations 'initial credibility'. I will assume that the claim that they have this status does not presuppose the view that they are incorrigible, indubitable, or true with certainty. Indeed, I even take it to be compatible with the coherentist view that it holds for any belief *p* that *p* is justified for a person *A* to the extent that *p* coheres with *A*'s other beliefs. If coherentism is correct, observational beliefs have initial credibility in virtue of the

fact that our theories about the world and our perceptual apparatus suggest that the beliefs are formed in a way that indicates that they are true.

3. Distorting Factors

Why are Greene's results supposed to undermine the claim that moral intuitions have initial credibility? Greene's results suggest that certain emotional responses have a causal role in the formation of at least some moral intuitions, such as the intuition that it would be wrong to push the stranger from the footbridge. Singer speculates about the evolutionary background of this mechanism. His idea is that the underlying emotional responses have evolved since they helped our ancestors to respond adequately in situations with a risk of violent conflict and where there was no time for much reflection. Due to the survival value of this propensity, it was passed on to further generations. The moral philosophers of the Stone Age (who thought too long about pros and cons in tricky situations) did simply not live long enough to get offspring.

Now, on one suggestion, the emotional responses underlying our intuitions represent a distorting factor on a par with the factors that lead us to discard observations, such as sub-optimal perceptual conditions. If someone claims to have seen a UFO we will be less impressed if we learn that he was heavily drunk. Similarly, if a moral intuition is influenced by the kind of emotional responses that prompt people's judgment about the footbridge, it too must be seen with suspicion, according to the first construal of the argument.

Why are the emotions supposed to be a distorting factor? Apparently, there is some research indicating that when people's judgments in other areas are affected by their intuitions, this detracts from their reliability, especially when they concern situations or cases that are significantly different from those that the feelings were formed to deal with.[5] We may trust the hunches of an experienced mountain guide when planning a trek on the mountain, but not when it comes to a walk in the Australian bush. The idea is that we can extrapolate on this research to reach a similar conclusion about our moral intuitions. Singer writes:

> There is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution as social mammals, primates, and finally, human beings. We should, with our current powers of reasoning and our rapidly changing circumstances, be able to do better than that. [Singer 2005: 348]

This may seem compelling, but some caution is needed. For, to begin with, it is not clear that conclusions from research that concern other areas can be extended to moral reasoning as well. After all, ethics is in many ways different, and it could be argued that the reliability of moral judgments, as contrasted with judgments in other areas, is in fact enhanced by certain kinds of emotional involvement. A

[5] For some research about intuitions (in the narrower sense, conceived as 'gut-feelings'), see [Baron 1998] and [Klein 1998].

well-founded evaluation of a moral dilemma usually requires information about which interests are at stake, and in order to gather such information it may help if we are capable of some amount of empathy. More generally, Singer's skepticism seems based on a rather crude picture of the role of emotions and their relationship to cognition and perception. Contemporary research suggests that emotions plays a crucial role in our cognitive endavours in that they help us to filter out irrelevant aspects which in turn may lead us to form more reliable answers to the questions we ponder (see, e.g., [Le Doux 1996]).

The idea that the engagement of one's emotions provides a distorting factor is therefore not obviously plausible. However, perhaps one can appeal to the fact that the emotional involvement means that it takes an *effort* to question the intuitions. This was corroborated by the fact that it took longer time for those subjects who did, after all, judge the footbridge case just like the switch case to reach their judgment (see [Greene et al 2001]). That is, it seems that, in the footbridge case, we must work against a certain automatic tendency. This tendency makes it easier to overlook relevant considerations and may therefore provide an obstacle to a reliable assessment.

4. Debunking Explanations

I think the argument now sketched at best gives very weak support for the claim that moral intuitions are not reliable. However, there is another, and better, way to construe the challenge.

As I wrote above, Singer thinks that Greene's findings fit within a broader evolutionary account of the origins of morality. In particular, they help to explain why the footbridge case is judged differently from the switch case. For in the switch case, there has not been a similar pressure to develop immediate emotional responses. Singer points out that, for most of the time in which humans have existed, they have lived in small groups, and violence was inflicted by 'hitting, pushing, strangling, or using a stick or stone as a club' [Singer 2005: 347f]. By contrast, the indirect way of killing people that the switch case represents is relatively new.

