College *of* Arts *and* Letters
Notre Dame Philosophical Reviews

UNIVERSITY OF NOTRE DAME

2005.08.11

Author ▼

Paul Weirich

Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances

Published: August 11, 2005

Paul Weirich, *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*, Oxford University Press, 2004, 278pp, $49.95 (hbk), ISBN 019517125X.

Reviewed by Adam Morton, University of Alberta

Circumstances are always nonideal and human beings are always very nonideal, so theories of ideal rationality are of limited use to us. They require us to do things we are not capable of, and, even more dangerously, they require us to do things we are capable of, but which will turn out badly unless we follow them with things that are too hard for us. That is why an analysis of an ideal rational agent is hard to translate into rules that we humans can take seriously. (This is similar to the reasons why "what would Jesus do?" is often a bad approach to a moral problem. Jesus could for example expose himself to temptations that the rest of us should stay well clear of.) Since we are very far short of being ideal agents, both in terms of our cognitive powers and in terms of our rational self-control, one might think that the standard account of ideal decision-making, expected utility maximization as worked out in Bayesian decision theory, would give very little advice for what individual human agents should do in particular circumstances. Paul Weirich's new book argues against this: it aims to provide decision-making rules for the nonideal case that are inspired by and in character not too different from the standard utility-maximizing principles. The book is admirably clear and full of stimulating suggestions. Anyone interested in the subject should read it. They should be warned, though, to set aside enough time as it is fairly slow going, not because it is excessively technical but because Weirich often takes some time to get to the point, and does not have a good sense of which details to spell out and which to leave to the reader to fill in. My own conclusion is that Weirich makes some extremely helpful points about the description of the nonideal situation, but has not really provided *rules* that we can follow in order to make the best decisions we can down here on earth. In fact I doubt that rules are what we should be looking for. It is not at all obvious that I am right, though: the question is important and deserves very careful consideration.

Weirich's most original and interesting idea occurs about half way through, after considerable stage-setting. When someone has made a decision we may evaluate it according to our standards of rationality. If conditions are not ideal some aspects of the decision may fail to live up to our standards in ways that only partially impugn the whole decision. So we need some scaled down evaluative concepts. He distinguishes between choices and decisions: a choice is a mental act, of selection of an act from a list of possibilities, while a decision is the process that led to that act. Sometimes the choice is right though the decision is flawed, as when you choose to get your money out of the stock market because of an ominous horoscope although you also have much evidence that the market is about to crash. So a choice can be correct though as a decision it is wrong, and then as a matter of bad luck it can turn out badly, as objectively the wrong thing to have done. Decisions are based on the agent's beliefs, and are only as good as them. So a decision can be good relative to bad beliefs, for example beliefs formed by irrational means. What an agent should decide or should have decided, as Weirich uses "should", is what is maximally rational given her possibly mistaken beliefs. "Should" goes with "recommended": the recommended decision is the one the agent should make given her mistakes. We may however "rate" the recommended decision as irrational.

Weirich can now recommend, in his terms, some rules for making the decisions, though not necessarily the choices, that one should. There are two central rules:

1) Decide in a way that maximizes utility after correcting relevant corrigible mistakes.

2) Decide in a way that maximizes utility after a reasonable effort to correct relevant corrigible mistakes.

1) applies to agents who make mistakes but have few constraints on their capacity to correct them. Facing a decision such an agent should consider her body of beliefs and her preliminary evaluations of outcomes, check the way they were acquired, correct those that she is capable of correcting, and then use a full maximization procedure. There are several kinds of mistakes such an agent cannot correct. One is mistakes that lead to beliefs or desires that the agent can now see to be reasonable: Weirich's example is preferring train travel out of an irrational fear of flying and then discovering the real comforts of trains. Another is mistakes that have altered the agent's options in irreversible ways. Another consists in bad habits that the agent is powerless to affect.

2) applies to agents who make mistakes and also have limited cognitive powers. This includes the case of agents who can summon immense cognitive power but possibly at an immense cost that may not always be worth paying. It requires agents to correct all the mistakes that it makes sense for them to correct, given their capacities and the costs of correcting them.

