Economics and Philosophy, **18** (2002) 303–328

DECISION THEORY AND THE RATIONALITY OF FURTHER DELIBERATION

Igor Douven
Erasmus University Rotterdam

Abstract

Bayesian decision theory operates under the fiction that in any decision-making situation the agent is simply given the options from which he is to choose. It thereby sets aside some characteristics of the decision-making situation that are pre-analytically of vital concern to the verdict on the agent's eventual decision. In this paper it is shown that and how these characteristics can be accommodated within a still recognizably Bayesian account of rational agency.

Only rarely when we come to face a decision problem are we presented with a full range of options right at the outset. More frequently in such situations searching for options is an integral and crucial part of the process of deliberation that lies before us. We have strong intuitions about the rational conduct of this part of the task as well as about the relevance thereof to the rationality of our eventual decision. For instance, we feel strongly that under certain circumstances it would be more rational to deliberate further on other courses of action that may be open to us in the situation we are in than to choose straight away one of the options that have already occurred to us. However, what is arguably the only well-developed formal decision theory to date, namely, Bayesian decision theory (hereafter BDT),[1] seems incapable of accommodating

An earlier version of this paper was presented at the Erasmus Institute for Philosophy and Economics. I am greatly indebted to Martin van Hees, the discussant on that occasion, for his extremely insightful comments. I am also grateful to the audience for critical questions and remarks, and to Pieter Hofstra, Anita Keij, Theo Kuipers, Patrick Maher, Uskali Mäki, Roberta Muramatsu, Diederik Olders, Jan-Willem Romeyn, Mark van Atten, Jack Vromen, and two anonymous referees for valuable comments. A discussion with Jos Uffink about the subject matter of this paper was also helpful.

[1] As I use this name, it is synonymous with "expected utility theory". In particular, the "B" in "BDT" does not signify a commitment on my part to the eponymous principle of

these intuitions. As a result, the theory appears to lead to counter-intuitive verdicts in an important class of cases. The present paper aims to offer a new, refined decision theory that does do justice to the aforementioned intuitions and also yields the intuitively correct result in cases in which BDT fails to do so.

1. A PROBLEM FOR BAYESIAN DECISION THEORY

According to BDT, the rational agent chooses from his options the/an act that maximizes expected utility given his personal probabilities and his utilities, where the expected utility of an act a, $EU(a)$, is defined thus:

Definition 1.1

$$EU(a) \stackrel{\text{def}}{=} \sum_{w \in W} p(w)u(a.w).$$

Here "W" denotes the collection of all possible states the world can be in; "$a.w$" denotes the result of choosing a in w (often referred to as the outcome of a in w). Both $p(\cdot)$ and $u(\cdot)$ are total functions on W that represent the agent's probabilities and utilities, respectively. So, $p(w)$ measures the agent's probability that w is the actual world-state, $u(w)$ measures his degree of utility for w, or, put more colloquially, the intensity with which he desires w to be the actual world-state.[2]

Friends and foes alike have praised this theory for its elegance and simplicity (see, for instance, Lewis (1981, p. 5) and Simon (1983, p. 13)). At the same time, it is widely recognized that BDT operates under various idealizing assumptions and that, therefore, it is applicable, at least in any literal fashion, to what at most is an extremely restricted class of decision-making situations. This reliance on idealizing assumptions does, of course, not automatically invalidate BDT; if it did, then presumably most (and perhaps even all) scientific theories would be invalidated, too.[3] But, I claim, in the case of BDT there is (at least) one particular idealization involved that is, in fact, quite crippling.

As part of their critique of BDT, March and Simon (1958, p. 137) highlight the following feature of the theory:

Bayesianism according to which Bayes's rule is the only rational rule for belief change. On the contrary, I repudiate this principle; see Douven (1999, 2002a, 2002b). (I do take Bayes's rule to be *a* rational rule for belief change, however, see Douven (2002c).)

[2] For detailed expositions of the basic machinery of BDT see, e.g., Savage (1954), Luce and Raiffa (1957), Chernoff and Moses (1959), Ferguson (1967), Jeffrey (1983), Resnik (1987), Maher (1993) and Kaplan (1996).

[3] See on the issue of idealization in the sciences Krajewski (1977), Nowak (1980), Cartwright (1983, 1999), Kuipers (2000, 2001). Specifically with respect to idealization in economics see the papers in Hamminga and De Marchi (1994); also Mäki (1992) and Nelson (2001).

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

> When we first encounter [the Bayesian agent] in the decision-making situation, he already has laid out before him the whole set of alternatives from which he will choose his action. This set of alternatives is simply ''given''; the theory does not tell how it is obtained.

Bayesians tend to be silent on what, according to them, justifies this feature. It is most unlikely that they believe that in reality the agent generally *is* confronted outright with, to repeat March and Simon, ''the whole set of alternatives from which he will choose his action''. For it should be, and I think it largely is, fairly uncontroversial that such cases are rare;[4] deliberating about what actions are up to one is, in general, a key aspect of the practice of decision making (see Simon, 1983, p. 22). Probably, then, the feature must be regarded as an idealization: just take the set of options from which the agent makes his choice to be given, and pretend that the question of how the set was obtained by him, as indeed any other question pertaining to that set (other than questions pertaining to the expected utilities of the options in it), are immaterial to the rationality of the agent's choice.

To be sure, the idealization might be harmless (and even useful); it would be harmless if, in spite of it (and useful, if, thanks to it), we were still mostly able to derive correct results from BDT (or, if that makes sense for this theory, at least approximately correct results). But there is ample reason to believe that this is not the case. Indeed it seems that, due to this idealization, BDT can go badly wrong in decision-making situations of the most ordinary type.

This is most easily exhibited with the aid of an example. The following specifies a decision-making situation of an utterly normal sort: you face the problem of deciding what to do this evening. Until now you have hardly thought about what your options for the evening are. You have been able to think up some options, but none of them really appeals to you. However, you still have all afternoon to deliberate about alternatives and you are confident that, if you do so, you will come up with some options that are decisively better than the ones presently before you. Clearly, you might have better things to do this afternoon than deliberating about other ways of spending the evening, or you might find such further deliberation too much trouble, but, in fact, neither is the case.

