# How can we be moral when we are so irrational?[1]

*Nils-Eric Sahlin and Johan Brännmark*

ABSTRACT. Normative ethics usually presupposes background accounts of human agency, and although different ethical theorists might have different pictures of human agency in mind, there is still something like a standard account that most of mainstream normative ethics can be understood to rest on. Ethical theorists tend to have Rational Man, or at least some close relative to him, in mind when constructing normative theories. It will be argued here that empirical findings raise doubts about the accuracy of this kind of account; human beings fall too far short of ideals of rationality for it to be meaningful to devise normative ideals within such a framework. Instead, it is suggested, normative ethics could be conducted more profitably if the idea of unifying all ethical concerns into one theoretical account is abandoned. This disunity of ethical theorizing would then match the disunited and heuristic-oriented nature of our agency.

Philosophy and rationality are in a long-term relationship and have been so for thousands of years – since Plato's time at least. The core of this relationship is a commitment by philosophers to advance their investigations by rational means. And although there are parts of philosophy where people have let go of this commitment, the discipline as a whole is still faithful to it and will, we hope, stay that way. But rationality is not just an ideal shaping and guiding the discourse within the discipline; it is also something that tends to figure prominently in the content or presuppositions of philosophical theories. For instance, in areas concerned with human action, like ethics and decision theory, theories tend to be constructed with a specific kind of agent in mind, the rational person. Although there are substantial disagreements about the role of reason in human action, often couched in terms indicating some variation on the eighteenth-century dispute between sentimentalists like Hume and rationalists like Kant, these disputes tend to focus on whether reason sets ends. It is more or less assumed that actually achieving our ends is something that we do largely in accordance with certain principles of rationality. At the same time, a wide range of empirical studies have shown that our behaviour is often far from that of someone like Rational Man; and although the extent to which empirical findings should be treated as relevant to theory construction in ethics and decision theory is certainly open to discussion, it would be somewhat peculiar if the growing body of empirical studies of human decision-making really had little or no impact on the relevant areas of philosophical investigation.

In this paper we focus on the impact of scepticism about human rationality on theory construction in normative ethics. We will start by discussing the relevance of empirical research to what is fundamentally an investigation of an entirely different domain, the normative. Having argued that it is indeed relevant, we outline, first, an account of what is involved in rational agency. Here we aim to work with a relatively thin conception of rationality: one that is clearly embraced in mainstream decision theory and neutral with

[1] In no particular order, the authors wish to thank Melissa Finucane, Göran Hermerén, Isaac Levi, Anna-Sofia Maurin, Johannes Persson, Paul Robinson, Paul Slovic, Niklas Vareman, and Annika Wallin for valuable comments.

respect to the main traditional theories within normative ethics. We then rehearse some of the relevant empirical findings and describe the picture of human agency suggested by them. Following this, we raise doubts about the extent to which traditional ethical theories can possibly fit this picture; and, finally, we suggest an alternative model for ethical theorizing. While we certainly agree that philosophical investigations can proceed by rational means, we would still suggest that, when it comes to the investigation of normative ethics, we should not have rational agents in mind. As can be seen already from this outline, our argument covers a lot of ground and so we cannot provide more than a survey of it. We hope the survey will, however, strongly suggest that there is a real need to take seriously the possibility that normative ethics should be largely reoriented in order to better fit the realities of human behaviour.

## The Relevance of Empirical Findings to Normative Ethics

Normative ethics stands out precisely because it is devoted to the normative. This fact might make some people doubt whether empirical findings can be relevant to it at all. As any first-year student of moral philosophy will tell you, it is impossible to derive an "ought" from an "is". This old Humean dictum certainly does present a logical truth,[2] but the question is whether the methodology of moral theory really can support the conclusion that nothing descriptive ever has any bearing on the kind of normative theory we should embrace. There are least three reasons for being wary of taking such a stance.

To begin with, there is a question here about the point of moral theory. Quite a few moral theorists have accepted the idea that the point of the enterprise is ultimately to guide us as agents seeking to do the right thing. This kind of view can be found already in Aristotle and has been defended by contemporary theorists like James Griffin (1996) and Christine Korsgaard (1996). There is reason to evaluate ethical theories at least partly in terms of a practicality criterion: *ceteris paribus*, a moral theory is adequate to the extent that it provides guidance to human agents. One could certainly dispute such a demand by claiming that moral theory ultimately seeks merely to represent an independent moral order: some things simply are good or bad, some actions simply are right or wrong, and a good moral theory is one that gets these matters right. If we human beings happen to be unable, deliberatively, to handle these moral truths in a capable way, then that is a sad fact about us, but it does not impinge on the enterprise of moral theorizing. Still, even if we accept that ultimately it might turn out that the only tenable moral theory is one that actually provides little or no guidance, it would seem a reasonable methodological principle to investigate, in the first instance at any rate, ethical theories with the potential to provide us with some reasonable level of guidance.

---

[2] Hume's dictum is a mantra with border conditions. Bergström (1992), for example, argues that no *pure* value judgment could logically follow from a consistent set of purely empirical premises. But with disjunctive premises or conclusions, where the disjuncts are empirical statements and value judgments, the dictum can be questioned. For example: "X is not rational" implies "Either is X not rational or X should be honoured".

Recently, Pekka Väyrynen (2006) has argued that if we accept something like the practicality criterion, an important dimension in terms of which such a criterion should be applied emerges in the form of an availability condition – what he calls the Cognitive Condition: "For any strategy S for acting well and any type of moral agent A, S is available to A for use in her practical thinking only to the extent that satisfying the conditions for using S in one's practical thinking lies within the limits of A's cognitive capacities." As Väyrynen notes, the cognitive capacities of individual agents will differ depending on their innate abilities and a variety of circumstantial factors, so the interesting category here is that of normal human beings. While normality might be difficult to get an firm grip on, if it turns out that regular violations of axioms of rationality are the norm rather than the exception, moral theories resting on a rationality-centred picture of human agency will tend to struggle to fulfil the Cognitive Condition and hence be unlikely to guide us morally in a meaningful way.

A second consideration concerns another famous dictum which moral philosophers generally accept, namely the Kantian one that "ought implies can".[3] In its simplest form this just means that if people cannot do something, it cannot be the case that they ought to do it. This auxiliary principle bridges the is/ought gap. It allows us to draw some conclusions about the normative realm using findings in the empirical realm. Of course, one could reject the dictum and hold that sometimes we are to blame for not doing things even where it would be strictly impossible for us to do so. Tough luck, one might say, but morality really is *that* unyielding. Most of us still accept Kant's dictum, however, and the reason we do so is that moral rightness and wrongness are tied to the practice of holding each other responsible. Given this it seems absurd to hold people responsible for failing to do something that it would be impossible for them to do.[4] While the first consideration concerned moral thinking from the agent's perspective, this consideration can be said to concern moral thinking from the perspective of the bystander.

