RISK, SOCIETY AND POLICY SERIES
Edited by Ragnar E. Löfstedt

The Ethics of Technological Risk

Edited by
Lotte Asveld and Sabine Roeser

earthscan
publishing for a sustainable future

London • Sterling, VA

First published by Earthscan in the UK and USA in 2009

Copyright © Lotte Asveld and Sabine Roeser, 2009

All rights reserved

ISBN: 978-1-84407-638-3

Typeset by 4word Ltd, Bristol
Printed and bound in the UK by TJ International Ltd, Padstow
Cover design by Yvonne Booth

For a full list of publications please contact:

Earthscan
Dunstan House
14a St Cross St
London EC1N 8XA, UK
Tel: +44 (0)20 7841 1930
Fax: +44 (0)20 7242 1474
Email: earthinfo@earthscan.co.uk
Web: **www.earthscan.co.uk**

22883 Quicksilver Drive, Sterling, VA 20166-2012, USA

Earthscan publishes in association with the International Institute for Environment and Development

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

The ethics of technological risk/edited by Lotte Asveld and Sabine Roeser.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84407-638-3 (hardback)
1. Technology–Risk assessment. 2. Ethics. I. Asveld, Lotte. II. Roeser, Sabine.
T174.5.E84 2008
174'.96–dc22
2008033036

The paper used for this book is FSC-certified.
FSC (the Forest Stewardship Council) is an
international network to promote responsible
management of the world's forests.

FSC
Mixed Sources
Product group from well-managed forests and other controlled sources
Cert no. SGS-COC-2482
www.fsc.org
© 1996 Forest Stewardship Council

10 Welfare Judgements and Risk

Greg Bognar

Welfare judgements

When we try to decide what to do, or we evaluate actions and policies, we are often concerned with welfare effects: how good an action or policy is for those who are affected. In such cases we attempt to make *welfare judgements*.

There are different kinds of welfare judgements. In many cases, we are interested in comparative judgements: how well off certain people are or will be compared to some other situation or compared to some other people. Such judgements are often elliptical – cast in non-comparative language. Other times we may be interested simply in whether something is good or bad for some person without comparing it to something else or someone else's lot. Welfare judgements may also concern the past or the present – when for instance we evaluate how well a person's life is going, or has been going, for that person – but they often concern the future: in such cases, we ask how well off people may be, given that this or that action or policy is chosen. These are prospective welfare judgements.

In recent years, philosophers have expended considerable effort clarifying and proposing theories of welfare.[1] In contrast, they have devoted much less attention to the problem of welfare judgements. One might think that given a theory of welfare, it is a straightforward task to evaluate actions, policies, institutions and the like. Thus, we do not need a separate model of welfare judgements – having the best theory of welfare at hand, we can make all the welfare judgements we want to make. But even though theories of welfare and models of welfare judgements are not unrelated, they are different. A theory of welfare is a view of that in virtue of which certain things are good (or bad) for people; it gives an account of what we mean when we say that a person is well off or her life is going well for her. It provides the basis on which we can evaluate welfare. A model of welfare judgements, on the other hand, tells us how we can make these evaluations: it answers questions about the scope and limits of the judgements we can make on the basis of our account of well-being.

Another way in which theories of welfare and models of welfare judgements differ is this: a theory of welfare is usually understood as an account of what is *intrinsically* good for a person; it is not concerned with what is good only as the means (for achieving what is intrinsically good) for a person. But when our objective is to promote welfare, we often have to deliberate about means as well as ends. Even though such deliberation makes essential reference to what is intrinsically good for the people we are concerned with, we need to take into account how different means might contribute to what is intrinsically good for them. In other words, we need to deliberate about what is merely instrumentally good.

As an example, suppose you accept a simple hedonist account on which a person's well-being consists of that person's having some conscious mental state. Then facts about that mental state are the truth-makers of claims about welfare: you are well off just in case you have this mental state. But your theory of welfare is not going to be sufficient to evaluate all welfare claims. How can we tell what your level of well-being is? How can we tell how that level compares to your level of well-being at other times or to the level of well-being of other people? Can we compare not only levels, but also units of welfare, such that we can determine how much better off you are than another person (or how much better off you are at one time than at another time)? In order to determine how to answer such questions, and whether they can be answered on your theory at all, we need a model of welfare judgements.[2]

As a matter of fact, there is one model of welfare judgements which is often implicit in discussions of well-being as well as in our ordinary, everyday practice of making welfare judgements. On this model, very roughly, welfare judgements are made and justified by appealing to what people in more privileged epistemic and cognitive circumstances would want, desire, prefer, seek or choose for themselves. That is, welfare judgements involve idealization. This model is most naturally connected to informed desire or full-information accounts of well-being.[3] It is not incompatible, however, with any theory of well-being which allows that the desires or preferences formed in these circumstances can serve as *indicators* of a person's well-being, hence welfare judgements can be made in terms of them – even if there may be other, perhaps often more adequate, methods for making those judgements, or the judgements made on this model track the truth of welfare claims less than perfectly reliably.[4]

I shall call this sort of model of welfare judgements the *ideal advisor model*. One of the considerations in its favour is that it seems to give a reasonably adequate picture of how people ordinarily make welfare judgements. When giving advice to someone, we often appeal to what that person would want if she was in a better position to decide what to do; normally, we give more weight to the advice of those who are better informed, more experienced, or make a sound case for their opinion; and we might prefer not to make decisions when we are tired, depressed, inattentive or lack information.