Now suppose that, all the foregoing notwithstanding, you make your choice for the evening instantaneously. Then, pre-theoretically, your

[4] Games of chance may be such rare cases. This is no coincidence; as Hacking (1975, p. 63) notes, one of the guiding ideas in the development of decision theory was that ''games of chance [can] serve as models for other problems about form of decision under uncertainty''. But it is evident that many decision problems fit the model of a game of chance very badly.

choice appears irrational, regardless of which particular option you choose; given the characteristics of the situation, you are simply making an overhasty decision. However, it will be noted that, as it stands, there is nothing in BDT that could prevent you or anyone else from taking the options before you now as the "given" ones in this instance. Thus, if you choose the option that maximizes expected utility relative to this set, then, according to BDT, your choice is rational – contrary to our intuitions. Quite evidently, BDT can go wrong in this way in any situation in which we are not presented from the beginning with a full set of options, that is to say, it can go wrong in what arguably is the vast majority of decision-making situations we encounter in reality.

What we can learn from the example just presented, I think, is that if we bluntly equate rational choice with expected utility maximization relative to a given set of options, where whatever set of options the agent happens to choose from may qualify as given (in the Bayesian sense), we completely neglect certain aspects of the procedure of deliberation that really are of primary importance to our verdict about the agent's decision (see in the same vein Laville, 2000). A theory imploring us to pretend that these aspects are not important at all must then ineluctably go wrong, or so it seems.[5]

But, putting the problem thus might already seem to suggest in what direction a remedy is to be sought: in any case we should *not* say that any set of options the agent happens to choose from can count as given in the relevant sense. What we should say instead, though, is not immediately clear. I will consider three candidate answers to this question. I will also consider a more radical response to our problem. None of these responses, I argue, provides us with a satisfactory solution. My own solution will consist in the decision theory to be presented in Section 3. As the reader will see, this theory need not invoke the notion of a given set of acts in the first place.

Going higher-order. By way of a first answer, consider Schick's (1997, pp. 9f.) suggestion that for a collection of options to count as given the agent should have *decided* to make his choice from just that collection. If we follow this suggestion, then, since first, the decision to choose from a

[5] The problem for BDT exhibited here seems to be a decision-theoretic cousin of the problem of epistemic indolence presented and canvassed in Foley and Fumerton (1982). The epistemic problem is this: many hold that in order to rationally believe a proposition it is sufficient that one has evidence confirming that proposition ("confirming" here meaning that the evidence makes the proposition highly probable, and not just more probable than it initially was). But what if this evidence is radically incomplete and one is known to have been indolent with regards to gathering evidence? Under such circumstances it would seem wrong to say one is rational in believing the proposition, just as in our example it seems wrong to say that your instantaneous choice is rational.

particular collection of first-order acts seems itself amenable to evaluation by BDT, and second, it seems arguable that any first-order decision that is dependent on an irrational second-order decision is itself irrational, it is no longer implicated that if you decide to choose from the rather unattractive options for the evening you have considered so far, and you choose the best of those, your first-order choice is rational by Bayesian standards. In fact, it seems likely that, the situation being as it is, the second-order decision and, hence, the first-order decision as well, will be irrational according to BDT.[6]

However, this line of reasoning faces an immediate difficulty. Analytic philosophers have become accustomed to the fact that when, in response to certain theoretical difficulties a levels distinction is introduced, the problems that thereby get solved at level n tend to reemerge at the next higher level $n + 1$. And it seems that in the present case this would, in fact, occur. After all, a second-order decision will once again involve a decision, this time a third-order decision, to choose from a particular collection of second-order acts. By a similar principle, the second-order decision will be rational only if the third-order decision is. To evaluate the latter decision by BDT, a fourth-order decision seems to be required, and so on. In short, it appears that an appeal to a second-order application of BDT in this fashion will initiate a regress of higher and higher order applications of BDT. See on this and similar regress problems that may arise for BDT, Savage (1954, p. 30), Raiffa (1968, p. 266), Johansen (1977, p. 144), Elster (1983, pp. 17f.), Resnik (1987, pp. 11f.), Mongin and Walliser (1988), and Smith (1991). It is currently unclear whether these problems constitute decisive arguments against BDT. Some, for instance, Mongin and Walliser, believe they do. Resnik (1987), on the other hand, thinks it may be possible to solve such problems by relying on what he calls policies informing us when to deliberate and when not. However, it is evident that in relation to such policies questions may arise – such as, *How to decide which policy to use?* or *How to decide when to reassess a policy?* – that may themselves give rise to regress problems. Resnik (p. 12), in effect, admits as much. I suggest

[6] The present understanding of a second-order decision as a decision concerning the modelling of a first-order decision-making situation should be distinguished from Sunstein and Ullmann-Margalit's (1999) understanding of the same term. On their use of the term (p. 7), "[s]econd-order decisions . . . involve the strategies that people use in order to avoid getting into an ordinary [i.e., first-order] decision-making situation in the first place". Such strategies may include delegating responsibilities and adopting routines or rules (e.g., a teacher might adopt the rule of letting a student pass an exam dependent on the outcome of a coin flip). For present purposes, second-order decision making of the kind canvassed by Sunstein and Ullmann-Margalit can be set aside, since the problem for BDT presented in the text applies with equal force to it (is it rational to settle for a particular strategy even if many alternative strategies have been left unconsidered?).

that, so long as this point has not been clarified, we explore alternative responses to our problem.

Making the set of acts logically exhaustive. So far, it has only been assumed that the acts in what was referred to as "the given set" are alternatives, that is, that they are mutually exclusive acts. The essence of Schick's proposal is that this set of alternatives is made exhaustive relative to the agent's second-order decision that the choice be made from exactly that set. Now we saw that Schick's proposal is problematic. It is not hard, however, to make the set of acts (*any* given set of acts, that is) *logically* exhaustive in the sense that the agent must, by way of logical necessity, and not by virtue of some self-imposed constraint, choose one of the acts in the set.[7] The trick, of course, is that the agent adds an act that we might simply term "some other act" to the acts that have previously occurred to him (supposing these do not already constitute a logically exhaustive set). So, if we were to demand that for a set of alternatives to qualify as "given" the alternatives must be logically exhaustive, that would not restrict BDT's scope in the least. Although that demand would imply that the rational agent always "closes off" the class of alternatives before he makes his choice, due to the trick just mentioned, this is an utterly simple task, one that can readily be accomplished by any agent in any decision-making situation.