Now, given a responsibility-oriented rationale for the Kantian dictum, weaker versions of it will also tend to have some appeal. In its standard form, it is usually applied on a situational basis, but one could argue, more generally, that if a set of moral principles or ideals fits ill with our deliberative capacities, it cannot be the case that we ought to live by them. Note that here the point is not about the motivational "can" – whether we can bring ourselves to effectively will certain things – but rather about the cognitive "can":

---

[3] While it is true that Kant put this forward in the second Critique (1788), it should be noted that for him it figured as the major premise in an argument of affirming the antecedent rather than, as is usually the case nowadays, denying the consequent.

[4] There is a possible exception to this, namely moral dilemmas, situations where one ought to do A, ought to do B, and cannot do both. But these are still cases in which both A and B, taken individually, are possible actions; so if we choose to do A, we could be blamed for not doing B precisely because B was open to us. If we accept that moral dilemmas are possible, we accept the idea that there might be situations where we cannot avoid blame or wrong-doing; we do not *therefore* need to reject the idea that wrong-doing or blame presupposes possibility.

given the way human agency functions, there might be certain theories that we cannot live by given our cognitive limitations. To blame us for not living by such standards would then be like blaming small children for not being able to reason like adults. Of course, most children are at least on a path to adulthood; but if we have reason to believe that adults simply are not on a path leading to rationality, classically understood, and are instead moving towards a range of local improvements in their methods of coping with a range of diverse daily tasks, we should perhaps look for ethical theories that start out from such a conception of agency.

The third consideration is more overarching and concerns general methodology. If moral theorizing were about getting an independent normative order of things right, our methodological options would be somewhat slim – in fact, we would seem to be left with little more than the hope that nature has granted us a moral sense with which we can intuit what is good and what is right. If, by contrast, moral theory is about improving actual practices of moral deliberation and argument, the reflective-equilibrium procedure seems to be the obvious route to follow. Following Norman Daniels (1979), it has been commonplace to distinguish between two approaches to reflective-equilibrium. The first pursues narrow reflective equilibrium. In it principles and considered judgments are gradually adjusted with respect to each other until we reach a stable state. While most moral theorists tend to say little about the methodology they adhere to, many would probably say that this is what they do if pressed. At any rate, something like the reflective-equilibrium procedure certainly seems to capture what is actually going on in much contemporary moral theory, where principles are suggested and then tested against our intuitive assessments of a range of cases, and where one is expected to be prepared to revise some of those intuitive assessments, though certainly not all of them.

The wide reflective-equilibrium procedure adds a third category to the process, namely that of so-called "background theories". According to Daniels, these theories might include such things as "a theory of the person, a theory of procedural justice, general social theory, and a theory of the role of morality in society (including the ideal of a well-ordered society)." Obviously, background theories of this kind will tend to be fairly abstract and general, but they are clearly the type of theory on which empirical findings will have bearing. At this point, those who want to keep their moral philosophy pure will probably retort that there really is no need to opt for the wide rather than the narrow reflective-equilibrium procedure. The question is whether this response is tenable.

The reflective-equilibrium procedure was never intended as an innovation; rather, it sought to make explicit what most moral and political philosophers had already been doing implicitly. So one of the claims being made, when the ideal of wide reflective equilibrium is advanced, is that background theories already play a role in moral theory; it is just that their role remains implicit even when people make explicit their reliance on the narrow process of adjusting principles and considered judgments.

Background theories are important in that they provide a general picture of the kinds of people, practices, and societies that the rules of morality are supposed to regulate. Probably, we already grasp of these matters to an extent, and that is why we can so effortlessly focus exclusively on moral theory as such, but that effortlessness should not be allowed to conceal the importance of background theories. Assigned the task of modifying the rules of a sport, would we be satisfied with a copy of the rulebook, some examples of rulings, and then trying to improve on that? If we would, it was probably a bad idea to hand us the job.

This is not to deny the Humean dictum. We cannot deduce normative principles from empirically informed background theories. The point is that background theories play a crucial role in determining the proper domain of the principles. A certain set of principles might be better or worse for a variety of reasons, but first it has to be to the point. Our choice is not between relying on background theories or not doing so. It is between relying on them implicitly or explicitly. When we do normative ethics, we always have one or another agent in mind.

In the next two sections we will first delineate the kind of agent that philosophers often have in mind when reasoning about human action, many of them will explicitly rely on it, but even more of them will implicitly rely on something at least quite similar to it; we will then briefly go through some empirical findings that raise doubts about this standard account. Some who are familiar with this area might find that we go into unnecessary detail about matters that are well-established, others will no doubt find that it deserves a much fuller treatment. We have tried to strike a balance here and readers who would have liked a fuller treatment should keep in mind that the main point being made is just that in areas of philosophy, particularly normative ethics, which rely on background accounts of human agency, there is a real need to think much more about how normative theory should be done if standard account of human agency are perhaps mistaken.

## Rational man and Normative man

While philosophers have had rational agents in mind for a very long time, it was not until the twentieth century that they began, more formally, to conceptualize what it means to be such an agent. Before we turn to discuss the extent to which people in general tend to violate axioms of rationality, we must first spend some time on the axioms deployed in the more formal endeavour. Of course, it is impossible to say whether earlier philosophers, like Hume and Kant, would have accepted something like these axioms, but we will assume that what took place in the twentieth century was not an invention of a new philosophical conception, but a formalization of an old one.

In “Truth and probability” Frank Ramsey (1926/1990) describes a completely rational decision maker – let us call her, him or it “Rational Man”. Ramsey showed how

Rational Man's beliefs and desires can be measured with a betting method. He showed that if Rational Man follows a few intuitive principles of rational behaviour her, his or its "degrees of belief" can be represented by a probability measure. Furthermore, he showed that Rational Man consistently avoids Dutch books; and that having a coherent set of beliefs is a necessary and sufficient condition of doing so.

Ramsey also proved a representation theorem stating that Rational Man's preferences can be represented by a utility function determined up to a positive affine transformation. This theorem guarantees the existence of both a probability function and an unconditional utility function. The expected utility defined by those functions represents Rational Man's preferences.

Ramsey's Rational Man would argue that not all probability assessments are rational, contrary to what some Bayesians tend to believe. Rational Man should be well-calibrated. If, for example, an accepted (physical) theory tells us what the chances are of a state of affairs obtaining, Rational Man should adjust her, his or its beliefs accordingly.

In other words, Ramsey teaches us that, once we have described Rational Man's degrees of belief and preferences, it looks as if she, he or it maximizes expected (subjective) utility and acts in agreement with an expected utility theory. The fundamental idea of expected utility theory is that two main factors determine our decisions: our desires and our beliefs. And expected utility theory provides us with a model of how to handle our beliefs and desires, since it provides us with an account of how they combine in rational decisions. In any given decision situation, Rational Man chooses the alternative with maximal expected utility (the principle of maximizing expected utility).