A further consideration in favour of the ideal advisor model is that it is relatively uncontroversial insofar as it does not stipulate that there are pursuits or goods in terms of which welfare judgements should be made. By starting from people's actual desires and preferences and determining how these would change in more ideal conditions, the model does not include *substantive judgements* – judgements about the content of desires and preferences. All we need to rely on in order to evaluate people's well-being are rationality and facts.

Nevertheless, I will argue that neither of these considerations hold for some kinds of welfare judgements. The ideal advisor model fails to capture an important component of what we do when we make such judgements. It also fails as their philosophical model, since these judgements are inevitably substantive. The root of these problems is that the assessment of risks plays a central role in these kinds of welfare judgements.

The welfare judgements I have in mind are *comparative* and *prospective*. Such judgements are perhaps the most important kind there is: people make them all the time when they want to decide what may be best for them or for others. Politicians and other public officials make them when they formulate policies. Economists and other social scientists make them when they evaluate those policies.

I begin by characterizing the version of the ideal advisor model which seems to be the most plausible. If this version is not the best formulation, it can at least serve as a starting point for my arguments. The last section explains why I do not think the model can be reformulated to meet my objections, and it concludes by discussing the general implications of my argument.

The ideal advisor model

On the ideal advisor model, judgements about a person's well-being are made in terms of the preferences that person would have were she in ideal conditions for forming her preferences.[5] There are, of course, other possible models which make use of the same idea, depending on how the ideal conditions for forming preferences are specified. For instance, one version may hold that a person is in ideal conditions for forming preferences if and only if she is in Buddhist meditation. What sets the ideal advisor model apart is that its ideal conditions are *epistemic* and *cognitive*: on the one hand, the person in ideal conditions for forming preferences is informed about her circumstances, range of options and the possible consequences of her choices; on the other hand, when she forms her preferences, she does not make any mistakes of representation of fact or errors of reasoning, and she is free of distorting psychological influences such as depression, anxiety, stress and the like. In short, the person in ideal conditions is informed and rational in some appropriate sense.

The ideal advisor model involves specifying counterfactuals about preferences. It takes the person's actual preferences and determines how they would change if the person was given information about her situation, provided that she does not make any mistake of reasoning and avoids other sorts of cognitive error. It is usually assumed – and I will also assume this – that the counterfactuals about the person's preference changes can be evaluated on some best theory for the evaluation of counterfactuals, whatever that theory is.

To save words, I will say that the preferences of a person in ideal conditions for forming preferences are the preferences of the 'ideal advisor' of that person.

Of course, distinguishing between the preferences of the 'actual person' and the preferences of her ideal advisor is no more than a useful metaphor that makes the exposition easier. Both are the preferences of the same person: the former are her actual preferences, the latter are the preferences she would form were she informed and rational in the appropriate sense.[6]

What is the appropriate sense? Since the epistemic and cognitive conditions of ideal preferences can be interpreted in different ways, the ideal advisor model has further sub-versions. On the interpretation I will be working with, the person in ideal conditions for forming preferences is *fully informed* and *ideally rational*. Let us consider these conditions in turn.

'Being fully informed' is usually understood to require that *all relevant* information is available to ideal advisors. Any piece of information is relevant which could make a difference to the preferences of the person, and all such information should be made available, since the recommendations based on the preferences of the person's ideal advisor would have less normative force if she was to work with limited information only.

Being fully informed, however, cannot mean that ideal advisors are omniscient. They cannot have certitude of what *will* happen given their actual person's choice; they only know what *is likely to* happen, given that choice. This restriction is necessary because if the model assumed that ideal advisors are omniscient, then it would involve counterfactuals which *in principle* cannot be evaluated. In that case, the ideal advisor model would be useless for making prospective welfare judgements (even though it may continue to be useful for making welfare judgements which concern the present or the past). Of course, it is true that we hardly ever have all the information to determine what we would prefer if we were fully informed and ideally rational, even if no prospective judgements are involved. But the idea is that as long as the relevant information is in principle available, we can make rough-and-ready judgements and use the model as a heuristic device. If, however, the model was such that in many cases we could not even in principle evaluate the counterfactuals about preference changes, there would seem to be a fatal flaw in it. Since many of our welfare judgements are indeed prospective, the restriction is inevitable.[7]

Therefore, an ideal advisor knows what options are open to her actual counterpart, the relevant features of the choice situation and the objective probabilities with which the possible outcomes might obtain – and, since she is ideally rational, she forms and handles subjective probabilities appropriately when objective probabilities of some options cannot be obtained.

Consider now the cognitive condition. The ideal advisor model treats ideal advisors as ideally rational; but it is controversial what rationality is and what it is to be ideally rational. For ideal advisors, rationality may consist of forming rational preferences, and representing and processing information appropriately. Or it may also consist of some further cognitive capacities. I propose therefore to make the following distinction. The cognitive capacities of ideal advisors include that:

(a) they form their preferences according to the canons of a fully developed theory of rational choice;

(b) in addition, they have further cognitive capacities.

By a ‘fully developed theory of rational choice’, I mean a formal theory that tells rational agents how to order their preferences under conditions of certainty, uncertainty and risk. This theory also involves rules for assigning to and updating the probabilities of uncertain options. I will call this complete theory of rational choice *R* for short. Needless to say, we do not now have such a fully developed theory, but rather we have a number of competing theories. Nevertheless, an intuitive idea of what such a theory would look like in broad outline is this: *R* tells a rational agent how to form her preferences with a view to maximize their satisfaction in decision problems, including problems in which her choice may be influenced by states of nature or the consequences of the choices of other agents, and it tells her only this much. In contrast, by ‘further cognitive capacities’ I mean cognitive capacities that are not part of that theory, even though they are necessary for an agent to have in order to be able to employ that theory. These further cognitive capacities are needed by the agent to be able to understand her situation – to describe and represent the options and possible strategies, the influences of the choices of other agents and so on. As it were, (a) enables a rational agent to make a choice, while (b) enables an agent to understand what is involved in making that choice.