The problem with this response, however, is that the act to be annexed is an act only in a very degenerate sense. Consequently, it will, in general, be extremely difficult, if not downright impossible, to evaluate its consequences. This is not necessarily the same as saying that it would be hard or impossible to attach a utility to such an act. Surely we can attach some *number* to it – if only by turning a sort of roulette wheel, say – and perhaps we could simply take that to be the act's utility.[8] The problem would remain, though, that there seems to be no

[7] Textbook examples of decision problems, which typically involve logically exhaustive sets of acts – "Order tuna sandwich/Do not order tuna sandwich" – may give the impression that it is a *formal* requirement of BDT that the set of alternatives be logically exhaustive (some authors also seem to be under this impression; see, e.g., Simon, 1983, p. 13 and Smith, 1991, p. 198). It is important to notice, however, that this impression is wrong. There is nothing in the way BDT is set up that prevents application of the theory to non-exhaustive sets of alternatives. Nor have I ever encountered a requirement of logical exhaustiveness in any of the more formal presentations of the theory.

[8] Perhaps – whether we really can assign utilities to such acts will depend on the exact interpretation of the concept of utility, an issue that is still very much open to debate (see, for instance, Weirich, 1986, 2001, Hansson, 1988, Kusser and Spohn, 1992, Hampton, 1994, Dreier, 1996, Niiniluoto, 1999, p. 159, and Hacking, 2001, pp. 100f, for some – sometimes very different – views on this issue). If, for instance, utilities are considered as representing the agent's feelings (as, e.g., Romer, 2000 seems to suggest), then it is evidently nonsensical to suppose that they can be determined by means of a random

method for assigning utilities to such "acts" which could be reasonably regarded as being a part or a prerequisite of a process of *rational* decision making.[9]

Adding the act of further deliberation. The idea of adding an act to the alternatives already contemplated may take a more subtle form. Instead of demanding that the rational agent always make the class of alternatives exhaustive before making his choice, BDT could demand that the agent always add further deliberation as an act to the ones already in view.[10] The act of further deliberation certainly is an act in the truest sense of the word and, thus, one might expect that it should be possible to evaluate its outcomes. However, while this suggestion goes in the right direction, it still does not take us very far.

To see why not, note that nothing in BDT secures any link between the utility of the act of further deliberation and any of the features of the acts already in view, or, indeed, any of the other intuitively relevant features of the situation the agent is in. What is more, BDT may let elements slip into the determination of that utility that pre-theoretically should *not* be taken into account. For instance, as far as this theory is concerned it may be that, because you happen to be of the incontinent type (say), making an instantaneous choice of any of the alternatives for the evening you have so far considered ranks higher on your utility scale than postponing choosing in order to search for better alternatives, even though you think it likely that such a search would be successful. But surely incontinence cannot make your immediate choice rational in any plausible sense.

Bayesians may concede that this problem can arise within BDT, but see it as an inevitable consequence of the fact that BDT is a purely

device, even in principle. (Some believe that utilities are just theoretical posits that do not stand in need of any interpretation; this view seems to underly, e.g., Ramsey's (1926) and Savage's (1954) work in decision theory and is still not uncommon, as Rabin (2000) reminds us. But aside from the difficulties generally related to instrumentalist interpretations of theoretical terms, on an instrumentalist reading of utilities, decision theory can only be used as an explanatory device, and not as a guide to decision making (see, e.g., Satz and Ferejohn, 1994), which is clearly how I want to understand decision theory in this paper.)

[9] This problem is analogous to a well-known problem in Bayesian confirmation theory, namely, that of "the utter intractability of the likelihood [of the evidence] on the catch-all" (Salmon, 1990, p. 275; see also Earman, 1992, p. 168). The catch-all hypothesis, which basically says that a hypothesis other than the ones that have been explicitly formulated and considered is true, is, of course, a hypothesis only in the sense in which our "some other" act is an act (Salmon aptly calls it "a hypothesis only in a Pickwickian sense").

[10] I am indebted to an anonymous referee for this suggestion. A somewhat similar strategy has been proposed by Schmidtz (1992, p. 446); he refers to it as "optimizing . . . of a more subtle variety" (the comparison is, of course, with optimizing as defined by standard BDT).

instrumental, ''means–ends'' theory of rationality which imposes no constraints on the agent's utilities (nor, for that matter, on his probabilities). And while such an instrumentalist conception may not be to everyone's liking,[11] the problem presented in this section would lose much of its interest if it appeared to be no more than one of the ways in which a long-recognized feature of BDT may manifest itself. Our problem does deserve special attention, however. For, whereas it may generally pose extraordinary problems to go beyond an instrumentalist approach to rationality,[12] a theory that is more-than-instrumentalist in its assessment of the rationality of further deliberation seems well within reach. The reason for this is that we simply *know*, at least in broad outline, how a decision theory ought to behave with respect to the act of further deliberation – we know, for instance, that incontinence should not affect the rationality of further deliberation – and this, in my opinion, is a claim that can only justifiably be made for few other acts. As will shortly become apparent in Section 2, this knowledge enables us to lay down quite specific conditions under which further deliberation is rationally mandated. And doing so will let us devise an explicit method of assessing, in what, arguably, is the intuitively correct way, and for any given decision-making situation, the value of further deliberation in that situation, a value that can then be used to determine the rationality of choosing the act of further deliberation in that situation (Section 3).

I do not espouse the present proposal, then, not because I think it is entirely misguided, but rather because I think we can do better.

[11] Or rather, it *is* not to everyone's liking; see Elster (1983, Chapter 1) and Maher (1993, pp. 29–33). See Eells (1982) for a defence of BDT's instrumentalism; see also van Fraassen (1989, Chapter 7).

[12] Maher's (1993, pp. 29–33) *qualified Bayesianism* is an attempt to do just that. On this account, probabilities and utilities can themselves be irrational, and only if they are rational will maximizing expected utility relative to them yield a rational choice. Unfortunately, Maher does not have any concrete proposal as to what constraints probabilities and utilities should satisfy in order for them to qualify as rational, and it seems doubtful that anyone will be able to come up with general constraints that can count on wide approval. Also, whether talk of putting further constraints on probabilities and utilities makes sense at all hinges, among other things, on how we are to interpret the notions of probability and utility, a question that, to date, still remains to be answered. (See the references given in footnote 8 on the question of the interpretation of utility. Gillies (2000) contains an excellent survey of the various interpretations of probability that have been proposed so far, as well as of the problems faced by each of these; we shall also encounter some of these interpretations in Section 4.) For example, some Bayesians hold that we come to have ''prior'' probabilities while at our mother's knee, and that our further epistemic life solely consists of ''conditionalizing'' these probabilities on the basis of whatever evidence we obtain. On this view, it is in an important sense not up to us what probabilities we have at any point in time. Accordingly, it would on this view seem to make little sense to insist that our probabilities ought to satisfy certain constraints in order for them (or for us, or our actions) to qualify as rational.