Description of the decisions of Rational Man makes it appear (*de dicto*) that she, he or it follows a set of axioms or rules. In Ramsey's presentation there are eight of them. They can be separated into three types: behavioural, ontological and mathematical. Axioms of the first type restrict Rational Man's choices and preferences (behaviour). Ontological axioms tell us what there is. Mathematical axioms give Rational Man imposing computational powers and the observer of her, his or its behaviour enough mathematical muscle to prove the representation theorems that narrate the behaviour.

In the present context it is the rationality rules that are of interest. Basically there are two of them: *ordering assumptions* and *independence assumptions*.

Transitivity, which is an ordering assumption, is one kind of a rationality rule. Rational Man's preferences are assumed to be transitive, and we too, want our preferences to be transitive. If they are not, we risk becoming money-pumps.

The independence axioms of decision theory and game theory – e.g. Leonard Savage's Sure-Thing Principle – are another classical type of rationality axiom. Savage's

principle tells us that if an alternative *A* is judged to be as good as another *B* in all possible states and better than *B* in at least one, then a rational decision maker will prefer *A* to *B*. Savage (1954/1972) illustrates the principle with the following case (a burning question, also, of today):

"A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness of the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate was going to win, and again finds that he would do so. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, … [E]xcept possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance". (p. 21)

Similarly, moving to game theory, and assuming stated probabilities, the so-called strong independence axiom, tells us that for all outcomes *A*, *B*, *C*, and probability $p > 0$, *A* is better than *B* if and only if a prospect, *A* with probability *p* and *C* with probability 1 - *p*, is better than a prospect, *B* with probability *p* and *C* with probability 1 - *p*.

Savage formulated the Sure-Thing Principle in 1954. His view of probability, by his own account,[5] derives mainly from the work of De Finetti and is inspired by Ramsey's forceful discussion. Savage's theory, set out in *The Foundations of Statistics*, is in many respects similar to Ramsey's, but instead of giving us a narrative of an ideal he presents a normative theory: that is, Savage's theory tells us how we ought to choose. Ramsey's theory, by contrast, is descriptive. If we assume, however, that the decisions of the ideal agent, Rational Man, are the best that can be made, a normative interpretation of Ramsey is not far away.[6]

That man acts to maximize his own well being, given the information he has, is an old idea. It is found in the works of J. S. Mill, and we know it as the *Homo economicus* hypothesis. With Savage's theory *Normative Man* was reborn and the assumptions behind Normative Man were made into a testable psychological hypothesis.[7] Now the thought was that human reasoning and decision-making could roughly be modelled by subjective expected utility theory, i.e. a Ramsey-De Finetti-Savage type theory. Among psychologists, generic expected utility theory has since received wide approval as *the* normative model of rational decision-making.

---

[5] Savage (1954/72), pp. 4 & 7. See Levi (2004) and Sahlin (1990), Sahlin (2003).

[6] See Sahlin and Vareman (2008) for discussion of the various types of decision theory (descriptive, normative, and prescriptive).

[7] See Sahlin, Wallin and Persson (2008).

The Sure-Thing Principle and the ordering axioms are accepted by many (perhaps most) decision theorists. Though they sometimes appear counter-intuitive, they are considered cornerstones of rationality: principles that Rational Man follows and Normative Man ought to follow. Are they also cornerstones within the context of ethical theory? This is difficult to tell since ethical theorists tend to operate with a largely unexamined conception of agency, but basic instrumental rationality would still seem to form a core conception uniting ethical theorists from the main traditions. And since it is difficult to evaluate a conception of agency we can at best guess at, we have chosen here to focus on the decision-theoretical counterpart.

## THE IRRATIONALITY CLAIM

It has been claimed that we are irrational, the suggestion being that we do not stick to the behavioural axioms of classical theories of rationality, and that we frequently break the most basic and obvious laws of logic. This claim has been disputed and is frequently rejected by philosophers. Today, however, experimental data rather unambiguously show that we are no ideal decision makers – that we cannot claim to be intuitive logicians. Those who deny this seem to be quibbling about mere words rather than substance. Let us briefly review a number of classical and well-known examples of irrational behaviour.

*The certainty effect*. Probably the best-known violation of a classical behavioural axiom is the certainty effect, highlighted in various versions by Allais (1979) and Kahneman and Tversky (1979).

Kahneman and Tversky asked subjects to choose between four prospects. First, they made a choice between (A) and (B); then, between (C) and (D). The prospects were constructed as follows:

(A) 2500 (Israeli pounds) with probability 0.33, 2400 with probability 0.66 or 0 with probability 0.01.
(B) 2400 with certainty.
(C) 2500 with probability 0.33 or 0 with probability 0.67.
(D) gives 2400 with probability 0.34 or 0 with probability 0.66.

They found that 82% of the subjects preferred B to A, and that 83% preferred C to D – choices which, in combination, imply a pair of incompatible utility inequalities.

Clearly, it is the Sure-Thing Principle that is violated.[8] It is this axiom that asserts that if two alternatives have a common outcome for a particular state of nature, the ordering of the alternatives should be independent of that outcome. In other words, more than 80% of us are not fully rational.

---

[8] For critical discussion of this view see Levi (1997), ch. 10. See also Seidenfeld (1988).

*The conjunction fallacy*. Linda, Tversky & Kahneman (1983) told their subjects, is 31 years old, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. The subjects were then asked which of two alternatives – (A) Linda is a bank teller and (B) Linda is a bank teller and is active in the feminist movement – was the more probable. In this case 85% of the subjects answered (B). But (B) is the conjunction of (A) and "Linda is an active feminist", and a conjunction cannot be more probable than one of its conjuncts.

It has been questioned (see, for example, Levi (1985) and Gigerenzer (1991)) whether this behaviour involves a violation of probability theory, whether it really shows that we are irrational, or less than fully rational.[9] It has been asked whether, in one way or another, the behaviour can be explained away. On its face, however, it is a direct violation of the fundamental axioms of classical theory: Rational Man would be unwilling to join the 85% of respondents under suspicion.

*The base-rate fallacy*. In a study David Eddy gave physicians the following information. For a woman aged 40 who participates in routine screening the probability of breast cancer is 1%. If a woman has breast cancer, the probability is 80% that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 10% that she will still have a positive mammogram. Physicians were asked to consider a woman of 40 with a positive mammogram. Eddy then asked what the probability is that she in fact has breast cancer.

In Eddy's study 95% estimated the probability to be between 0.7 and 0.8. The correct value is 0.075. Eddy's observation, that we neglect base-rates, has been confirmed by several other studies. We make erroneous probabilistic inferences. Our guesstimates stand in glaring contrast with Rational Man's calculated estimate and fail to accord with traditional Bayesian probability (decision) theory.