Therefore, we can think of the difference between (a) and (b) this way. The former says that ideal advisors follow the norms of rationality, hence their preferences must be formally representable on theory *R*. The latter describes what capacities an agent must have to count as rational – what it takes to be able to follow the norms of rationality and to have preferences representable by *R*.

Of course, we do not make welfare judgements in exactly these terms. But this model is intended to capture the central features of what we do when we advise ourselves or others: we appeal to having the relevant information (insofar as we can have it), processing that information correctly and weighing the options on the basis of that information. Also, more needs to be done to work out the details of the model: we would have to specify more precisely what relevant information is, say more about the theory of rational choice used in the model, and spell out the details of component (b) in the cognitive condition.

In what follows, however, I am going to bracket (b).[8] In order to make my argument against the ideal advisor model of welfare judgements, all I need to suppose is that ideal advisors are ideally rational in the sense given by (a) – that is, they follow the norms, and their preferences satisfy the axioms, of *R*.

A COUNTERARGUMENT

Consider the following example. I am faced with the choice of what career to pursue in my life. For simplicity, I assume that only my success in my chosen career determines how well my life goes. Now suppose that due to my circumstances, inclinations and talents, the two relevant options open to me are becoming a philosopher and becoming a concert pianist. In order to decide which of these would be better for me, I turn to my ideal advisor. My ideal advisor knows the following. I have talent for both pursuing an academic career in philosophy and a career in the performing arts as a pianist. But he also knows that my talent for philosophy is somewhat modest: I can become a reasonably successful, average philosopher, and therefore have a reasonably good life. If, on the other hand, I pursue a career in music, I have the ability to become an exceptionally good pianist and have an immensely rewarding life.

There is, however, a problem. If I do decide to pursue the career in music, there is a high likelihood that I will develop rheumatoid arthritis in my fingers in a few years – which will destroy my career completely, and I will end up with a miserable life. My ideal advisor knows that the probability that I do develop this condition after a few years is in fact 0.9 – since he fulfils the epistemic condition, that is, he has full information of the possible consequences of my choice and the relevant probabilities. As I suppose now, he also uses a fully developed theory of rational choice, *R*, to form his preferences over what I should prefer and choose. In other words, he knows that the decision problem I face is the one depicted in Figure 10.1.

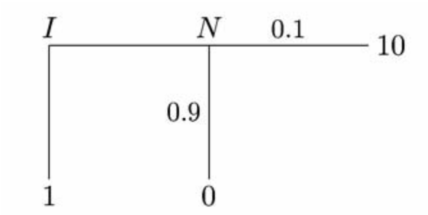

Figure 10.1 *Career choice*

The numbers 0, 1 and 10 represent my well-being: how well my life may go for me overall if I choose to become a philosopher or a concert pianist.[9] Node *I* shows my move, and node *N* shows Nature's 'move'. If I move 'down', that is, become a philosopher, my life will be alright, although not great. If, on the other hand, I move 'across', it is Nature's move – 'she' will either move down, with the consequence that I develop rheumatoid arthritis, or she will move across, in which case I do not develop the condition. There is a 0.9 probability that Nature moves down, and a 0.1 probability that she moves across. If she moves down, my career is ruined and my life will be miserable. Should she, however, move across, my life will be extraordinarily good. Note that the *expectations* of the two prospects are equal. If I move down, I realize a life with 1 'unit' of well-being. If I move across, my expectation of well-being is 0.9 × 0 + 0.1 × 10 = 1 as well.

The preference my ideal advisor forms over these prospects determines which one is likely to be better for me – whether it is the sure prospect I can choose by moving down (that is, by becoming a philosopher), or the lottery prospect I can choose by moving across (that is, by trying to become a concert pianist). So what will his recommendation be?

Now, we know that my ideal advisor forms his preference based on the norms and axioms of *R*. But in order to form his preference over the prospects I am now facing, he has to use some norm or principle of *R* that tells him how to form his preferences when I have to choose between a sure and a lottery prospect with the same expected values. Thus, he needs a norm or principle that tells him what risk-attitude he should have towards these prospects – more generally, he needs what could be called a *principle of reasonable levels of risk-taking* towards well-being, which I will abbreviate as P.[10]

Let me ask the following question: will principle P be a part of a fully developed theory of rational choice *R*, and what are the consequences of its inclusion or omission for the ideal advisor model?

First, suppose that P is *not* part of *R*. In this case, my ideal advisor will not be able to form a preference in cases like my career choice. Since no principle determines which prospect he should prefer, he cannot give recommendations to the actual person. He cannot give recommendations since he cannot compare the prospects from which the actual person has to choose with *R*. And he cannot say that the prospects are equally good since their expected values are equal. That is, he is not indifferent between the prospects, since that would presuppose a principle P: that you should be indifferent between prospects with equal expected values.

It might be objected that I may already have some risk-attitude towards these prospects, and it will be taken into account as an input to the determination of my fully informed and ideally rational preferences. But when we want to assess our actual preferences, we want to assess our preferences over prospects as well. What I am asking my ideal advisor to do in this case is precisely to tell me whether my preference based on my risk-attitude would be one that I would embrace in ideal conditions, and, if not, what sort of risk-attitude I should have when forming preferences over these prospects. Consequently, if P is not part of *R*, then the ideal advisor model is underdetermined: when the actual person has to choose from risky prospects, the theory does not specify what the person would prefer were she fully informed and ideally rational.