Settling for less. Partly as a response to the problem for BDT discussed in this section, a considerable number of authors have rejected the Bayesian approach to rational agency altogether.[13] Most of these authors have pinned their hopes on some version of the so-called satisficing approach to rationality, according to which rational people in general do not look for the optimal solution to their decision problems but are satisfied with one that is ''good enough''. More exactly, satisficing can be described as a two-step procedure in which the agent searches for options until he finds one whose expected utility meets or exceeds his preset level of aspiration, which he then chooses.[14] As a standard of rationality, this is evidently more responsive to human psychology than the Bayesian rule of optimizing.

Responsiveness to human psychology, and particularly human cognitive limitations, ranks high, or at any rate should rank high, among the desiderata of decision theories.[15] Acknowledging this is not to abandon the project of conceiving a normative decision theory (contrary to what is sometimes suggested). Rather, it seems that *only* a theory of rationality that attends to human psychology can have normative force. After all, the time-honoured principle that ''ought'' implies ''can'' would be totally empty if ''can'' were not to refer to what is possible for actual human beings to accomplish. A second valid and important attribute of the satisficing approach is that the rationality of continuing one's search for further options is related to the goodness of the acts one has already contemplated.

But while I am sympathetic to the satisficing approach, in my view, it is neither necessary to depart as far from BDT as the advocates of satisficing propose we do, nor – at least at the time of writing this – is it really sufficient to replace BDT with a satisficing theory, if what we are after is a (more) viable account of rationality.[16] To start with the latter, the

[13] See, for instance, March and Simon (1958), March (1978), Simon (1979, 1983), Elster (1983), Giere (1988, 1999), Herrnstein (1988) and Slote (1989), to name but a few.

[14] See, e.g., March and Simon (1958, pp. 140f.). It should be noticed that the agent's level of aspiration may vary from one decision-making situation to another. (Simon (1955, pp. 14ff.) even seems to allow that it may fluctuate during one and the same decision-making situation, but he is rather vague on this point.) Therefore, some authors prefer to describe satisficing as a three-step procedure in which, as a step preceding the ones cited in the text, the agent starts by setting his level of aspiration for the given decision-making situation; see, e.g., Pettit (1984, p. 166).

[15] See also Moser (1990, p. 6): ''[A]dequate principles of rational decision making should be psychologically realistic relative to actual human decision making''.

[16] Byron (1998) also makes a convincing case for the claim that satisficing is a philosophically unstable position. Specifically he argues that, although rational behavior may be satisficing with respect to some local goal, it must be what he calls optimific, i.e., it must optimize with respect to one's global goal, which may roughly be characterized as that one's life is going well. It should be emphasized, though, that it is decisively not

proponents of satisficing will readily agree that so far their theory of rationality has not left the programmatic state.[17] In fact, it is rather misleading to speak of a satisficing *theory* of rationality; what exist today are, at best, sketches of sketches of such a theory. (Simon's (1983, p. 23) remark that his satisficing account of rationality lacks "the beautiful formal properties of [BDT]" is best regarded as an understatement). As to the former, it may be interesting to see that, while the decision theory to be developed in the following sections stays broadly within the Bayesian framework – it can, for instance, still be regarded as an optimizing approach to rationality – it is capable of honouring what were just identified as valid insights of the satisficing approach. In particular, it will appear to do perfectly well without the idealizing assumption highlighted earlier in this section. Moreover, it will turn out that the new account meets (at least to some extent) certain further wishes of the proponents of satisficing as well.

The foregoing are the only possible responses to our problem I can think of that at least have some prima-facie plausibility. Still, it cannot be excluded that there are other and more successful ways to explicate the notion of a given set of options than the ones I have considered. Also, there may be rebuttals to my objections to the above responses that I currently fail to see. Hence, I am not in a position to claim that the theory to be developed in the remainder of this paper is the only solution to our problem, nor can I claim that it is necessarily the best. I do believe, though, that the theory offers a solution that is both mathematically elegant and psychologically plausible, and that this is sufficient to make it worth considering.

2. THE VALUE OF FURTHER DELIBERATION

We have said that from a pre-analytic standpoint, rational decision making is not just governed by the principle of expected utility maximization; factors related to the process of deliberating about options count heavily, too. But, as was seen in the previous section, BDT cannot – or, at least, not in any discernible way – deal with these latter factors. To incorporate them within a broadly Bayesian setting, I propose to define rational action not directly, but only indirectly, in terms of the expected utilities of the acts the agent considers choosing from. The intermediate concept to be utilized is that of the *value of further deliberation* for a given act a. This concept captures, for an act a, the value of deliberating further about alternative courses of action over choosing a right away.

Byron's aim (at least not in the paper referred to here) to defend BDT as the correct account of optimizing.

[17] Giere (1993, p. 176), indeed frankly admits that he and other opponents of Bayesianism have so far had very little to offer by way of an alternative theory.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

I suggest that, given some decision-making situation and given an option *a* the agent considers in that situation, whether it is worth the agent's while, and, if so, what it is worth to the agent to deliberate further instead of choosing *a* right away, should depend on the following three factors. First, it should matter how "good" or "satisfactory" the agent finds *a*, in the following sense: *ceteris paribus*, the better the option (according to the agent), the lower the value of further deliberation should be (for, it will be less urgent for the agent to come up with something better). Secondly, there is the agent's estimate of his ability to think of a better alternative for *a*: *ceteris paribus*, the more confident he is that he can come up with such an alternative, the more worth his while it should be to deliberate further instead of choosing *a* now. And thirdly, the cost of further deliberation has to be considered: *ceteris paribus*, the lower the cost of deliberating further instead of choosing *a* straight away, the greater the value of further deliberation should be for *a*.

It may be arguable that, even though what it is worth to deliberate further instead of choosing some act instantaneously *largely* depends on the aforementioned factors, certain other factors have a role in determining that value as well. Also, the factors mentioned may have been stated somewhat crudely. For example, for our judgement of what it is worth to deliberate further instead of choosing now, it seems it matters not just *how likely* we believe it is that further deliberation will allow us to make a better choice later on, but also *to what extent* the later choice may be better. However, as a first approximation, let us assume that the three previously mentioned factors are *all* the factors that are, or should be, involved in assessing what it is worth to go on deliberating further about other alternatives, and that they are involved exactly as stated. It will further be assumed that the utilities the agent attributes have a greatest lower bound; for mathematical ease, we take this to be 1 (it follows from a well-known result in decision theory that any utility function that has a greatest lower bound can be transformed into an equivalent one that has 1 as its greatest lower bound – see, for example, Ramsey (1926), Jeffrey (1983, Chapter 2), Resnik (1987, Chapter 4); also below).