It is obvious that this type of "irrational" inference, where we take insufficient account of known statistical evidence, can, directly or indirectly, cause harm or lead to

---

[9] It is important to keep in mind that within decision theory several alternative conceptions of rationality have been suggested and discussed. We have, for example, Herbert Simon's concept of bounded rationality, and Gerd Gigerenzer's concept of ecological rationality. (For discussion, see, for example, Gigerenzer et al. (1999) and A. Wallin (2003).) In this paper we will focus on what might be called "classical" conceptions of decision rationality and logical rationality. In light of the experimental findings it seems safe to say that philosophers (or logicians and statisticians for that matter) are as irrational (or rational) as the man in the street. Remember that some of the very best statisticians and decision theorists, faced with Ellsberg's problem, violated Savage's sure-thing principle. See Ellsberg (1961/1988) and, for discussion, Gärdenfors & Sahlin (1988) and Levi (2001).

maltreatment – and, if worse were to come to worst, could even lead to death. Failure to be fully rational has moral implications.[10]

*Overconfidence*. Psychologists have shown that we are overconfident. They have established this by asking subjects a series of knowledge questions. For example: Which city has more inhabitants? The subject is offered two alternative answers – in this example, Hyderabad and Islamabad. The subject is then asked how confident he or she is that the answer is correct, expressed on a scale of 50–100% in steps of 10%. The problem is that confidence does not match the relative frequency of correct answers. When an average subject claims that he or she is 100% confident that the answer is correct, the relative frequency of correct answers is around 80%. When the confidence estimate is 90%, the number of correct answers drops to around 75%, and so on.

This "bias" is not a direct violation of one or other axiom of decision theory. It is possible to be overconfident and at the same time remain perfectly coherent; and it is the latter that the classical theories demand – not that we are goal-rational, just that we are internally rational. But empirical findings strongly indicate that we are not well calibrated. As Ramsey pointed out, if one is a decision maker, being well calibrated is a good thing.

*The selection task*. The selection task was designed by P. C. Wason (1966). The subject is shown a set of four cards placed on a table. Each card has a vowel on one side and a number on the other side. The visible faces of the cards show A, 4, D, and 7. The subject is asked which cards he or she wants to turn over in order to determine the truth of the following proposition: If a card shows a vowel on one face, its opposite face shows an even number.

The most frequent answers are A or both A and 4. The correct answer is A and 7. Incidentally, it can be added that one of the two authors of this paper has, in teaching philosophy of science since the early 1980s, given his students the selection task. By now the informal results of well over 1000 respondents (undergraduates, graduates, researchers and faculty members) confirm Wason's original result. People exhibit fairly limited logical ability, and it seems not to matter how far you have advanced in your studies or career, or what type of training you have (humanities, medicine, engineering – or mathematics, for that matter).

Wason's selection task triggered a vast debate and discussion. In the course of this it has been shown, among other things, that people do better if the task is contextualised, and constructed in terms of social relations or everyday activities. But the fact remains: we are not intuitive logicians. We break the rules of logic. If logic is a prerequisite of rationality, we are less than fully rational.

---

[10] Levi has argued that the base rate fallacy is not a fallacy, i.e. not a violation of the classical theory. he claims that if we take the reference class problem seriously the experimental data do not show what they are supposed to show. See, for example, Levi (1981).

*Valuing lives*. Fetherstonhaugh, Slovic, Johnson and Friedrich (1997) have studied a phenomenon called "psychophysical numbing". Subjects were asked to point out the number of lives a medical research institute would have to save to get a $10 million grant. When 15 000 lives were at risk the median value was 9000 lives saved. When 290 000 were at risk, the median value was 100 000 lives saved. As Slovic (2007) points out, "respondents saw saving 9000 lives in the "smaller" population as more valuable than saving ten times as many lives in the largest" (p. 85).

Empirical studies of life-saving lead us to expect, Slovic says, that there will be more support for saving 80% of 100 lives that are at risk than there is for saving 20% of 1000 lives at risk. We prefer to reduce deaths from 4000 to 3000 rather than reduce deaths from 103 000 to 102 000. This contradicts the familiar nostrum that every human life is of equal value.[11] Slovic argues that studies of this kind show a cumulative collapse of compassion: the value of life-saving dramatically decreases as the number of lives saved increases. We simply fail to comprehend big numbers – an empirical fact with moral implications. The Darfur crisis is but one example.[12] The explanation given is that the affective response is greatest when the group consists of one (identifiable) individual, and that this response declines as the number of group members increases.

Affects are, according to contemporary cognitive psychology, a conscious or unconscious experience of goodness or badness and very context-dependent. Finucane et al. (2003) argue that growing quantities of empirical data support the claim that affective states are a key driver in choice and decision-making. The life-saving intervention studies appear to reveal a phenomenon known as "proportion dominance". Saving nearly all of the individuals in a small group is far more important, as a decision trigger, than the expectation that one will save the maximum number of individuals.

It is argued that a decision maker who behaves like Normative Man, or Ramsey's Rational Man, reasons logically and consciously, encodes reality in symbols, words and numbers, and seeks justification. These types of process are generally slow. Affective decision-making, on the other hand, encodes reality in images, metaphors and narratives. It reacts swiftly on the basis of memories or images of good and bad, pleasure and pain – a process that is generally swift. Affective decisions can lead to actions that are judged in retrospect to be irrational.[13]

---

[11] Of course, deontologists and consequentialists disagree over whether it is acceptable to kill some people to rescue a larger number (i.e. disagree on the question whether the duty of beneficence is itself morally constrained in certain important ways). The cases under consideration here avoid this issue and merely involve rescuing different numbers of people.

[12] See Slovic (2007).

[13] For discussion of the affects and decision-making, and so-called System 1 and System 2 theories, see, for example, Finucane et al. (2003), Kahneman (2003), Slovic (2007) and Sahlin et al. (2008).

*One-eyedness*. Numerous experimental studies have shown that we generate too few, and excessively narrow, hypotheses. In addition, we collect information and evidence in favour of our guesses that is too narrow, selected because it is readily available, and skewed in favour of our existing beliefs. Once we have a preferred hypothesis, we look for confirmatory evidence, neglecting countervailing evidence. We are, in other words, short-sighted and one-eyed when it comes to strategies of discovery and justification.

This idiosyncrasy need not make us irrational, if by that we mean incoherent and illogical.

Let us sum up this section. Almost half a century of psychological research has shown that, as decision makers, we are short-sighted, one-eyed and prone to a serious error of refraction. We definitely are not blessed with the sharp eyes of Ramsey's Rational Man, or of Economic Man. In this respect we are all more or less irrational.