In order to avoid this problem, it is natural to assume that P is part of *R*. Thus, in the remainder of this section, I test the hypothesis that P is part of *R* for different versions of P. I argue that if it is, then ideal advisors may form preferences such that the welfare judgements which are based on them do not seem to be correct. I show this by arguing that if P is part of *R*, then actual persons might reject, for good reasons, the recommendations based on the preferences of their ideal advisors.

I will assume, for now, that if P is part of *R*, then it can be any of three simple principles. (I will discuss the possibility of more refined principles in the next section.) P might tell rational agents to be risk-averse towards well-being. Or it might tell rational agents to be risk-neutral towards well-being. Finally, it could tell rational agents to be risk-seeking towards well-being. However, I only mention this last possibility to discard it at the outset. I suspect that it would be quite extraordinary for our ideal advisors to tell us to take risks comprehensively. It is hard to see how a principle to seek risk could be a principle of rationality. Consider again the choice I have between becoming a philosopher and pursuing a risky career as a concert pianist. Suppose now I have the talent of a genius for playing the piano. If I do not develop rheumatoid arthritis, I will not only become a great concert pianist, but I will be the greatest concert pianist of the time: a talent like me is born only once in a century. Unfortunately, I am even more likely, in this scenario, to develop the condition in my fingers. Suppose the probability of this is 0.999 now. There is, however, a very low – 0.001 – probability that I do not develop the condition, and my life will be exceptional: its value will be not 10, but 1,000. The expectations of becoming a philosopher and risking the career in music are again equal. It nonetheless seems, given the extremely low likelihood of pursuing the concert pianist career successfully, that my ideal advisor could not recommend to take that risk. But, in any case, I will give a general argument against *any* principle later.

Let me now consider the remaining two cases. Suppose, first, that principle P of *R* tells rational agents to be risk-averse towards well-being. In particular, it tells rational agents that when they are faced with a sure prospect and a lottery prospect of the same value, they should prefer the sure prospect – in short, rational agents play it safe.

I will, once again, argue through an example. In the example of the choice between becoming a philosopher or a concert pianist, the outcome of the choice of pursuing the latter was influenced by factors outside of my control – by the state that may result following a 'move' by Nature. But our choices are not influenced by states of nature only. They may also be influenced by the consequences of the choices other people make. When giving us advice, our ideal advisors must take into account these influences as well.

Look at Figure 10.2 now. In this situation, there are two individuals, *A* and *B*. I suppose *A* is female and *B* is male. The first number at the endpoints stands for the value of the outcome for *A*, and the second number stands for the value of the outcome for *B*. Thus, *A* has a choice at node *A*: she can either move down, in which case she receives 1 unit of well-being and *B* receives 0; or she can move across. If *A* moves across, there is a 0.5 probability that she will receive 0 and *B* will also receive 0, but there is also a 0.5 probability that *B* gets to make a choice: he can also move down or across. (What p and $(1 - p)$ stand for will become clear later.) If *B* moves down, *A* receives 0 and *B* receives 3 units of well-being. If, on the other hand, he moves across, there is a 0.5 probability that they both receive 2, and it is equally likely that *A* receives 2 and *B* receives 4.

I will assume that initially both *A* and *B* take the recommendations of their respective ideal advisors as authoritative – that they recognize the preferences of their ideal advisors as *reason-giving* – and they mutually know that they do. Furthermore, their ideal advisors form their preferences according to the canons of *R*, and *R* contains P: a principle that tells rational agents to be risk-averse towards well-being. What will the recommendations of the ideal advisors be?

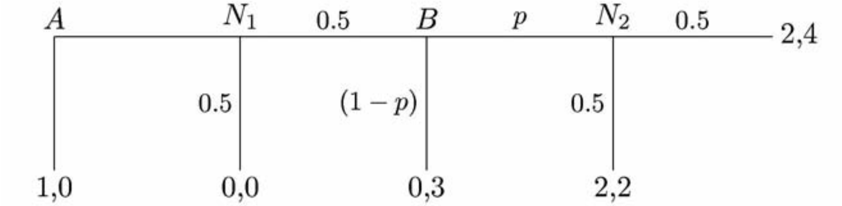

Figure 10.2 *A problem for actual persons*

Look at the situation from the perspective of *B* first. His ideal advisor reasons that if *B* moves down at his node, he will receive 3 for certain; if he moves across, he may receive either 2 or 4 with equal probabilities. The expectations of these two prospects are equal. But *B*'s ideal advisor follows P, which says that you should be risk-averse towards well-being. Hence, if *B* ever gets a chance to make a move, he should move down.

Consider *A* now, who will definitely have a chance to make a move. She can either move down or across. If she moves down, she will get 1 for sure. In order to find out whether she should move across, she will reason the following way:

> If I move across, Nature will either move down or across. If Nature moves down, I receive 0. If Nature moves across, *B* will make a move. He will either move down or across. If he moves down, I again receive 0. If he moves across, Nature will move again, but that move is irrelevant, since no matter what happens, I receive 2. So what I can expect if I move across partly depends on *B*. Suppose *B* moves across with probability p, and he moves down with probability $(1 - p)$. My expectation if I move across therefore is:
>
> $0.5 \times 0 + 0.5((1 - p) \times 0 + p(0.5 \times 2 + 0.5 \times 2)) = p.$
>
> So whether I should move across depends on what *B* is likely to do, whether he is willing to move across. But I know that he will take the preference of his ideal advisor as the reason for his move. And I also know that his ideal advisor forms his preference according to a principle of risk-aversion towards well-being, that is, he will prefer him to move down, should he get a chance to move. Hence I know that he would move down, that is, I know that $p = 0$. So I should move down myself.