We first introduce some notation. Further deliberation about other options cannot go on indefinitely (if there is time for it at all). For instance, if you are to decide what to do this evening, where what you consider choosing from includes going to the opera, and you believe that the ticket office for the opera house closes at 4 p.m., then by, say, 3.59 p.m. at the latest your mind will have to be made up. Other acts among those you consider may require your mind to be made up even earlier (or at any rate that is what you may believe), but if not, let us call 3.59 p.m. the *time limit* for further deliberation in this case. If the agent is faced by a decision problem at t, then "$\mathbf{t}$" denotes the time limit for

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

further deliberation in the given situation.[18] "CA_t" denotes the class of alternatives the agent perceives at time t; apart from the requirement that the acts in this class indeed be alternatives, the only formal requirement on CA_t is that it be non-empty. Furthermore, deliberation consumes resources that might be spent otherwise. Let "c_t" then denote the cost (in utiles) the agent attaches at t to deliberating until **t** about possible alternatives to the acts in CA_t. And finally, where a is some act in CA_t, "$p_t(\varphi(a))$" denotes the agent's personal probability at t that further deliberation during the time he believes to be available for that (i.e., during the time interval t–**t**) will bring to his mind at least one alternative act b such that $EU(b) > EU(a) + c_t$.

We can now formally define the value of further deliberation until **t** at t for an act a, $V_t(a)$, as follows:

Definition 2.1

$$V_t(a) \stackrel{\text{def}}{=} \frac{p_t(\varphi(a))}{EU(a)}.$$

Thus, the value of further deliberation till **t** for an act a at t will be greater the likelier the agent at t thinks it is that he can come up with something better than a before **t** at an acceptable cost, and will be smaller the greater the expected utility of a is. Hence the defined function weighs the factors determining the value of further deliberation in the intuitively right way.[19] Strictly speaking, we could formulate our theory by means of the

[18] For simplicity's sake, I will assume that the agent is always certain about when the time limit expires (which is not to say he is necessarily right about that – see Section 4). Clearly, in reality, this is not true for every, or indeed any, agent. However, it is possible (though, as far as I can see, quite cumbersome) to adapt the definition of value of further deliberation so that it can account for an agent's uncertainty regarding the time limit.

[19] It is not difficult to modify this definition so that the value of further deliberation also comes to depend on the size of the possible gains from further deliberation. There is, in fact, a variety of ways of doing this, but here is one that allows us to take into consideration n degrees of betterness of acts, for arbitrary n. Let there be n degrees of betterness (characterized either qualitatively or quantitatively in terms of units of utility), numbered from 1 till n, with higher numbers corresponding to higher degrees of betterness, and such that a's being better than b to a degree of $i + 1$ implicates a's being better than b to a degree of i (this is like saying that, for instance, an act that is much better than some other act also is moderately better than that same act). Furthermore, let "$p_t(\varphi_i(a))$" denote the agent's probability at t that before **t** he can think of an act that is better than a to a degree of i. Then we can define $V_t(a)$ as $d_t(a)/EU(a)$, with $d_t(a) = \left[\sum_{i=1}^{n} p_t(\varphi_i(a))\right]/n$. It is immediate that the numerator of the latter fraction will be higher – and thus that, given this definition, the value of further deliberation for an act a will be higher – if the agent thinks it likely that his search will result in a much better act than a than if he thinks it likely his search will result in an only moderately better act than a, all else being equal.

In this paper my main concern will be with delineating the decision theory to be presented shortly from BDT, and not so much with distinguishing between more or less

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

foregoing definition alone. However, in order to facilitate presentation of the theory, we explicitly define the value of further deliberation at t (*simpliciter*), $V_t(\mathrm{CA}_t)$, as the least value of further deliberation for any of the acts considered at t, or, more formally:

Definition 2.2

$$V_t(\mathrm{CA}_t) \stackrel{\text{def}}{=} V_t(a) : a \in \mathrm{CA}_t \;\&\; \neg\exists b \in \mathrm{CA}_t(V_t(b) < V_t(a)).$$

The following proposition, stated without proof, enumerates some immediate consequences of these definitions (for proposition 2.1.1 and 2.2.2 recall that, by stipulation, utility functions have 1 as their greatest lower bound):

Proposition 2.1

1. $\forall a, t : 0 \leq V_t(a) \leq 1.$
2. $\forall a, t : 0 \leq V_t(\mathrm{CA}_t) \leq 1.$
3. $\forall a, t : V_t(a) = 0 \Longleftrightarrow p_t(\varphi(a)) = 0.$
4. $\forall a, t : V_t(a) = 1 \Longleftrightarrow p_t(\varphi(a)) = 1 \;\&\; EU(a) = 1.$
5. $\forall t\big[V_t(\mathrm{CA}_t) = 0 \Longleftrightarrow \exists a(a \in \mathrm{CA}_t \;\&\; V_t(a) = 0)\big].$
6. $\forall t\big[V_t(\mathrm{CA}_t) = 1 \Longleftrightarrow \forall a(a \in \mathrm{CA}_t \Rightarrow V_t(a) = 1)\big].$

Two comments are in order before we can formulate our new decision theory. First, on standard BDT, two utility functions are equivalent if and only if one can be obtained from the other by some positive linear transformation, that is, if there are real numbers x and y with $x > 0$ such that

$$u'(\cdot) = xu(\cdot) + y.$$

But it is not true that such functions are equivalent with respect to definitions 2.1 and 2.2. Let utility functions $u(\cdot)$ and $u'(\cdot)$ both have 1 as their greatest lower bound, and let $u'(\cdot) = xu(\cdot) + y$. It is easily seen that, if $x > 1$, the expected utility of the acts considered will, given definition 2.1, have more import relative to the estimated chance that further deliberation will be successful if the agent's utilities are represented by $u'(\cdot)$ than if they are represented by $u(\cdot)$; if $0 < x < 1$, the expected utility

refined versions of the new theory. In particular, this means that the consequences to be derived from the latter follow regardless of whether the simple definition or the refined definition of $V_t(\cdot)$ is assumed. (This is due to the fact that, mathematically speaking, the only properties of $V_t(\cdot)$ involved in these derivations are those stated in proposition 2.1 below, which – substituting "$d_t(a)$" for "$p_t(\varphi(a))$" in 2.1.3 and 2.1.4 – remains valid given the refined definition.) For the purposes of this paper the reader can therefore take the value of further deliberation either as being defined by definition 2.1 or as being defined along the lines suggested in this note, as he or she prefers (in the latter case "$d_t(a)$" has to be substituted for "$p_t(\varphi(a))$" throughout).