On the other hand, it must be observed that the experimental findings point in various directions. There is more than one concept of rationality under attack here. First, we are irrational because we are not coherent, i.e. because we do not follow the rules of rational behaviour that appear in traditional decision and probability theory. Second, we are irrational because we break some basic rules of first-order logic, the core of all rational thought. Third, we are irrational because we do not have, and seem not to desire to have, calibrated beliefs. Fourth, we are irrational because we make decisions too fast – decisions which, reflecting on what we did, we realise, contradict our beliefs and values.

There may be ways to explain away some of these findings and insist that the studies do not show what they give the impression of showing, but the problem is just that the findings are as consistent as they are persistent. The mark of a reasonable framework for understanding human agency is not the capacity for finding within it alternative *possible* explanations for findings which the framework at the outset did not lead us to expect. We ourselves would favour a more piecemeal approach in this area, but piecemeal attempts to explain away the data in order to save a favoured account, will always tend to seem a bit like exercises in drawing epicycles. This is not to say that the standard account has been conclusively disproved, because it has not, but for the remainder of this paper it will be assumed that the claim of irrationality is largely corroborated by empirical findings: that in actual human behaviour we face what might be called *the fact of irrationality*. This is obviously not the same as saying that all of us are more or less irrational more or less all of the time. The point is rather that the picture of human agency that emerges from the empirical results is not one of human beings (as it were) halfway there to rationality, i.e. of beings whose problem-solving capacities largely build on a dim grasp of certain rules of rationality and logic, a grasp that we can possibly improve on, but which is essentially already there somehow. The emergent picture is rather one of beings possessing a virtual tool-box of heuristic devices, affects and instincts. These devices might be perfectly explainable by evolutionary processes;

they perform well in some contexts and lead us into error in others. Given such a picture, irrationality involving violations of classic axioms of rationality will be an endemic and ineliminable feature of human behaviour. This does not necessarily mean that there is anything wrong with us. What it does mean is that the background conception of human agency we should have in mind when thinking about normative ethics is not so much Rational Man as something like Heuristic Man: a person whose competence is context-dependent and whose deliberative skills do not necessarily transfer from one context to another; a person whose modes of thinking will inevitably be deeply shaped by a range of factors beyond agency as such.[14] Equally, the picture we should *not* rely on is that of an agent with a unified set of principles of rationality she, he or it simply applies to every type of situation. Of course, as decision theorists we might, perhaps, say with Isaac Levi (1997) that the behavioural axioms are for "rational angels", but for ethical theorists who profess to address us the problem is potentially a deep one: this is just not how we work.

## Moral Agency and Ethical Theory through Thick and Thin

In most cases background theories tend to reside, somewhat inaccessibly, in… well, the background. This means that although, to the reader of books and papers in moral theory, it will usually be reasonably clear what kinds of pictures of human agency different authors are working with, it remains very difficult to criticize them for relying on a specific account of human agency. And there is a further difficulty here insofar as it would hardly be feasible to look at all versions of the major types of theory in turn, one at a time. So we find ourselves not only with the need to generalize, but under pressure to do so with respect to things that are most often merely implicit. From the point of view of enquiry, clearly, this is far from ideal, but it should be kept in mind that our aim here is primarily to raise certain doubts, albeit potentially very serious ones, rather than to establish anything conclusively.

We would say that, within mainstream ethical theory, approaches to human agency, whether implicit or explicit, tend to fall into two main categories. To begin with there is a "thin" approach.[15] The emphasis here is *agency* as such. The aim is to provide an abstract characterization of what is involved in being an agent, the idea then being that humans happen to exemplify this notion. This approach is evident in the work of Kant, where being human simply means being a finite rational being, or more precisely, a

[14] We will not enter into it here, but psychologists have also shown that in decision situations gender, race, and social status influence risk-taking. See, for example, Finucane et al. (2000) and Slovic (2000).

[15] "Thin" and "thick" are not used here to distinguish between that which is purely normative and that which is both normative and descriptive; rather, it has to do with the degree of anthropological detail in accounts of human agency.

being that combines rationality with the possession of inclinations.[16] Even in a more fleshed-out Kantian picture of "humanity" abilities such as the capacity to represent ends and act on principles remain central. These are features, very clearly, that could be possessed by members of species other than our own, *homo sapiens* – something which can be seen as an advantage of the theory, as a sign of its universality. However, the problem is just that a general construal of agency with application to a range of possible species promotes a picture of agency that does not really capture actual human agency. How can we be thought to act on principles, in any interesting sense of the notion, when our deliberative capacities, as judged by traditional models of rationality, are as regularly erratic as they have been shown to be?

The standard interpretation, among Kantians, of the role of maxims is that we more or less always act on them,[17] and what this means is that for adult human beings there is always a generalistic aspect to our actions, a principled element. Even if we do not always act explicitly on the basis of some sort of policy, it is reasonable to insist that policy-like states lie behind our more specific choices. Children, it is said, are not capable of this, but somewhere along the line we supposedly acquire the relevant capacity. But the psychological research outlined above indicates that, on a line from children to Rational Man, human adults do not really move that far. In fact, it can be asked whether human development is mainly a matter of moving along such a line at all, rather than acquiring a wider and wider range of diverse tools with which we can handle ourselves in a wider and wider range of situations. It might be reasonable to conceive of someone like Rational Man always acting on maxims, but there seems to be little reason to suppose that actual human beings exhibit, in their behaviour, the kind of principled generality that would licence a self-understanding in which we always act, at least implicitly, on maxims – and this means that the categorical imperative is not a meaningful moral test for beings like us.[18]

Consequentialism is not consistently explicit, in the way Kantianism is, about its preferred image of agency; but it is difficult not to see, in the consequentialist approach, a universal rendering of traditional theories of rationality, with the agent being asked to

---

[16] As is shown by Jerome Schneewind in *The Invention of Autonomy* (Cambridge: Cambridge University Press, 1998), Kant's ethics should be understood against a background including the debate, between voluntarists and antivoluntarists, about whether God is above the moral law, or whether human beings and God basically belong to the same moral community. Kant adopted the latter stance, and with that came a conception of agency attempts to capture the agency both of God and of us. For Kant, the main difference here is that we have inclinations, which is why the moral law takes the form of an imperative for human agents; but the fact that we are so (very) far from God-like deliberative powers should surely be an enormously important factor in any attempt to understand human moral agency.

[17] For an influential interpreter, see Onora O'Neill (1989), Part II.

[18] At least not the Formula of Universal Law. It might be possible to construe a version of Kantian ethics strictly in terms of the Formula of Humanity that is compatible with a more realistic psychological account, but that would involve quite a radical turn from traditional Kantian ethics.

maximize overall, rather than his individual, good.[19] This move to a universalism involves some difficult issues in its own right, primarily having to do with the possibility of interpersonal utility comparisons and how aggregation across individuals is to be best understood; but even setting those (slightly) complicated matters aside, one might question the wisdom of any attempt to construct a moral theory relying on a more complicated counterpart of what might very well already be a misguided approach to individual agency. In the consequentialist framework, contexts enter in only as areas where the theory is to be applied. However, contexts are not just situations in which we happen to find ourselves; rather they seem to enter already in the formation of human decision-making capacities. Why, then, should we adopt a single and unified picture of ideal moral deliberation?