This assumes, of course, that the preferences of the ideal advisors, as well as the reasons for those preferences, are known by the actual persons. This assumption enables us to check whether *A* and *B* can endorse the preferences of their ideal counterparts. The argument I make is that they have reasons not to. They have reasons to reject the recommendations of their ideal advisors.

To return to *B*. He realizes that if they both act in accordance with the preferences of their ideal advisors, he will never have a chance to move. But getting a chance to move by *A*'s moving across would be at least as good for him as not getting a chance to move by *A*'s moving down. That is, if only he got a chance to move – even if he *actually* could not move because Nature moved down at N_1 – he would not be worse off, and possibly he could end up being much better off. In short, he would not be worse off if *A* moved across, irrespective of what happens afterwards. And *B* starts to think now, and comes up with an idea.

Suppose *A* and *B* can communicate, and do it without costs. Then *B* can make the following offer to *A*: 'I promise to move across if I get a chance to make a move.' *B* has nothing to lose with promising this, since if *A* accepts the offer, he might end up better off, and if she rejects it, he ends up no worse off. The idea behind the offer is that by cooperating in ways not embraced by their ideal advisors, they might fare better than by strictly following their recommendations.

When *B* makes his offer, he promises that he will not act in accordance with the preference of his ideal advisor. In effect, he promises that at his move he will *not* be risk-averse towards well-being. In other words, he promises to reject the reasoning based on P. Instead of being risk-averse, he becomes risk-seeking, and he makes it the case that $p = 1$. We can think of *B*'s offer as choosing a *risk-disposition* towards well-being at the start of the choice problem: if he makes the offer, he promises to become risk-seeking, and if he declines to make the offer, he remains risk-averse. Similarly, we can think of *A*'s decision whether to accept the offer as choosing a risk-disposition which determines which way she moves at node *A*: on the one hand, if she accepts the offer, she becomes risk-seeking towards well-being and moves across; on the other hand, if she rejects the offer, she becomes risk-averse towards well-being and moves down.[11]

For the sake of the argument, assume that the agents are *transparent*, that is, their risk-dispositions are known with certainty. Of course, in many situations agents are not transparent, thus their risk-dispositions are not known with certainty. In such cases, whether *A* can accept the offer depends on how she evaluates the risk of accepting it, given her probability assessment of *B*'s risk-disposition – thus, whether she accepts the offer depends on the *degree* to which she is willing to become risk-seeking. However, at least in this case *B* has a reason to become transparent – as a way of assuring *A* that the promise of moving across at his node will be kept.

Assume also that *B*'s offer to commit himself to be a risk-seeker is *credible*: once he chooses his risk-disposition, he sticks with it, and he does move across at node *B*. Of course, prior commitments are not always credible. When the time for action comes, agents may find that they are better off breaking a prior promise. However, at least in this case, once *B* has chosen his risk-seeking disposition, he has no obvious reason to change it later. In other words, we may suppose that *B*'s risk-disposition is stable, in which case *A* can count on *B* to move across at node *B*. If risk-dispositions are less than perfectly stable, whether *A* can accept the offer depends on how she evaluates the risk of accepting it, given her probability assessment of *B*'s stability of risk-disposition – thus, whether she accepts the offer once again depends on the *degree* to which she is willing to become risk-seeking. Hence, *B* has a reason to develop a stable risk-disposition.

But should *A* accept the offer after all? If she moves down, she will get 1 for certain. If she moves across, she also expects 1, since now she knows that *B* will move across – that $p = 1$. Why would she reject the offer? One consideration is that her ideal advisor, who is ideally rational in the sense given by *R*, tells her to be risk-averse towards well-being, so she should still move down at node *A*, regardless of *B*'s promise. This consideration, however, is decisive only if she continues to believe that P is a principle of rationality or that it is relevant to deciding what she should do. But *B* now rejects P, and for sound reasons. With the offer, *A*'s situation has changed. Figure 10.3 illustrates how her original choice problem is simplified, given that the offer is made.[12]

B's offer requires that the actual persons cooperate by harmonizing their risk-dispositions towards well-being in ways not embraced by (and not open to) their ideal advisors. The offer works only if *B* gives up the recommendation based on the preferences of his ideal advisor by rejecting P, and *A* too gives up the recommendation based on the preferences of her ideal advisor, also rejecting P. The offer requires that they both become risk-seekers – that they both become irrational in light of *R*. If they do, *B* can end up better off: once *A* moves across and

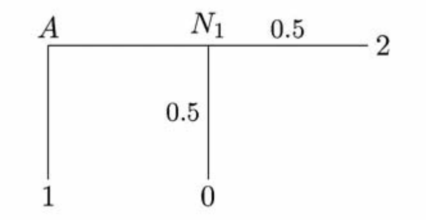

Figure 10.3 *The problem simplified*

if he gets lucky, he is guaranteed a pay-off of at least 2, and, with some further luck, 4. Moreover, *A* is no worse off, since the expectations of moving down and moving across are now equal, and she has no reason not to reject P and become a risk-seeker.