In one passage, it seems that Singer thinks that this explanation rules out that there is a morally relevant difference between the cases. He writes that

> the salient feature that explains our different intuitive judgments concerning the two cases is that the footbridge case is the kind of situation that was likely to arise during the eons of time over which we were evolving; whereas the standard trolley case describes a way of bringing about someone's death that has only been possible in the past century or two.

And he asks, rhetorically,

> what is the moral salience of the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred years ago? I would answer: none. [Singer 2005: 348]

However, Singer's way of stating the challenge is misleading. The fact that this particular difference has no moral salience does not imply that there is no relevant difference. The error stems from an ambiguity concerning the phrase 'explain an intuition'.

Suppose that *A* believes that the earth is round. To explain this belief could either be to explain why *A* entertains it, or to explain why the proposition that constitutes its content is true. An explanation of why *A* entertains the belief will presumably invoke assumptions about his education or upbringing. An explanation of the truth of the proposition that constitutes its content would rather invoke assumptions about cosmology and physics. Similarly, an explanation of the intuition that it is wrong to push the stranger (but right to flip the switch) could either be an explanation of why people have come to form that conviction or an explanation of why it *is* wrong to push the stranger (but right to flip the switch).

Now, the explanation of people's intuitions that Singer offers—the one that appeals to the fact that the switch case represents a way of killing that was not around when humans evolved—is of the first kind. It is an explanation of why people *think* that it is wrong to push the stranger but right to flip the switch. And one can accept this explanation without committing oneself to a particular view about why it *is* wrong to push the stranger but right to flip the switch. Thus, one can agree that 'the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred

years ago' has no moral relevance, and still think that there is a relevant difference between the cases.

However, there is a better way to bring out the significance of the explanation. The question is if moral intuitions can serve as evidence. One way to explain why observations have such a role in science is to say that, in many cases, the fact that we make an observation is best explained by assuming that the observation is true. For example, when sensory stimulations prompt us to believe that there are people around us, that is usually due to the fact that there *are* people around us.

What Greene's results suggest, however, when combined with the evolutionary story, is that the same does not hold for moral intuitions. In particular, it suggests that we can explain why we intuitively judge the footbridge case and the switch case differently without assuming that there is a morally relevant difference. Therefore, the fact that we do judge these cases differently provides no reason to think that there is a difference, or to reject principles (such as utilitarianism) that entail that there is none.[6]

I shall call an explanation of this kind a 'debunking' explanation.[7] Consider a fact *F* that is offered as evidence for a theory *T*. A debunking explanation of *F* is an explanation that does not entail that *T* is true or significantly likely. To provide such explanations is a common way to question the relevance of considerations

[6] So construed, Singer's challenge is simply a version of Harman's well known argument against moral realism. See [Harman 1977: Chapter 1].

[7] For a use of the phrase 'debunking explanation' similar to mine, see [Lillehammer 2003].

offered as evidence. For example, they are used for questioning witness testimonies in legal cases.

Twenty years ago, the Swedish Prime Minister Olof Palme was shot dead on a street in Stockholm when he was walking home in the evening together with his wife Lisbeth. The police eventually caught a man for doing it, a man by the name Christer Pettersson. The case was tried in court and Pettersson was found guilty, primarily on the basis of Palme's widow's testimony, as there was no physical evidence that tied him to the crime. Lisbeth Palme had identified him in a line-up, in which he appeared with a number of other men.

However, when the case was tried in the Court of Appeal, Pettersson was acquitted. For it was found that Palme's widow had got certain information before the identification that pointed her in the direction of Pettersson. She had been told that the suspect was an alcoholic and a homeless person, and Pettersson was the only such person in the line-up (the rest were police officers). Moreover, as Lisbet Palme was a social worker, she was familiar with signs of alcoholism. Her testimony was therefore not considered reliable.