In contrast with the approaches of Kantians and consequentialists, there is a "thick" approach to human agency in which the emphasis is on the *human* element. The goal here is to understand our agency largely in terms of the emotional or affective capacities that we possess as members of a particular species, *homo sapiens*. This approach is most clearly seen in virtue ethics, especially in the Aristotelian tradition, although the ethics of care might also be taken to fall within this category.

If the main problem with thin theories is that they tend to be too ideal, a potential problem with thick theories is that they are not ideal enough. This is not the complaint that such theories are unable to provide us with ethical guidance,[20] but rather that, for a thick theory to be normative at all, it would have to involve not just a characterization of our affective capacities, but also an understanding of what constitutes good exercise of them. Of course, in Aristotelianism this understanding is presented in terms of the way in which reason gradually, over time, influences our affective capacities so that they become more accurate and insightful. Now, even if Aristotelianism takes a range of actual aspects of human psychology into account, it should be noted that it still operates with the notion of a unified ideal agent. Even if the *phronimos* is perhaps thought to be an unreachable ideal, the basic idea is still that a context is something into which the agent enters and then applies universal powers of reasoning. However, there is really no basis for thinking that the development of human beings runs along a track towards

---

[19] Consequentialism and decision theory have much in common. Good arguments against the axioms of the theories of rational decision-making are often, though not always, also good arguments against consequentialist theories. A difference is that decision theories more often deal with uncertainties: consequentialist theories sometimes ignore this complication. Someone might therefore argue that the problems of rationality listed above, the examples of irrational behaviour, are problematic for theories trying to incorporate uncertainty. This is only true up to a point. Experimental findings showing that we have intransitive preferences, that we reverse our preferences in an irrational way, that we fail to grasp big numbers, and that our utility (preference) assignments (constructions) are far too context-dependent, all hit consequentialist theories as much as they hit theories of rational decision making. See, for example, Slovic (2000) and Lichtenstein & Slovic (2006) for relevant empirical studies.

[20] Such worries are ably addressed in Ulrik Kihlbom (2000).

such a unified and balanced power of judgment.[21] Thus although the thick approach looks more promising than its thin cousin, its most prominent version, Aristotelianism, continues to still presuppose, at its core, a picture of human rational powers that would not seem to have any reasonable ground.[22]

Moreover, some have argued that while it involves a richer moral psychology than Kantianism and consequentialism, Aristotelianism relies on a bad psychological model when it comes to character traits: it assumes a model according to which human beings have stable and general character traits of a kind that, as a matter of fact, we simply do not have.[23] This is not the place to review the growing literature on this particular issue, though it might be noted that in many ways this attack on Aristotelianism has a certain kinship with the ideas put forward here – namely, that the way in which human decision-making is driven by concrete heuristics, and is context-specific all the way down, makes the ideal of the *phronimos* seem an unpromising one.

## A Possible Recourse: Two-Level Approaches

While the thick approach is closest to what will, in the end, be suggested here as a possible way forward, thin theorists have a response available that must be addressed before we proceed. For many thin theorists would not actually contend that the core theory should be used to guide our choices in any thorough-going sense. One can find both consequentialists and Kantians advocating some kind of two-level approach in which high theory is supposedly demarcated in a way that at least partially separates it from most of our daily routines and practices.

---

[21] It is well known that people improve in their decision making and problem-solving if the tasks are contextualised, and constructed in terms of, for example, social relations or everyday activities. But it seems safe to say that rationality is not a question of training. If that were the case one would expect scientific training to ensure that one does better in the selection task; but this seems not to be the case (see above). We know that we do not make decisions that accord with the decision maxims. But even more interesting is the fact that good arguments for the ideal rarely convince those to whom they are put of the correctness and applicability of the theory. In fact arguments for rationality seem to have a direct and negative effect on "rational" behaviour. Arguments for irrationality seem to do two things: first, they strengthen the convictions of those already irrational, and second, they make the rational irrational. See Slovic & Tversky (1974).

[22] When discussing Aristotelianism, we should keep in mind Aristotle's discussion in Topics. In book III of Topics Aristotle lays the foundation of modern preference logic. He works with two concepts of preferability: "worthy of choice" and "better", one situation-specific and one absolute, one describing common man and one describing, one could say, a rational angel. Among the principles Aristotle discusses is the contra-position principle, a thin version of Savage's sure-thing principle, the substitutivity condition, and a maxi-max rule of choice. These are principles which we can assume an ideal agent follows, but which are problematic, or not fulfilled, in terms of "worthy of choice". See Sahlin (1993).

[23] See John Doris (2002).

The consequentialist two-level approach builds on a vital distinction between accounts of right-making characteristics and decision procedures.[24] While the former provide necessary and sufficient conditions of an action being morally right, decision procedures are intended to be used in everyday life. The idea is that we need not demand, of the former, that it can be utilized as the latter;[25] it might be enough that the former can be used in the cool hour of reflection in order to evaluate in a general way our decision procedures and think about how to improve on them.[26] A consequentialist of this kind could accept something like Väyrynen's Cognitive Condition for the kinds of decision-making strategy that the consequentialist account of rightness should help us select.

Of course, this still leaves us with the problem that someone should ideally be able to evaluate our decision procedures in the light of a criterion of rightness, and accordingly there will always be a worry about whether such a massive calculative effort can be carried out to a satisfactory degree by creatures like us, even in a careful and collective effort. The admission, or discovery, that only angels could successfully execute this procedure would probably sever the link between consequentialist theory and our everyday practices; and then there might no longer be any substantial sense, even an indirect one, in which the theory provides us with guidance.

At a more principled level, one might ask whether the consequentialist criterion of rightness does not still presuppose something like Rational Man as a background theory of ideal agency within the context of the reflective-equilibrium procedure with which we are presumably trying to proceed in our ethical theorizing. If we *were* beings located somewhere on a developmental path towards this ideal, that might be reasonable. But if human deliberative capacities are primarily context-based and situational, then, while we might theoretically understand its workings, Rational Man might not be a relevant

---

[24] The *locus classicus* of this distinction is Eugene Bales (1971).