These are all the adaptation standard utility theory needs for the non-ideal case! They lead to a

principle of acceptability: A rational decision maximizes utility with respect to circumstances in which unacceptable mistakes are corrected.

Weirich does not give any precise definition of when a mistake is uncorrectable, absolutely or in terms of its costs. I imagine he sensibly thinks that this is not a topic for precise definitions. The principle of acceptability is deceptively simple; when unpacked in terms of the layers of definition behind it, it makes an interesting substantive claim about nonideal rationality. It says, roughly, that one should follow a decision-making procedure that corrects as many deviations from full standard utility maximization as one is capable of and can afford. This is not a trivial suggestion. Consider three ways in which it might be wrong.

First, the decision costs might be structured in such a way that the process of maximizing expected utility, even leaving uncorrected deviations from it that would be expensive to fix, would give worse results than some other procedure. This is the territory in which satisficing lives. Weirich's general line on satisficing is that one should incorporate the costs of making the decision into one's calculations. (There are cases in which maximization is structurally impossible, and in these cases he sees satisficing as a generalization of maximization.) This presumably means that the agent should consider the acts of gathering the information and making the calculations involved in maximization as subject to the principle of acceptability, so that a maximization argument might show that a satisficing shortcut is the best course to take. Suppose that this is correct. Maximization has then handed over the actual executive power to another procedure. This could happen on a larger scale, with a variety of other procedures, over a longer period of time. Perhaps it would be rational early in life to make a maximizing meta-decision never again to maximize.

Second, the agent might not be capable of carrying out Weirich's procedure correctly, identifying with any success or accuracy which mistakes are correctable and what the costs of correcting them are. Weirich does not discuss this possibility, but I suspect he would think that such an agent is incurably irrational. I think such an agent looks a lot like you and me.

Third, the decision in question might be part of a sequence of decisions, such that the agent can now see that the best course is to choose A now and then B at a later stage, when in fact at that later stage she will disastrously not choose B. Not choosing B could be an easily correctable mistake, but one the agent will not correct. I'm not sure how Weirich would handle this kind of case, either in the case in which the agent can anticipate the problem or that in which she cannot.

These may or may not present insuperable problems for the principle of acceptability. In following chapters Weirich shows that the principle has some attractive consequences. He shows that it allows for decision-making dilemmas, in which whatever an agent does her act will be irrational, though a rational decision is possible for her. And he makes interesting connections with situations, like the Newcomb problem, where it is rational to acquire a disposition to follow a pattern of decision each instance of which may be irrational. In some such cases the decision that accords with the disposition may be rational although the act it leads to is not. In the final chapters of the book Weirich applies the same framework to game theory. He first generalizes the point of view I have been describing to cases where an agent's choices change the structure of the choice situation, requiring

that agents choose in such a way that their subsequent acts are in harmony with the values that result from them. He then uses this to provide a justification for taking Nash equilibria as solutions to games in cases where the reasons for this conventional assumption are usually rather obscure. Readers interested in solution concepts in game theory, in dynamic choice problems, and, especially, the connections between them will find a lot of useful material here. I think it is possible that the material in the last three chapters could be used to defend Weirich's point of view against worries arising from cases of the three kinds I described above. But I do not see well enough how the arguments would go.

If there are rules for the decisions of nonideal agents then they will most likely run along the lines Weirich describes. So it is important to test his theory: if it fails then most likely there are no rules. Would that be a disaster? I think not: contrast rational decision with rational belief. Few philosophers believe that there are rules that tell us what to believe. There are principles that tell us how to evaluate the strength of evidence, and less definite principles that tell us how to gather evidence, and even less definite principles that tell us how to distinguish serious hypotheses from those not worth considering. All these are valuable and non-trivial, but they don't add up to rules for belief. The really fundamental question that Weirich brings to the foreground is whether decision is the same. If we want to articulate principles for judging the rationality of agents, taking into account their limitations and the difficulties of the situations they find themselves in, should these principles take the form of imperatives?

ISSN: 1538 - 1617
College of Arts and Letters
Notre Dame, IN 46556