These considerations undermine the evidentiary value of the widow's testimony since they provide material for a debunking explanation. For, although the explanation that appeals to the police's indiscretion does not exclude that Pettersson killed Palme, it doesn't assume it either. Therefore, since the truth of that explanation couldn't be ruled out, the mere fact that the widow pointed Pettersson out in the line-up wasn't thought to provide sufficient reason for thinking that he killed Palme. And since that was the prosecution's strongest card,

Pettersson was released. Similarly, since Greene's results provide material for a debunking explanation of people's tendency to think that it would be wrong to push the stranger but right to flip the switch, that piece of 'evidence' (i.e., the fact that they have this tendency) could also be questioned. This is the upshot of the second version of the challenge that appeals to Greene's results. I will return to the second version in section 6. In the next section, where I discuss the implications for reflective equilibrium, I shall mainly be concerned with the first; the one that conceives of the emotional influence as a distorting factor.[8]

5. Reflective Equilibrium

In some passages, Singer's scepticism against the method of reflective equilibrium seems to stem not only from concerns about the reliability of our intuitions but from the view that it somehow misconstrues the whole point of formulating normative theories. Thus, he writes that the analogy between testing normative theories against our intuitions and testing scientific theories against our observations, is fundamentally misconceived, since

> [a] normative theory […] is not trying to explain our common moral intuitions. […] For a normative moral theory is not an attempt to answer the

[8] Since Singer explicitly says that his criticism against intuitions has more general implications for moral methodology (in that it is supposed to undermine the method of reflective equilibrium), I shall ignore the possibility that he merely wants to question certain particular intuitions, namely those that are supposed to cast doubt over his own favorite principle (utilitarianism).

question ‘Why do we think as we do about moral questions?’ [Singer 2005: 345]

Instead, a normative theory is an attempt to answer the question ‘What ought we to do?’, and Singer suggests that the advocates of the method have overlooked that obvious fact.

However, this reasoning is fallacious, due to the ambiguity of ‘explain a belief’ mentioned above. To explain someone’s belief could either be to explain why he holds the belief or to explain why the proposition that constitutes its content is true. The fact that a normative theory is not meant to explain why we *have* our intuitions does not exclude that it should explain the *contents* of those intuitions. And it is only in the latter sense that normative theories should explain our moral intuitions, according to the method of reflective equilibrium.

But the challenge that has to do with the reliability of moral intuitions still remains. Why trust our intuitions, in view of the mechanisms that Greene and his collaborators have uncovered?

Singer acknowledges that Rawls stressed that it might occasionally be reasonable to reject some intuitions rather than the theory in case of conflict. In such cases, we should go ‘back and forth’, and both modify the theory and discard some of the conflicting judgments, until coherence is achieved. However, even given this feature of the method, it entails, according to Singer, that too many of our intuitions will have to be preserved. For he thinks that the criticism against intuitions that Greene’s results lend support for shows that a method of moral

reasoning is plausible only if it allows for the possibility that we end up with a moral theory that conflicts with *all* our 'common' or 'ordinary' moral intuitions. For example, he thinks it must not exclude that a plausible answer to the question 'What ought we to do?' is to say 'Ignore all our ordinary moral judgments, and do what will produce the best consequences'. And he adds:

> My point is that the model of reflective equilibrium, at least as presented in *A Theory of Justice*, appears to rule out such an answer, because it assumes that our moral intuitions are some kind of data from which we can learn what we ought to do. [Singer 2005: 346].

The qualification about *A Theory of Justice* is prompted by the fact that, in more recent discussions of the method, versions have been developed that appear to assign less weight to intuitions. In particular, he alludes to the distinction between 'wide' and 'narrow' reflective equilibrium that has been stressed by Norman Daniels.[9] The difference between a narrow and a wide equilibrium is that, in order to reach the latter state, our principles must not only cohere with our intuitions but also with certain 'background theories' (such as a conception of a person, or a theory about how moral beliefs are reliably acquired). This increases the revisionary element of the method. For, if the principles we ponder obtain support from such theories, then we have a stronger reason to hold on to them when they

[9] See [Daniels 1979]. However, Daniels stresses that the distinction was implicit already in [Rawls 1971], and explicit in [Rawls 1974/75].

conflict with intuitions. However, Singer believes that Daniels' move saves the method only at the cost of making it 'close to vacuous' (see [Singer 2005: 349]). Singer's claim about the method of reflective equilibrium can accordingly be stated: It is either implausible, as it is too conservative relative to our moral intuitions, or (almost) devoid of content. In the rest of this section, I argue that this claim is false.