Their ideal advisors, in contrast, cannot cooperate in the same way. They cannot transform their situation in order to open up the possibility for realizing higher gains. Since ideal advisors, by definition, are 'in the grip' of their rationality, and their rationality, by hypothesis, prescribes risk-aversion towards well-being, their preferences cannot yield the recommendation to cooperate. *A* and *B* will realize this, since their ideal advisors are transparently risk-averse and their risk-dispositions are fixed.

B has a reason to reject the recommendation of his ideal advisor and make this known to *A*, because if he follows that recommendation, he forgoes benefits he might otherwise be able to obtain through cooperation. His ideal advisor, since he is ideally rational and this is known, cannot credibly commit himself to move across at node *B*. Thus, his offer would not be accepted. Actual persons therefore may be able to come to agreements which would be foreclosed to them if they were ideally rational.

A has no reason to accept the recommendation of her ideal advisor because she may fail to see why she should be risk-averse in a situation like the one depicted in Figure 10.3. She notices that there are many situations in which it is better not to follow the recommendation given by *R* (*B* is in such a situation), and she may start wondering why P should be considered a principle of rationality at all – or, if it is a principle of rationality, why a principle of rationality should determine what is likely to be better for her.

Notice that the argument is not that ideal advisors can *never* employ commitment or other mechanisms to come to advantageous agreements. Rather, the argument is that in virtue of their ideal rationality, certain mechanisms whose use can make people better off are foreclosed to them. Therefore, their preferences may fail to determine what is likely to make less than ideally rational persons better off.

Finally, let me ask what happens if principle P prescribes risk-neutrality towards well-being instead of risk-aversion. Actually, nothing changes in this case. *A* and *B* may similarly reject the recommendations of their ideal advisors. In order to see this, return to Figure 10.2. For *B*, the expectations of moving down and across at node *B* are equal, and if he adheres to the preferences of his ideal advisor, then he will, say, toss a fair coin to decide which move to make. That is, $p = 0.5$. Hence *A* will move down, because by doing so she expects 1. Once again, *B* will realize that this way he will never get to make a move, and he is thus eager to give up P and convince *A* to move across. He tells her that he will not follow the principle, attempting to make the expectations of *A* equal. If *A* listens to the recommendation of her ideal advisor, she will also toss a fair coin to decide whether to move down or across. But that is not good enough, for why jeopardize their cooperation by staying risk-neutral? She could make sure their cooperation gets a better chance to kick off if she also abandons P, and acts as a risk-seeker. She once again has reason to think that principle P does not determine what is likely to be better for her, given that there are situations when it is more beneficial to give it up – to become irrational in light of *R*. She may tell herself that sometimes it is better not to listen to what you would advise yourself to do were you ideally placed to give yourself advice.

Reasonable risks

In this section, I first consider the possibility that *R* contains some more refined principle of reasonable levels of risk-taking towards well-being. I argue that no such principle is possible, at least not within the context of rationality. I conclude by exploring some of the implications of my argument.

Of the three basic candidates for P, we found that one, risk-seeking, is implausible on its own right, and it is possible to construct situations in which the two other principles also become implausible – because actual persons may find incentives to abandon them. This means that if *R* contains any of these, actual persons may fail to take the preferences of their ideal advisors as reason-giving, since rejecting these principles can lead to outcomes which are better for them. At the same time, if *R* does not contain a principle that prescribes preference formation in risky situations, then ideal advisors cannot form preferences in these situations and the prospects remain incomparable. Either way, the ideal advisor model of welfare judgements fails to resolve which prospect is likely to be better. Hence we cannot base our prospective welfare judgements on the preferences people would form were they fully informed and ideally rational.

On the one hand, perhaps one can bite the bullet and argue that the ideal advisor model is incomplete for prospective welfare judgements. One may argue that there is a good reason for this: when we are faced with a risky situation, there is no answer to the question which prospect would be better, since a person's well-being is uncertain until the risk (or uncertainty) is resolved. Hence, the prospects are incomparable with respect to well-being.

This objection rests on an understanding of what we do when we try to make prospective welfare judgements that is different from mine. It assumes that these judgements try to determine how well lives *will* go, given that some action or policy is chosen. In contrast, I take it that prospective welfare judgements are about how well lives are *likely* to go, or can be expected to go, given that the action or policy is chosen. While it is certainly impossible to determine the answer to the former question until the risk is resolved, it should not be impossible to determine the answer to the latter question. In this respect, prospects do not seem to be incomparable. After all, when I am pondering whether I should try to be a reasonably good philosopher or an exceptionally good piano player (with a predisposition to develop rheumatoid arthritis), my complaint is not that I cannot compare these prospects – my complaint is that what risk-attitude I should have towards these prospects does not seem to be a matter resolvable by a principle of rationality. Comparing prospects is the most we are able to do, and we cannot avoid making these comparisons.

On the other hand, one could propose that principle P, in a fully developed theory of rational choice, will be much more complex. It will specify some particular *level* of risk-taking. So, for instance, it will tell me to try to become a concert pianist if the likelihood of developing rheumatoid arthritis is within tolerable limits, but choose the career in philosophy if its probability is too high. That is, the principle prescribes some rational level of risk-taking. What you should do, then, depends on the riskiness of the prospects you face.

But it is doubtful that you should have the same level of risk-taking in all situations, regardless of the context. Different attitudes towards risk seem warranted when you decide how to invest your savings for your old age and when you play a friendly poker game. So a complex principle must differentiate between different reasonable levels of risk-taking for different *objects* of preference. For example, the principle could say that you should be risk-averse when making a career choice, to ensure that your life does not turn out to be very bad. Hence I should choose to become a reasonably good, although not great, philosopher. But this complex principle could also tell me to be more risk-seeking when I am faced with the choice between staying at my current academic post or accepting a job offer from another country, where I may do my best work, but there is a fair chance that I will not be able to integrate into the academic community there, and my work will go poorly.