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

will have less import given $u'(\cdot)$ than given $u(\cdot)$.[20] One consequence of this is that it does not hold that if $V_t(a) = r$ given that the agent's utilities are represented by some function $u(\cdot)$, $V_t(a)$ will equal r given any positive linear transformation of $u(\cdot)$.[21]

In determining an act's value of further deliberation, is it possible to lay down generally how heavily the expected utility of that act should weigh relative to the agent's probability that he can come up with a better alternative? If so, then we could put further constraints on utility functions. However, I do not see any basis for doing that. Perhaps the best that can be said here is that everyone should choose a transformation of the agent's utility function according to how relevant to the value of further deliberation he or she feels the expected utility of acts is (keeping in mind that the greatest lower bound on the utility function should equal 1). Alternatively, one could choose a transformation in accordance with how the *agent* feels about this point, but, unless one is oneself the agent, this would evidently be harder to carry out practically.

Secondly, I should like to point to the fact that the function here called the value of further deliberation only measures the value of further deliberation about possible, unconsidered alternatives to the acts already considered in some decision-making situation. However, further deliberation can be about many things. It can be about whether one really prefers swimming to going to the opera, about whether one really took account of all available information, about whether one should not, before making a decision, gather additional information, about whether the acts one considers choosing from really are *alternatives* (or whether perhaps one can choose two, or even more, of them), and so on. Just as we have clear intuitions about the circumstances under which further deliberation concerning alternative courses of action is the rational thing to do, so we have clear intuitions concerning the circumstances under which further deliberation about these other issues becomes rational. A complete account of rational agency should certainly do justice to all those intuitions and, thus, should ideally be formulated in terms of a function measuring the value of further deliberation quite generally (or in terms of several functions each measuring the value of further

[20] Since it must hold that $y = 1 - x$, lest $u'(\cdot)$ violates our constraint on utility functions that they have a greatest lower bound of 1, y must be 0 if $x = 1$, so that in that case $u(\cdot) = u'(\cdot)$.

[21] If one finds this confusing, or, for some other reason, prefers to stick to the traditional manner of representing utilities by a function that is unique up to linear transformations, one can define a new function $f(\cdot)$ such that $f(a) = ru(a) + s$ with r and s so chosen that the new function satisfies the constraints that in the text are imposed upon utility functions. To determine the value of further deliberation for a given act we must then, in the denominator of the fraction in definition 2.1, consider not the act's expected utility but the property measured by a new function (say) $g(\cdot)$ such that $g(a) = \sum_{w \in W} p(w)f(a.w) = rEU(a) + s$.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

deliberation about some particular issue). I must confess, however, that presently I have no clue as to what such a function would have to look like (nor of how to combine in an elegant fashion a number of distinct functions measuring the value of further deliberation about separate aspects of the decision-making situation). Thus, in the sequel by ''value of further deliberation'' we will mean the function as specified by definition 2.1, the function, that is, measuring solely the value of further deliberation about potential unconsidered alternatives. Inevitably this means that the decision theory to be presented is still not quite complete. Yet it is decisively more complete than standard BDT and, hopefully, is also a first step toward a truly complete decision theory.[22]

3. DELIBERATION AMENDED DECISION THEORY.

The notion of value of further deliberation has been defined in BDT's most basic terms. Furthermore, the theory to be built on that notion will be seen to be in partial agreement with BDT. For these reasons we can conceive of the theory to be presented as a reformulation of BDT, rather than as a wholly new approach to the rationality of action. The reformulation of decision theory I want to propose – I call it Deliberation Amended Decision Theory, or DADT for short[23] – defines rationality of choice by means of the following two axioms:

Axiom 3.1 *The rationality of choosing act a at time t is a linearly decreasing function of* $V_t(a)$.

[22] In terms of the literature on idealization referred to in note 3, the new decision theory will only be a *partial concretization* (or, in Mäki's (1992, 1994), vocabulary, a *partial de-isolation*) of BDT: compared to the latter it takes into account an important extra factor relevant to decision making, the factor related to the process of searching for alternatives, but it is still idealized in certain respects. This might seem rather disappointing. But then it should be realized that step-wise concretization is the normal procedure in science, i.e., scientists typically start out with some highly idealized theory and then attempt to work into the theory, *one at a time*, whatever relevant factors the theory initially neglects. To give a famous example, the theory of ideal gases involves many false idealizing assumptions, like for instance that the molecules of a gas have zero volume and that they do not exert any attractive forces. Clausius formulated a predictively more accurate version of this theory by correcting for the volume of the gas molecules. Somewhat later, van der Waals achieved a further concretization of the theory; the equation that now goes by his name also takes into account the intermolecular forces (as well as certain other factors which both the ideal gas law and Clausius' emendation of it neglect). But we know that even the van der Waals equation is not a full concretization of the ideal gas law (e.g., it assumes that the molecules of a gas obey the laws of classical mechanics, which is not entirely true). For more detailed illustrations of the strategy of concretization, see Kuipers (2000, Chapter 11) on the development of the early quantum theory and the development of capital structure theory in economics.

[23] This name was suggested to me by Uskali Mäki.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

Axiom 3.2 *The rationality of choosing to deliberate further at t is a linearly increasing function of $V_t(\mathrm{CA}_t)$.*

Given these axioms, rationality is no longer a matter of all or nothing, as it is on standard Bayesian decision theory, but of more or less: acts are more or less rational as the value of further deliberation can be smaller or greater. There are limit points. We can say that choosing an act a at t is *fully*, or *maximally*, rational just in case $V_t(a) = 0$, and that it is *fully* irrational, or *minimally* rational, just in case $V_t(a) = 1$. Likewise, it is *fully* rational to deliberate further at t exactly if $V_t(\mathrm{CA}_t) = 1$ and *fully* irrational to deliberate further at t exactly if there is an $a \in \mathrm{CA}_t$ such that $V_t(a) = 0$.[24] These limit points are related in the most obvious way: there is some act $a \in \mathrm{CA}_t$ such that it is fully rational to choose it at t if and only if it is fully irrational to deliberate further at t (this follows from the foregoing axioms and proposition 2.1.5). And, it is fully rational to deliberate further at t if and only if for all $a \in \mathrm{CA}_t$ it is fully irrational to choose a at t (from proposition 2.1.6 and the axioms). More generally, it can readily be seen to hold that

Proposition 3.1 *The rationality of further deliberation (simpliciter) at t and the rationality of choosing at t the or a best act in CA_t are inversely proportional to each other.*

The following propositions state some further consequences of DADT. The proofs assume that the agent is cognitively rational (this includes the assumption that he obeys Lewis's (1980) principal principle, which, very roughly, says that one's degree of belief in a proposition given that the objective probability or chance of that proposition equals x, should equal x).