One thing to note about Singer's view is that it is merely our 'common' or 'ordinary' intuitions whose rejection a plausible method must allow for, not *all* our intuitions. Indeed, Singer assumes, or wants to leave open, that there are some intuitions (those he calls our 'more reasoned conclusions') that are not vulnerable to the criticism (more about those later). Moreover, the intuitions he appears to be most sceptical about, and those that the brain research seems most relevant to, are the ones that constitute mere spontaneous reactions or gut-feelings.

One problem with Singer's reasoning is that this set does not correspond well with the subset of our intuitions that are taken as evidence by the theory of reflective equilibrium, namely our 'considered moral judgments'. The considered judgments of a person are, roughly, those that are held with some confidence, not distorted by self-interest and prejudice, and based on well-grounded information and sound inference patterns (see [Rawls 1971: 20-21, 47-48]). The idea is to filter out those intuitions that we have some particular reason to be suspicious against. Therefore, our considered judgment may well include intuitions of the kind Singer is more sympathetic toward (those that are 'more reasoned'), whereas

mere spontaneous 'gut-reactions' are excluded. So Singer's criticism largely misses the target.

Another problem with his reasoning is that the notion of a considered moral judgment is open to revision. Consider the finding that some of our intuitions are influenced by certain evolved emotional responses, and suppose that this means that they are not reliable. Maybe Rawls did not think about that when defining the concept of a considered moral judgment. But there is no reason why new knowledge should not lead us to revise the definition. Rawls wanted to exclude judgments that are formed under the influence of distorting factors. So, if influence of the kind Greene has uncovered is one such factor, we should accommodate his results by excluding intuitions thus influenced from the set of our considered moral judgments. This is congenial with the dynamic nature of the method, and with the central idea that we should always be prepared to make further revisions in view of new considerations.

The method of reflective equilibrium can therefore accommodate the first version of the challenge that is based on Greene's results. Is it thereby devoid of content? Well, what does that mean? On one interpretation, a method of moral argumentation is 'vacuous' if there is no moral theory or principle that the method excludes our ending up with. But this notion of vacuity is irrelevant. The method of reflective equilibrium is an *epistemological* theory, not a normative one, and it is as an epistemological theory that the claim about vacuity should be assessed. I suspect that Singer's failure to see this has to do with the fact that he associates the method with anti-utilitarianism, perhaps as it was introduced by Rawls. But

there really is no such connection. A utilitarian may well accept that method, and attempt to use it for justifying her position.

Moreover, conceived as an epistemological theory, it entails several distinctive and controversial claims. For example, as it is a form of coherentism, it is incompatible with foundationalist ideas about justification. Foundationalists hold that there are beliefs that have a privileged status, in that they are justified independently of their relations to other beliefs, and that other beliefs are justified for that person only if they obtain support from such 'basic' beliefs. Coherentists, by contrast, deny that there are any basic beliefs, and insist that it holds for each of a person's beliefs that it is justified if and only if it coheres with the rest of his beliefs.

There has been some criticism against the claim that the theory of reflective equilibrium is distinctly non-foundationalist.[10] It is acknowledged, of course, that Rawls did not regard considered moral judgments as incorrigible. But it has been held that the theory presupposes that those judgments are privileged in the more modest sense that they have *some* degree of credibility independently of their relationship with other beliefs. Robert Ebertz mentions a number of considerations that are supposed to show this.

For example, Ebertz appeals to the fact that, at the start of the process of seeking reflective equilibrium, our considered judgments do not as yet have any support from any normative theory. However, we are still required to assign

[10] See, e.g., [Ebertz 1993] and Holmgren [1989]. Similar doubts have been expressed by Singer in that he writes that foundationalism is merely the 'limiting case' of the method (347).

weight to them when exploring such theories. Moreover, although further investigation may prompt revisions among our initial considered judgments, we are still, at the end, required to have *some* considered judgments to support our principles. Ebertz suggests that these requirements presuppose that such judgments are assigned an independent credibility (see [Ebertz 1993: 202]).

But Ebertz's arguments are not persuasive. For although the considered judgments we start out with are not yet supported by any normative theory, they obtain support from second-order beliefs about the reliability of (first-order) moral beliefs. It is these beliefs that explain why we should filter out those judgments that are affected by bias, ignorance of non-moral considerations, etc, and it is our considered judgments' coherence with them that, according to the method of reflective equilibrium, justifies our assigning initial weight to those judgments.[11] No assumption about *independent* credibility needs to be made. And as for the requirement that we, in the end, must still have some considered judgments, this is accounted for by the fact that coherence requires some complexity, in order for the relevant relations of coherence in our set of moral beliefs to obtain.