When we give advice in real life, we do distinguish between reasonable levels of risk-taking. We believe people should not jeopardize their health with smoking, and that they should save up money for their old age. But very often, we also advise people to take risks. We think it is a good thing to travel to other countries, to take reasonable financial risks, we admire people choosing risky professions. Hence, the proposal goes, a fully developed theory of rationality will incorporate a principle of reasonable levels of risk-taking for different objects of preference.

The problem with this proposal is twofold. First, given that there is no consensus on reasonable levels of risk-taking either in everyday situations or in philosophy, and it is controversial what these levels should be, it is hard to see where we could find the resources to formulate this complex principle. Even if disputes about what risk-attitudes are reasonable in different situations can be resolved, we end up on this proposal with a 'principle' which is merely an infinitely long conjunction associating reasonable risk-taking levels with descriptions of possible objects of preference and, perhaps, also contexts of choice. And this leads to the second problem, because it is hard to see how such a complex 'principle' could be considered a *principle*; and even if it was one, why it would be a principle of *rationality*.

Of course, one could argue that we need a more robust conception of rationality which incorporates principles of reasonable risks. But this just recreates the problem within this more robust theory of rationality. Moreover, on this proposal, prospective welfare judgements unavoidably involve substantive judgements – judgements about risks associated with the contents of preferences. If the complex principle was a principle of rationality, then the norms of rationality would appeal to objects of preference. In this case, when we evaluate the counterfactuals about a person's preference changes on the ideal advisor model in order to determine which prospect is better for the person, then at least sometimes we have to appeal to normative facts: what determines what can be expected to be better for the person is not the person's fully informed and ideally rational preferences, but, at least partly, normative facts about the objects of those preferences.

In my argument, I assumed a 'fully developed' theory of rational choice – a theory that we don't yet have, and perhaps we never will. This raises the question of how my argument relates to extant theories of rational choice, familiar from economics and game theory. Inevitably, all I can do here is to make some brief and cursory comments.

How do extant theories of rational choice deal with risk-attitudes? Typically, they take these attitudes as exogenous to the theory. That is, risk-attitudes are given either as empirical assumptions about people's reactions to risk in the context of particular goods (for instance, money), or they are incorporated into the utility functions which represent the decision-maker's preferences. Putting aside certain coherence assumptions on preferences, this means that risk-attitudes are not subject to norms of rationality. Hence, extant theories of rational choice typically cannot say anything about the reasonableness of people's risk-attitudes.

Perhaps this is not a serious problem for descriptive applications. But in normative contexts, and in particular in the context of making prospective welfare judgements, we are often interested in the reasonableness of risks – as in my example of a choice between becoming a philosopher or a concert pianist. In these cases, we cannot take preferences as given, and we cannot simply make assumptions about people's risk-attitudes. If we want to further develop some extant theory of rational choice for the purposes of the ideal advisor model, we need to appeal to normative facts about the objects of preferences.

If what I have argued in this chapter is correct, then prospective welfare judgements pose a special difficulty for the ideal advisor model.[13] Since such judgements involve the assessment of risks, the model is either deficient or needs to be augmented with appeals to substantive claims about the reasonableness of risks. Such claims, in turn, at least partly depend on the pursuits or goods which are the objects of the preferences. Moreover, the ideal advisor model cannot be an adequate model of how people ordinarily make prospective welfare judgements in their everyday lives. People reason about risks, and they reason about them in substantive terms; this sort of reasoning is not captured by the model.

In recent decades, we have learned a great deal about the heuristics people use when they form preferences in risky situations as well as the mistakes they are prone to make.[14] Most of these studies focus on how people judge probabilities. From a normative perspective, the ideal advisor model may help to avoid those mistakes when reasoning about risks. But it is not sufficient to base our judgements on the ideal advisor model: in order to make prospective welfare judgements, we also need to develop substantive criteria to distinguish between reasonable and unreasonable risks.[15]

References

Anderson, E. (1993) *Value in Ethics and Economics*, Harvard University Press, Cambridge, MA

Arneson, R. J. (1999) 'Human flourishing versus desire satisfaction', *Social Philosophy and Policy*, vol 16, pp113–142

Brandt, R. B. (1979) *A Theory of the Good and the Right*, Clarendon Press, Oxford

Brandt, R. B. (1992) 'Two concepts of utility', in R. B. Brandt (ed.) *Morality, Utilitarianism, and Rights*, Cambridge University Press, Cambridge

Brink, D. O. (1989) *Moral Realism and the Foundations of Ethics*, Cambridge University Press, Cambridge

Cowen, T. (1993) 'The scope and limits of preference sovereignty', *Economics and Philosophy*, vol 9, pp253–269

Darwall, S. (1983) *Impartial Reason*, Cornell University Press, Ithaca, NY

Darwall, S. (2002) *Welfare and Rational Care*, Princeton University Press, Princeton, NJ

Enoch, D. (2005) 'Why idealize?', *Ethics*, vol 115, pp759–787

Finnis, J. (1980) *Natural Law and Natural Rights*, Clarendon Press, Oxford

Gauthier, D. (1986) *Morals by Agreement*, Clarendon Press, Oxford

Griffin, J. (1986) *Well-Being: Its Meaning, Measurement, and Moral Importance*, Clarendon Press, Oxford