Proposition 3.2 *For all a, t: if $a \in \mathrm{CA}_t$ does not bear maximum expected utility relative to CA_t, it can never be fully rational to choose a at t. Put contrapositively, if choosing an act $a \in \mathrm{CA}_t$ is, on our theory, fully rational, then choosing a at t maximizes expected utility relative to CA_t.*

Proof: Suppose $a \in \mathrm{CA}_t$ and a does not maximize expected utility. Then the chance or objective probability that the agent will come up with something better before **t** at no cost at all is 1 (this is true even if $t =$ **t**: the agent already *has* come up with it). So (assuming the principal principle) $p_t(\varphi(a))$ will be 1. Hence, by proposition 2.1.3, $V_t(a) \neq 0$, and, thus, by axiom 3.1, choosing a at t cannot be fully rational. ■

[24] It may accord better with ordinary language use to call choosing a simply rational (dropping "fully") if $V_t(a) = 0$ and irrational otherwise, but increasingly so, as $V_t(a)$ approaches 1. For further deliberation the converse may hold: it may be most natural to say that further deliberation at t is *ir*rational exactly if there is an $a \in \mathrm{CA}_t$ such that $V_t(a) = 0$ and rational otherwise, though increasingly so, as $V_t(\mathrm{CA}_t)$ increases.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

Proposition 3.3 *For all a, t: if $a \in \mathrm{CA}_t$ does not maximize expected utility relative to CA_t, the rationality of choosing a at t is a function solely of $EU(a)$.*

Proof: We saw in the foregoing proof that $p_t(\varphi(a)) = 1$ for any act $a \in \mathrm{CA}_t$ satisfying the antecedent. Let a and b be two such acts. Then $V_t(a) = {}^1EU(a)$ and $V_t(b) = {}^1EU(b)$, so that

$$V_t(a) \leq V_t(b) \iff EU(a) \geq EU(b).$$

Hence, by axiom 3.1, the rationality of choosing a at t is at most as great as the rationality of choosing b at t exactly if the expected utility of a at t is at least as great as the expected utility of b at t. In other words, for non-maximizing acts the rationality of choosing them is a function of their expected utility only. ■

Proposition 3.4 *For all a, b, t: if $a, b \in \mathrm{CA}_t$ and $EU(a) = EU(b)$, then choosing act a at t is just as (ir-)rational as choosing act b at t.*

Proof: Suppose $a, b \in \mathrm{CA}_t$ and $EU(a) = EU(b)$. We must distinguish between the case in which there is a $c \in \mathrm{CA}_t$ such that $EU(c) > EU(a), EU(b)$ and the case in which there is no such c. In the first case, it follows from proposition 3.3 that the rationality of choosing a or b is a function of their expected utility only. Since their expected utilities are the same, the rationality of choosing one cannot differ from that of choosing the other. In case a and b have maximum expected utility relative to CA_t, we must consider $p_t(\varphi(a))$ and $p_t(\varphi(b))$. Since coming up with an act with greater (cost adjusted) expected utility than a simply *is* coming up with an act with greater (cost adjusted) expected utility than b, the agent (who is supposed to be cognitively rational) will deem the one as likely as the other, that is, it will hold that $p_t(\varphi(a)) = p_t(\varphi(b))$. But then, since also $EU(a) = EU(b)$, it must be that $V_t(a) = V_t(b)$, and hence that it is as rational to choose a at t as it is to choose b at t. So, whether a and b do or do not have maximum expected utility relative to CA_t, choosing one at t is as rational (or irrational) as choosing the other at t. ■

Proposition 3.5 *For all a, t: if a maximizes expected utility relative to CA_t and the agent knows or believes that a will maximize expected utility relative to $\mathrm{CA}_{t'}$ for all t' such that $t \leq t' \leq \mathbf{t}$, then choosing a at t is fully rational.*

Proof: Assume the antecedent. Then, where a is any best act in CA_t, $p_t(\varphi(a)) = 0$ will be a partial specification of the agent's belief state at t. Hence, by proposition 2.1.3, $V_t(a) = 0$ for any best act $a \in \mathrm{CA}_t$ and, thus, choosing it at t is fully rational. ■

So, for instance, if you believe that act a is the optimal solution to the decision problem you are facing, that is, that no alternative, whether already considered or overlooked so far, is better than a, then choosing a right away is the rational thing to do. Further deliberation about what

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

else may be up to you will seem to be a waste of time and energy. Also, if you know or believe that $t = \mathbf{t}$, then the proposition's antecedent will hold trivially for you (provided you are cognitively rational). Thus, in that case too, choosing $a \in \mathrm{CA}_t$ at t is fully rational if a maximizes expected utility relative to CA_t. Again this seems right: you believe you *must* choose now. What then could be more rational than to choose the or a best act of those presently before you? A third important class of cases in which the antecedent holds are those in which the agent knows CA_t to comprise *all* the available alternatives in the given situation. Deliberating further would in such cases be wholly nonsensical – there is nothing it could bring you.

It is well worth stressing that propositions 3.2, 3.4, and 3.5 establish some interesting partial correspondences between the new theory and BDT:[25] If we take CA_t in both propositions as the set of options given to the agent (in the sense of BDT's idealizing assumption discussed earlier), then BDT yields basically the same result as DADT in those cases in which one of the propositions' antecedents holds. The only difference is terminological in that BDT does not speak of fully rational choice – but if we assimilate the latter to BDT's ''rational choice'', as seems reasonable to do (see note 24), then the verdicts of both theories in the indicated cases read exactly the same. It is desirable for any formal decision theory to be in agreement with BDT on the points just noted, given that here the verdicts BDT yields are clearly unassailable.

Of course, there are also some unambiguous differences between the two theories. Presumably the most significant of these is the fact that DADT provides the right kind of answer in those cases that motivated our project, namely, the cases in which BDT's idealizing assumption that the agent is simply given his options cannot plausibly be assumed to hold. This will come as no surprise. After all, the whole project was aimed at developing a decision theory that incorporates exactly those facets of the process of deliberation which, we feel, crucially matter to the rationality of what the agent in the end decides to do, facets that BDT sidelines by making the idealizing assumption discussed in Section 1. And we defined our notion of value of further deliberation, and concomitantly the notion of rationality, in precisely such a way that these facets are given their due. Still, as an illustration of how DADT handles the kind of cases where, as we saw earlier, BDT failed so conspicuously, we briefly consider again our example from Section 1.