Another argument Ebertz uses in support of the claim that the method presupposes that considered moral judgments have some independent credibility is that, unless the method involves such an assumption, what it requires for justification would be too easy to attain. We could just pick the theory that our background theories seem to favour, apply it to various cases, and then replace

[11] See, e.g., [Brink 1989: 127 & 136] and [Sinnott-Armstrong 2006: Chapter 10]. Sinnott-Armstrong calls coherence between first-order beliefs and beliefs about the reliability of first-order beliefs 'second-order coherence'.

our initial considered judgments with the conclusions we thus reach. This would lead to instant coherence, without having to go through the 'back and forth'-process (see [Ebertz 1993: 204]).

However, it is doubtful if it is psychologically possible to just forget about one's initial considered judgments and replace them with the implications of the theory one explores. After all, those initial convictions are held with some confidence. But, even apart from that, things are not so easy. For 'replacing' the initial judgments in this way is justified, according to the theory of reflective equilibrium, only if it is allowed by the theories of reliability that account for the credibility of considered judgments in the first place.

But the anti-foundationalist element of the method is not its only distinctive feature. Another has to do with the concept of coherence that is used. Coherence is a matter of certain evidential and explanatory relations holding between the agent's moral views, where some explain and others are explained by the rest (relative to the agent's nonmoral beliefs). This entails that a reflective equilibrium, and thus also justification, is achieved only if the agent has come to accept certain general normative views. There are views that deny that the justification of moral beliefs requires acceptance of such principles. For example, it is denied by the approach called 'moral particularism' [see, e.g., Dancy 1993] that therefore sharply contrasts with the theory of reflective equilibrium. I conclude that the method of reflective equilibrium can accommodate the first version of the challenge that is based on Greene's results without dissolving into vacuity.

6. Can We Avoid General Scepticism?

Let us turn to the second version of the challenge that Greene's et al results lend support for, namely the idea that they provide material for debunking explanations. In this section, I want to address the question of whether one can use this line of reasoning without committing oneself to a more general scepticism about intuitions and the justification of moral beliefs.

Singer is open to the possibility that Greene's results might lead to a general worry about justification in ethics but wants to avoid such a conclusion. He concedes that he too must ultimately rely on intuitions (for example when defending his utilitarianism), such as 'the intuition that five deaths are worse than one, or more fundamentally, the intuitions that it is a bad thing if a person I killed' [Singer 2005: 350].[12] But the idea he pursues is that those intuitions are not vulnerable to the kind of criticism that Greene's and Haidt's research lends support for. The alleged reason is that they represent 'more reasoned conclusions'. This is shown, he thinks, by the above-mentioned finding about longer reaction times for those who do, after all, judge the footbridge case in the same way as the switch case. For, although they had the same emotional responses against pushing the stranger as the others, reasoning and reflection led them to a different answer. Morever, as for the intuition that it is a bad thing that someone is killed, he adds

[12] Singer is reluctant to call these beliefs 'intuitions'. But, given my definition of the term (and given Henry Sidgwick's), they clearly are. In any case, regardless of what we call them, the important question is if they avoid the criticism he raises against (other) intuitions.

that this intuition 'does not seem to be one that is the outcome of our evolutionary past' [Singer 2005: 350] since there is no reason to expect that a 'love for mankind, merely as such' would have evolved through natural selection. He concludes:

> [T]he 'intuition' that tells us that the death of one person is a lesser tragedy than the death of five is not like the intuitions that tell us we may throw the switch, but not push the stranger off the footbridge. It may be closer to truth to say that it is a rational intuition, something like the three 'ethical axioms' or 'intuitive propositions of real clearness and certainty' to which Henry Sidgwick appeals in his defense of utilitarianism in *The Methods of Ethics*. The third of these axioms is 'the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other'. [Singer 2005: 350f]

However, in what follows, I shall argue that the mere fact that reflection and reasoning has had a role to play in the formation of Singer's 'more reasoned' intuitions does not exclude that there is a debunking explanation of them too (or that the evolutionary theory has a role to play in such an explanation).