Hare, R. M. (1981) *Moral Thinking: Its Method, Levels, and Point*, Clarendon Press, Oxford

Harsanyi, J. C. (1982) 'Morality and the theory of rational behaviour', in A. Sen and B. Williams (eds) *Utilitarianism and Beyond*, Cambridge University Press, Cambridge

Kagan, S. (1992) 'The limits of well-being', *Social Philosophy and Policy*, vol 9, pp169–189

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

Loeb, D. (1995) 'Full-information theories of individual good', *Social Theory and Practice*, vol 21, pp1–30

McClennen, E. F. (1990) *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, Cambridge

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

Railton, P. (1986a) 'Facts and values', *Philosophical Topics*, vol 14, pp5–31

Railton, P. (1986b) 'Moral realism', *The Philosophical Review*, vol 95, pp163–207

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

Raz, J. (1986) *The Morality of Freedom*, Clarendon Press, Oxford

Rosati, C. S. (1995) 'Persons, perspectives, and full information accounts of the good', *Ethics*, vol 105, pp296–325

Scanlon, T. M. (1993) 'Value, desire, and the quality of life', in M. C. Nussbaum and A. Sen (eds) *The Quality of Life*, Clarendon Press, Oxford

Sobel, D. (1994) 'Full information accounts of well-being', *Ethics*, vol 104, pp784–810

Sumner, L. W. (1996) *Welfare, Happiness, and Ethics*, Clarendon Press, Oxford

NOTES

1 See, for example, Arneson (1999), Brandt (1979), Brink (1989, pp17–36), Darwall (2002), Finnis (1980, pp59–99), Griffin (1986), Kagan (1992), Parfit (1984, pp493–502), Railton (1986a), Raz (1986, pp288–320), Scanlon (1993) and Sumner (1996), among others.

2 Further questions can be raised about the *temporal unit* of welfare measurement. Suppose we want to establish how well off a person is, or how well off she is in comparison to other persons. Do we then assign a value to how well off she is for her whole life, for this very moment, or something in between? (See Brandt (1992) for a discussion of this and related problems.) These questions lead to controversial issues in metaphysics; for instance, whether persons persist through time or they are merely collections of persons at different time-slices. For a seminal discussion of such issues, see Parfit (1984).

3 For such accounts, see, for example, Brandt (1979), Darwall (1983, pp85–100), Hare (1981, pp101–106, 214–218), Harsanyi (1982), Railton (1986a) and Rawls (1971, pp407–424). For discussions, see Anderson (1993, pp129–140), Cowen (1993), Loeb (1995), Rosati (1995) and Sobel (1994).

4 For a discussion of the distinction between the two uses of idealization in value theory, see Enoch (2005).

5 Spelling out the model in terms of preference rather than, for instance, desire, allows us to make comparative welfare judgements, since preference is a comparative notion.

6 Note also that the ideal advisor's preferences range over the feasible options which are open to her non-ideal counterpart, even if that person is unaware of the availability of some of those options. This takes into account the possibilities that the person is unaware that some option is available to her or that she has not formed a preference over some of her options.

7 As a matter of fact, a similar restriction appears to be implicit in many informed desire theories of well-being. For instance, in Brandt's theory, the information ideal advisors have must be part of current scientific knowledge, available through inductive or deductive logic, or justified on the basis of available evidence (Brandt, 1979, pp111–113). In Railton's theory, even though ideal advisors have unlimited cognitive and imaginative powers, they have to base their preferences on factual and nomological information about the person's psychological and physical constitution, capacities, circumstances and history (Railton, 1986b, pp173–174). And in Harsanyi's view, a person's 'true' preferences are those that she would form if she reasoned with the greatest possible care about the relevant facts with a view to making a rational choice (Harsanyi, 1982, p55). None of these accounts of the ideal conditions include omniscience in the sense I am using the term.

8 The objections to the cognitive condition of full-information accounts of well-being and idealization in general tend to target component (b). For references, see notes 3 and 4.

9 We can think of these numbers as indices for valuable mental states, objects of intrinsic desires, items on the list of objective goods or whatever our theory of welfare proposes for having intrinsic value.

10 The expected values of the prospects do not have to be equal: principles for reasonable levels of risk-taking are relevant even if these values are unequal. I use the simplest case for purposes of illustration.

11 In order to model the offer, we can insert a node B_0 before node A in Figure 10.2, representing B's making the offer or declining to make the offer. The subsequent branches are the same on both branches leading from this node. The difference is in

the preferences that *B* forms over the prospects at node *B*, given the choice of his risk-disposition at the initial node.

12 Figure 10.3 shows only *A*'s perspective of the choice problem – assuming that *B* is transparent and his risk-disposition is stable – with the probabilities and payoffs relevant to deciding whether she should accept the offer (move across) or reject it (move down).

13 At least, they pose one sort of difficulty. I have left open the question whether other problems, familiar from the literature on strategic interactions, rational intentions and paradoxes of rationality, also raise difficulties for the ideal advisor model of welfare judgements (see, for example, Parfit (1984); Gauthier (1986); McClennen (1990)). The answer to this question at least partly depends on how we further specify component (b) when we give an account of the cognitive capacities of ideal advisors.

14 See, for example, the contributions to Kahneman et al (1982).

15 Earlier versions of this paper have benefited from discussions with Geoffrey Brennan, John Broome, Campbell Brown, János Kis, Michael Smith and an anonymous referee. I would also like to thank audiences at the Australian National University, the University of Adelaide, Central European University and the Conference on Ethical Aspects of Risk at the Delft University of Technology for useful comments.