[25] Let us look briefly at the distinction between right-making characteristics and decision procedures from a decision-theoretical perspective. In this case the former provides necessary and sufficient conditions for rational decision-making, and the decision procedures are rules for everyday life decision-making. This distinction mirrors the one we made between rational angles and human decision makers, or Aristotle's distinction between "better than" and "worthy of choice". It also highlights an interesting problem. Decision theorists have tried to modify the axioms of the theory of rational decision making in order to make the theory less demanding – e.g. making the theory more prescriptive than normative, or allowing for imprecise probability and value judgments. See Sahlin (1985), and Gärdenfors and Sahlin (1988) for discussion of competing nonstandard decision theories. The problem is that no matter what we choose to give up or change (the axioms of independence or the ordering), we, as decision makers, run into all sorts of more or less serious and unwanted difficulties. See Seidenfeld (1988). And it looks as if it is only by following the true path of rationality (i.e. adhering to classical theories of rational decision making) that we can steer clear of the snags. Theories of rational decision-making are far more formalized than normative moral philosophies. But it can be anticipated that normative moral theories, moving in a similar direction, away from the ideal agent's absolute rationality, will run into analogous problems. Consequentialist theories will definitely do so.

[26] One could certainly take a further step, maintaining that, *qua* account of right-making characteristics, a theory can be acceptable even if yields the recommendation that we would be better off without it: see Derek Parfit (1984). Given the general methodological considerations above, such a stance would, however, be highly unsatisfactory.

ideal. If the distinction between decision procedures and the criterion of rightness could be mapped on to a distinction between our agency and the agency of an improved version of us, then certainly, at least within the context of a wide reflective-equilibrium procedure, it might make sense to adopt a two-level approach. But if our agency is not merely a slightly inferior version of Rational Man, but requires to be understood quite differently, the consequentialist two-level theory (at least) would seem to represent an unpromising avenue of research.

The Kantian version of the two-level approach deploys a distinction between ideal and non-ideal theory.[27] An ideal theory specifies the principles regulating a society where there is complete, or at least almost complete, compliance with the dictates of morality, and where no obstacles prevent the realization of ideal institutions. Non-ideal theory deals with how we should proceed if we find ourselves in a world where there are problems with compliance and/or the conditions needed for establishing the appropriate institutions. Armed with this distinction, of course, one might interpret the fact that we are often irrational as a problem of the second kind: our irrationality stands in the way of realizing (anything remotely similar to) a Kingdom of Ends here on Earth.

The problem with this approach is that, unlike the consequentialist two-level approach, Kantian high theory is still supposed to perform a guiding role in everyday life, even if we will often have to rely on non-ideal theory there as well. The reason for this is that non-ideal theory is not intended to be complete; it is just intended to be a complement. So Kantians are vulnerable to a dilemma: either they develop non-ideal theory to the point where it is more or less complete in its own right (in which case the approach as a whole will probably lose its distinctly Kantian character) or they stick with ideal theory of the kind intended to play a role in everyday life (in which case doubts inevitably arise about the viability of this ideal in view of our serious failings as rational beings).

## Outline of An Alternative Approach: Mid-Level Theory

These remarks on existing ethical theories are, of course, broad and sweeping. Many readers will undoubtedly conclude that their favoured theory is untouched by them. But our aim is not to disprove every current moral theory. It is rather to point to a peculiar fact about moral theorizing in general – namely, that so little attempt has been made to build theories by starting with the reasonable assumption that actual human agency does not approximate to the imagined agency of Rational Man. The argument here is not that the usual approach is an impossible one. It is rather that too much energy and time has been devoted to it rather than the investigation of other possibilities.

In comparing thin and thick approaches to human agency from this perspective, we are bound to feel that there is something right in the thick theory; the problem is just that, at least in its most common form, it still is a theory operating with a picture of human

[27] The *locus classicus* here is John Rawls (1971); see also Christine Korsgaard (1986).

beings as more or less rational (as this is traditionally understood). As already pointed out, actual human beings are probably better compared to the figure of Heuristic Man rather than the "almost there" Rational Man; and whereas most of this paper has been concerned with negative or destructive aspects of this observation, the most important question is what ethical theorizing ought to look like if we have the second picture of human agency in mind. One answer immediately suggests itself: if human decision-making is ultimately fragmented, in the sense that we have a diversity of concrete and often disunited ways of coping with different types of situation, then perhaps normative ethics should not be working towards unified and all-encompassing theories.

If we give up the hope of a grand theory uniting human problem-solving in the exercise of certain fundamental rules, another type of approach naturally invites itself – a more pragmatic approach. We can instead start by identifying typical human choice-situations and the types of strategy employed in them. From there we can move on, given a certain type of situation and the problem-solving strategies usually employed by humans there, to consider what would constitute an improvement in such strategies. The extent to which there can be improvements will not, however, be determined by an antecedent standard of rationality. Rather it will respect the actual layout of the types of situation in which we find ourselves. Room for ideals in such an approach remains. It is just that the ideals will arise out of the need to work our way out of particular clashes and conflicts between different aspects of our thinking rather than out of a will to achieve overarching unity.

It should be noted that surrender of the ideal of Rational Man as a background theory of human agency does not mean that we have to surrender rationality as an ideal in the conduct of philosophy. The latter is a slow, thoughtful, and collective effort, and just as other areas might have their own modes of thinking – modes that suit them – rationality, especially in playing by rules of logic, might very well be precisely the type of ideal that suits philosophy. It is just that there is a difference between rationally dissecting and considering different areas of human concern and activity, on the one hand, and having rationality as a standard defining what constitutes possible improvements of our actual decision-making processes within all of these areas, on the other. Changes to our decision-making that count as ethical improvements might or might not be available. To see whether they are, and to address the further question of wherein such improvements would consist, we would have to pay much closer attention to all the intricacies of such areas and to our actual practices of dealing with them than is usual in normative ethics today. Among current ethical theories, the one probably most in sync with this kind of methodological ideal is the ethics of care, for this does not seek an understanding of morality as a whole, but is focused on the ethical dimension of certain types of situation and relation.[28]

---

[28] For a leading example, see Nel Noddings (1984). However, even among some of its advocates, this feature of the ethics of care is sometimes seen as a drawback; and there are those who seek to expand it into a general ethical theory (e.g. Michael Slote (2007)). But what is really needed is not yet another grand ethical theory, but a greater range of mid-level theories.

This last approach, which might be called “mid-level theory”, is a piecemeal enterprise. It is difficult to say exactly what might come out of it, simply because there has been so little done resembling it within moral theory, but it is imaginable that something like the ideal of reflective equilibrium could still function as a model for it. It is just that one of the roles that would now be played by background theories is that of delineating the types of choice situation and scenario in which different problem-solving abilities and ideas (which might work satisfactorily in some contexts) will be out of their depth and perhaps clash with the workings of other abilities and ideas. One thing that needs to be emphasized, however, is that a piecemeal approach need not be toothless: it is quite possible that large areas of common-sense morality and current ethical practice implicitly rely on outmoded pictures of human agency, and that reform would in fact greatly improve them.