One of the strategies Singer uses in trying to compromise intuitions is to point out that they are a heritage from our Christian past:

> On abortion, suicide, and voluntary euthanasia, for instance, we may think as we do because we have grown up in a society that was, for nearly 2000 years, dominated by the Christian religion. We may no longer believe in Christianity as a moral authority, but we may find it difficult to rid ourselves of moral intuitions shaped by our parents and our teachers, who were either themselves believers, or were shaped by others who were. [Singer 2005: 345]

This connects with his point in an earlier paper:

> Why should we not rather make the opposite assumption, that all the particular moral judgments we intuitively make are likely to derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past? In which case, it would be best to forget all about our particular moral judgments, and start again from as near as we can get to self-evident moral axioms. [Singer 1974: 516].

But the Christian influence is not only pertinent to our intuitions about suicide but also to Sidgwick's 'axiom' that 'the good of any one individual is of no more importance, from the point of view […] of the Universe, than the good of any other'. Already from the start, Christian ethics involved the belief that many differences that had previously been regarded as morally relevant, such as

ethnicity or differences in class, are not in fact so. Any person could be a Christian, and the love towards others that is prescribed by the faith should be extended beyond family and tribe, and even to people beyond the Christian community. This was something entirely new, and could not be found in, for example, Judaism or the pagan religions that at the time existed in the Roman Empire. Thus, consider this quote from a letter by the early bishop Cyprian of Carthage (born around 200 AD) to his congregation:

> [T]here is nothing remarkable in cherishing merely our own people with the due attentions of love, but that one might become perfect who should do something more than heathen men or publicans, one who, overcoming evil with good, and practicing a merciful kindness like that of God, should love his enemies as well […]. Thus the good was done to all men, not merely to the household of faith. (The quote is found in [Harnack 1908: 172-173].)

In fact, as Rodney Stark argues in his book *The Rise of Christianity*, this aspect of Christianity has probably strongly contributed to the fact that it became the dominant religion in the highly multi-cultural and multi-ethnic Roman Empire. The economic and political unity that Rome had created had led to a cultural chaos, where people with different gods, languages, and upbringings had been dumped together helter-skelter in cities and army units. It is easy to see that Christianity served an important function in this context, as it offered a universalistic and seemingly coherent morality stripped of ethnicity. It is equally

easy to see how this heritage, that as Singer stresses has had a deep impact on the culture of the present day, has encouraged the train of thought that leads to the conclusion that the good of no one is more important from a moral point of view than the good of any other, especially in the case of a philosopher like Sidgwick who so strenuously searched for consistency and generality. Indeed, replace 'Universe' with 'God', and you get a doctrine that will impress many a Christian. The emergence of this aspect of the Christian faith is clearly one step of the 'expansion of the circle' that Peter Singer so often writes about.[13]

But there are plenty of other possibilities. Thus, consider the intuition that the death of five is a worse thing than the death of one. If badness is conceived as a quantity, it is easy to see how the mere fact that the number five is greater than one can lead to that conclusion. Although this is a simple example, it illustrates something important, namely that, when we reflect on ethical issues, basically the same processes are at work as those that operate when we ponder other issues. We search for simplicity, coherence, and generality. Given just some prior convictions

[13] Of course, this is a debunking explanation only if we can explain the emergence of Christianity as a cultural force without assuming that any of its basic ethical beliefs are true. But since the explanation merely appeals to the social function of the religion, it satisfies this condition. Moreover, Singer needs to make exactly the same assumption in offering his debunking explanations of people's intuitions about suicide. Also, notice in this connection that Singer thinks that one consideration that helps to explain why humans have evolved a propensity for moral thinking is that it has helped them to solve various Prisoner Dilemma-type coordination cases, and that it therefore has been selected through natural selection (see [Singer 2005: 335f]). Such considerations could also be invoked in explaining the tendency (which is central in Christian ethics) to ignore differences between individuals that Sidgwick's axiom is all about.

to work with, these processes prompt us to move on and to reach further 'reasoned' conclusions.

For example, consider a recent example by the modern intuitionist Roger Crisp. Crisp suggests that the following principle can be grasped by intuition and is 'self-evident in being justified by that grasp, perhaps to the extent that it is entitled to be called knowledge' [Crisp 2006: 73]:

> *The Self-Interest Principle* (*SI*). Any agent at time *t* who has (*a*) a life that can go better or worse for her, and *(b*) a range of alternative actions available to her at *t* which will affect whether that life does go better or worse overall for her has a [not necessarily conclusive or overriding] reason to act at *t* in any way that makes her life go better overall […].

A debunking explanation of SI can be given along the following lines. It is safe to assume that at least some concern for one's self-interest is the result of evolutionary pressure, and the conviction that we have a *reason* to act self-interestedly can be seen as a way of verbalizing that concern, given the role of such judgments in planning and deliberation. The universal element of SI—the part that entails that it holds for everyone—needs another explanation. But then we can appeal to the cognitive processes mentioned above. We search for generality and coherence, and try to find relevant similarities and ignore irrelevant differences. If we restrict the scope of SI, we need an explanation in terms of relevant differences between the persons for whom it holds and those for whom it

does not hold. The universal version does not require such complexity, and is therefore attractive for the reflective mind that seeks simplicity. So, the fact that reflection on SI can prompt us to accept it comes as no surprise.

Notice that the fact that this explanation invokes cognitive processes makes it no less debunking than the ones that appeal to Greene's results.[14] The fact that we have formed a conviction as the result, partially, of a preference for coherence and consistency does not in itself indicate that it is true or likely. It does so only if combined with the assumption that the beliefs that provided the *starting points* of that process are true or likely (just as we can trust the conclusions reached through valid deductive reasoning only if we can assume that the premises are true). However, in the case of SI, we never have to make that assumption, as we have a debunking explanation of the starting point (the evolutionary explanation of our concern for our self-interest). So, since the assumption that a further conviction has been reached through the search for coherence does not by itself entail that it is true or likely, the conjunction of that assumption with the debunking explanation of the starting point constitutes a debunking explanation also of the conclusion. I conclude that Crisp's intuition about SI, as well as Singer's 'more reasoned conclusions,' are as vulnerable to the challenge from the availability of debunking explanations as the ones Singer wants to criticize.

[14] Nor does it exclude them from being intuitions in the sense defined above. The fact that a preference for simplicity is involved in their formation need not mean that they are inferred from some other principle,

7. Concluding Remarks

This conclusion suggests that the second version of the challenge threatens to collapse into a general scepticism toward moral intuitions (including Sidgwick's 'axioms' and Singer's 'more reasoned conclusions'). If one wants to avoid such global scepticism, one must show that there are intuitions for which no debunking explanation can be given or where the debunking explanations are inferior to non-debunking ones. In order to succeed with such an endeavour, it is not enough to show that some intuitions are formed in part through reflection and reasoning.

What more is required? Let us say that if an explanation of an intuition entails that it is true or likely then it is 'validating'. In my view, if an explanation appeals to the way the intuition was formed, it is validating only if combined with an account of *why* the fact that it was so formed makes it true or significantly likely. And in order for that explanation to be a plausible one, the account in question must have some degree of initial credibility.

Unfortunately, modern intuitionists largely ignore the task of developing such an account. The defence they offer consists mainly in the assurance that intuitionism is not committed to the controversial metaphysical and epistemological claims associated with the early intuitionists (such as the view that there is a special organ or faculty for grasping moral truths).[15] For example, the closest Roger Crisp comes in his recent book to offering an account of the required kind is when he tries to explain the emergence of the capacity to intuit mathematical truths. Part of the explanation is, according to Crisp, that a being

[15] See, e.g., [Crisp 2002: 57-63] and [Audi 2002].

who grasps that when just two out of three previously observed predators (tigers) have been seen to leave the vicinity there might be one left is 'more likely to survive than an innumerate creature' [Crisp 2006: 87]. He then suggests that once that capacity has evolved it might have been extended so that it allows us to intuit normative truths as well.

However, this is, in my view, far from satisfactory. For example, suppose that our capacity to form moral beliefs through intuition can be given an evolutionary explanation. This does not in itself indicate that the beliefs thus formed are true or significantly likely. It does so only if we can show that the survival value of the capacity somehow *depends* on the fact that the intuitions it produces are true.

Finally, what is the problem of a wholesale rejection of intuitions as a source of evidence? The obvious problem is that this leaves us with fewer resources if we want to argue that moral beliefs can be justified. One remaining option is to try to explain how moral beliefs can be justified on the basis of purely non-moral considerations. However, the prospects of such a project are, for familiar reasons, bleak.