What *will* have to go is the notion of one fund of principles and one fund of considered judgments that should be adjusted with respect to each other. The situation is much more complex than that, and any movement towards reflective equilibrium will in all likelihood take place by way of a series of demarcations and local fixes rather than in any grand unifying transformation of the system as a whole. In a way, the moral theorist we see coming out of this is less an architect trying to redesign an existing building in order to get rid of a few flaws in the original design and more the janitor trying to find workable solutions to a wide range of problems within a sprawling building complex for which there never was an architect in the first place.

## References


Allais, M. (1953). “Le comportement de l'homme rationnel devant le risque : Critique des postulats et axiomes de l'école américaine”. *Econometrica*, 21:503-46.

Bales, R. E. (1971). “Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure”. *American Philosophical Quarterly*, 8:257-65.

Bergström, L. (1992). *Grundbok i värdeteori*. Stockholm: Thales.

Daniels, N. (1979). “Wide Reflective Equilibrium and Theory Acceptance in Ethics”. *The Journal of Philosophy*, 76:256-82.

Doris, J. M. (2002). *Lack of character: personality and moral behavior*. New York: Cambridge University Press.

Eddy, M. “Probabilistic reasoning in clinical medicine: problems and opportunities”. In Kahneman, D., Slovic, P., and Tversky, A. (eds.). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Ellsberg, D. (1961). “Risk, ambiguity, and the Savage axioms”. *Quarterly Journal of Economics*, 75:643-669. Reprinted in *Decision, Probability, and Utility: Selected Readings*, Gärdenfors, P. and Sahlin, N.-E. (eds.), Cambridge: Cambridge University Press, 1988.

Fetherstonhaugh, D., Slovic, P., Johnson, S. M., and Friedfrich, J. (1999). “Insensitivity to the value of human life”. In *The Value of Life*, Hermerén G. and Sahlin N.-E. (ed.), Stockholm: Almqvist & Wiksell, 13-31.

Finucane, M. L., Slovic, P., Mertz, C. K., Flynn, J., and Satterfield, T. A. (2000). “Gender, race, and perceived risk: the ‘white male’ effect”. *Health, Risk & Society*, 2:159-72.

Finucane, M. L., Peters, E. and Slovic, P. (2003). “Judgement and decision making: The dance of affect and reason. In Schneider, S. L. and Shanteau J. (eds.), *Emerging perspectives on judgment and decision research*, Cambridge: Cambridge University Press, 249-67.

Gärdenfors, P. and Sahlin, N.-E. (eds.) (1988). *Decision, Probability, and Utility: Selected Readings*. Cambridge: Cambridge University Press.

Gigerenzer, G. (1991). How to make cognitive illusions disappear. (In W. Stroebe and M. Hewstone (Eds.), *Review of social psychology* (83-115), Chichester: Wiley).

Gigerenzer, G., Todd, P. M., & the ABC Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Griffin, J. (1996). *Value judgement: improving our ethical beliefs*. Oxford: Clarendon Press.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*, 263-291.

Kahneman, D. (2003). “A perspective on judgment and choice: Mapping bounded rationality”. *American Psychologist*, 58: 697–720.

Kant, I. (1788). *Critique of Practical Reason*, trans. L. W. Beck. Indianapolis: Bobbs-Merrill, 1956.

Kihlbom, Ulrik (2000). “Guidance and Justification in Particularistic Ethics”. *Bioethics*, 14:287-309.

Korsgaard, C. (1986). “The Right to Lie: Kant on Dealing with Evil”. *Philosophy & Public Affairs*, 15:325-349.

Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge: Cambridge university Press.

Levi, I. (1980). *The Enterprise of Knowledge*. Cambridge, Massachusetts: The MIT Press.

Levi, I. (1981). "Should Bayesians sometimes neglect base rates?" *The Behavioral and Brain Sciences*, 4: 342-3.

Levi, I. (1985) “Illusions about Uncertainty,” *British Journal for the Philosophy of Science*, Vol. 36: 331-340.

Levi, I. 1997. *The Covenant of Reason*. Cambridge: Cambridge University Press.

Levi, I. (2001). ”Introduction to D. Ellsberg, *Risk, Ambiguity and Decision*”, New York and London Garland Publishing, ix-xxxvii. (Introduction to Ellsberg’s Ph.D dissertation of 1962).

Levi, I. (2004) “The Logic of Consistency and the Logic of Truth”. *Dialectica*, 58: 461-82

Lichtenstein, S. & Slovic, P. (2006). *The Construction of Preference*. Cambridge: Cambridge University Press.

Noddings, N. (1984). *Caring: A Feminine Approach*. Berkeley: University of California Press.

O’Neill, O. (1989). *Constructions of Reason*. Cambridge: Cambridge University Press.

Ramsey, F. P. (1926). Truth and probability. (In D.H. Mellor (Ed), Ramsey, F. P. (1990). *Philosophical papers*, (1990), Cambridge: Cambridge University Press).

Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Sahlin, N.-E. (1990). *The Philosophy of F. P. Ramsey*. Cambridge: Cambridge University Press.

Sahlin, N.-E. (1985). "Three decision rules for generalized probability representations". *The Behavioral and Brain Sciences*, 4:751-753.

Sahlin, N.-E. (1993). "Worthy of choice". *Theoria* 65:178-91.

Sahlin, N.-E. (2003). *Frank Ramsey*, www.fil.lu.se/sahlin/ramsey, Lund.

Sahlin, N.-E. and Vareman, N. (2008). "Three types of decision theory". *Reasoning, Rationality and Probability*, Galavotti, M. C., Scazzieri, R. and Suppes, P. (eds.), Stanford: CSLI Publications.

Sahlin, N.-E., Wallin, A., and Persson, J. (2008). "Decision science: From Ramsey to dual process theory". Forthcoming *Synthese*.

Savage, L. J. (1954/1972). *The Foundations of Statistics*. New York: Dover Publications, Inc.

Seidenfeld, T. (1988). "Decision theory without 'independence' or without 'ordering'". *Economics and Philosophy*, 4:267-90.

Slote, M. (2007). *The Ethics of Care and Empathy*. New York, Routledge.

Slovic, P. and Tversky, A. (1974). "Who Accepts Savage's Axiom?" *Behavioral Science*, 19:368-73.

Slovic, P. *The Perception of Risk*, London: Earthscan.

Slovic, P. (2007). "'If I look at the mass I will never act': Psychic numbing and genocide". *Judgment and Decision Making*, 2:79-95.

Wallin, A. (2003). *Explaining everyday problem solving*. Lund: Lund University Cognitive Studies 99.

Väyrynen, P. (2006). "Ethical Theories and Moral Guidance," *Utilitas*, 18:291-309

Wason, P. C. (1966). "Reasoning about a rule". *Quarterly Journal of Experimental Psychology*, 20:273-281.