Acknowledgements

I am indebted to two anonymous referees for *Australasian Journal of Philosophy* for valuable comments that have significantly improved the paper. An earlier version was presented at seminars at Lingnan University, Hong Kong, University of Auckland, New Zealand, and University of Waikato, New Zealand. I also want to thank the participants of those seminars.

References

Audi, Robert 1996. Intuitionism, Pluralism and the Foundations of Ethics, in *Moral Knowledge?*, eds. Walter Sinnott-Armstrong and Mark Timmons, Oxford: Oxford University Press.

Audi, Robert 2002. Prospects for a Value-Based Intuitionism, in *Ethical Intuitionism: Re-Evaluations*, ed. Philip Stratton-Lake, Oxford: Clarendon.

Baron, Jonathan 1998. *Judgment Misguided: Intuition and Error in Public Decision Making*, New York: Oxford University Press.

Brink, David O. 1989. *Moral Realism and the Foundations of Ethics*, New York: Cambridge University Press.

Crisp, Roger 2002. Sidgwick and the Boundaries of Intuitionism, in *Ethical Intuitionism: Re-Evaluations* ed. Philip Stratton-Lake, Oxford: Clarendon.

Crisp, Roger 2006. *Reasons and the Good*, Oxford: Oxford University Press.

Dancy, Jonathan 1993. *Moral Reasons*, Oxford: Blackwell.

Daniels, Norman 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics, *Journal of Philosophy* 76: 256-282.

Ebertz, Roger P. 1993. Is Reflective Equilibrium a Coherentist Model?, *Canadian Journal of Philosophy* 23: 193-215.

Foot, Philippa 1967. The Problem of Abortion and the Doctrine of the Double Effect, *Oxford Review* 5: 5-15.

Greene, Joshua D. 2002. *The Terrible, Horrible, No Good, Very Bad Truth About Morality*, Ph.D. dissertation, Department of Philosophy, Princeton University.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J.D. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment, *Science* 293: 2105–2108.

Greene, Joshua D. and Haidt, Jonathan 2002, How (and Where) Does Moral Judgment Work?, *Trends in Cognitive Sciences* 6: 517-523.

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

Harman, Gilbert 1977. *The Nature of Morality*, New York: Oxford University Press.

Harnack, Adolf von 1908. *The Mission and Expansion of Christianity in the First Three Centuries*. London: Williams and Norgate (volume 1).

Holmgren, Margaret 1989. The Wide and Narrow of Reflective Equilibrium, *Canadian Journal of Philosophy* 19: 43-60.

Klein, Gary 1998. *Sources of Power. How People Make Decisions*, Cambridge, MA: MIT Press.

Le Doux, Joseph 1996. *The Emotional Brain* New York: Simon & Schuster.

Lillehammer, Hallvard 2003. Debunking Morality: evolutionary naturalism and moral error theory, *Biology & Philosophy* 18: 567-581.

Moore, G. E. 1903. *Principia Ethica*, Cambridge: Cambridge University Press.

Parfit, Derek 1984. *Reasons and Persons*, Oxford: Clarendon Press.

Rawls, John 1971. *A Theory of Justice*, Cambridge: Harvard University Press.

Rawls, John 1975/75. The Independence of Moral Theory, *Proceedings of the American Philosophical Association* 48: 5-22.

Sidgwick, Henry 1907. *The Methods of Ethics*, London: Macmillan, 7th ed.

Singer, Peter 1974. Sidgwick and Reflective Equilibrium, *The Monist* 58: 490-517.

Singer, Peter 2005. Ethics and Intuitions, *The Journal of Ethics* 9: 331-352.

Sinnott-Armstrong, Walter 2006, *Moral Scepticisms*, Oxford: Oxford University Press.

Stark, Rodney 1996. *The Rise of Christianity. How the Obscure, Marginal Jesus Movement Became the Dominant Religious Force in the Western World in a Few Centuries*. Princeton: Princeton University Press.

Tersman, Folke 1993. *Reflective Equilibrium. An Essay in Moral Epistemology*, Stockholm: Almqvist & Wiksell.

Thomson, Judith J. 1967. Killing, Letting Die, and the Trolley Problem, *The Monist* 59: 204-217.