Recall the main features of the situation: none of the options for spending the evening currently before you is appealing; you have all afternoon to search for better alternatives; and you believe that there is a

[25] Of course, proposition 3.3 reveals a correspondence between the theories, too, but not one that seems to be of much significance.

good chance that such a search will or would be worth the effort. Nonetheless, you decide on the spot what to do this evening. We feel strongly that you rushed into your decision and that it would be wrong to call it rational. But we saw that BDT's verdict may well be at odds with this intuition: if you choose the or a best option of the ones you consider, then, BDT, in its present formulation, seems to qualify your choice as rational. DADT, on the other hand, will certainly not do so. For, given the relevant characteristics of the situation, the value of further deliberation for even the best of the options that lay before you must be quite high for you. Hence, choosing even the best or one of the best of these will, by axiom 3.1, be quite irrational.

In addition to this, whether you decide to deliberate further about alternative options or to confine your choice to the few options you are now aware of, we need not invoke second-order decision theory – with the concomitant possibility of initiating a regress – to obtain a verdict about that decision. Since it would be quite irrational for you to choose the best of the small set of options you have so far considered, it follows immediately from proposition 3.1 that further deliberation at this time is quite rational for you.

To end this section, I would like to mention two additional differences between our theory and BDT that are both due to the fact that on our theory rationality is not a categorical concept.

First, DADT allows, much to its credit I believe, for comparative judgements of rationality of the kind we all frequently and quite naturally make.[26] Consider that if, in the above example, you were instantly to choose the best of the options you have pondered, we would find your decision irrational. But, supposing that these options were not all equally unattractive to you, we would find your decision even more irrational were you to choose, from among these acts, the one you deemed *worst*.

And second, while BDT condemns any non-optimizing behavior (whether or not it is satisficing) as irrational, on DADT, such behavior can be rational, albeit not fully rational. It thus seems that DADT makes it possible to accommodate to some extent the intuitions of those attracted to a satisficing account of rationality.

Better yet – better, at least, from the perspective of the latter group – it is quite straightforward to turn DADT into a fully-fledged theory of satisficing (although in the face of DADT I do not quite see what motivation there could be left for endorsing a satisficing account of

[26] Psillos (1999, p. 227) urges the same point with respect to epistemic rationality – "A full theory of rational belief . . . should allow for *comparative judgements*: some beliefs are more rational than others" – and castigates the Bayesian theory of epistemic rationality for only admitting categorical judgements of rationality.

rationality). To achieve such a theory, we merely need to determine that it is *fully* rational to choose the first act the agent comes up with whose associated value of further deliberation is below a certain threshold value. Not only would this yield a genuine *theory* of satisficing, the account would also incorporate a psychologically significant fact of actual human decision-making behavior that the advocates of a satisficing approach to rationality tend to neglect (in spite of the explicit concern of many of them to formulate a psychologically realistic theory). Remember, that according to these authors, the rational agent terminates his search for alternatives as soon as he finds an act that is good enough, in the sense that its expected utility meets or exceeds the agent's aspiration level. Now, suppose you come up with an act a you find good enough, but at the same time, are quite confident that spending just a little more time deliberating will suggest an act you will find much better than a. Then it would seem rash if you were to choose a, and I venture that it is unrealistic to suppose that agents typically do so. The adapted version of DADT suggested here, however, does account for that intuition: whether the value of further deliberation for a given act drops below a certain threshold value does not just depend on the expected utility of that act, but also on the agent's personal probability that he is able to find a better alternative (though it will be recalled from the first comment at the end of Section 2 that the extent to which the latter factor matters depends on the choice of unit of utility).[27]

4. OBJECTIFIED DELIBERATION AMENDED DECISION THEORIES.

As we introduced '**t**' in Section 2, "t–**t**" denotes the time interval the agent at t *believes* to have available for further deliberation; likewise, we have taken "$p_t(\varphi(a))$" in definition 2.1 as the agent's *personal* or *subjective* probability that further deliberation will bring to his mind some act with expected utility exceeding a's plus the cost of further deliberation. It is noteworthy, however, that these elements from the definition of $V_t(\cdot)$ can also be given objective (or more objective) interpretations. For instance, "**t**" could be interpreted as the *actual* (instead of the supposed) time limit for further deliberation at t, so that "t–**t**" would denote the time interval *actually* available for further deliberation at t; and "$p_t(\varphi(a))$" could be interpreted as the *objective* probability that, if the agent deliberates further, he will come up with at least one act with greater (cost adjusted) expected utility than a. Interestingly, these other interpretations of "**t**" and/or "$p_t(\varphi(a))$", and, hence, of the central terms occurring in axioms

[27] Note that DADT could be thought of as a special case of the theory of satisficing here envisaged, namely, the case in which 0 is imposed as a threshold on the value of further deliberation.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

3.1 and 3.2, do not so much yield other interpretations of DADT, as give us new decision theories.[28]

To demonstrate this, suppose that we interpret "**t**" and "$p_t(\varphi(a))$" in the objective way previously suggested. Then we have the following:

Proposition 4.1 *For all t and corresponding* **t***: deliberating till* **t** *will allow the agent to make a choice that is at least as rational as any choice he could make at t', with $t \leq t' \leq$* **t**.

Proof: The objective probability that at **t** the agent can arrive at some better act than the best act(s) in $\mathrm{CA}_{\mathbf{t}}$ equals 0. Hence, by proposition 2.1.5, $V_t(\mathrm{CA}_t) = 0$, so that by axiom 3.1 choosing any best act $a \in \mathrm{CA}_t$ will be fully rational. This holds regardless of whether there are acts in $\mathrm{CA}_{\mathbf{t}}$ better than the best in CA_t. ■

Without the supposition, proposition 4.1 would not hold. For instance, the agent may at t have completely mistaken beliefs about when the time limit will expire. If it expires earlier than he believes it will, he may very well let the opportunity to choose slip by.

It should be emphasized that the foregoing proposition does *not* implicate that, on the considered objectified decision theory, further deliberation is always fully rational. It may be that you can only choose fully rationally at **t** because you (fully) irrationally postponed your choice at t in order to deliberate further until **t**. Think again of the case where you already at t know that you have the optimal solution to your problem. Then although choosing it at **t** is fully rational – that is, at *that* time, there is nothing you could do that is more rational than choosing the optimal solution – the fact that you deliberated till **t** is still fully irrational.

However, a decision theory thus objectified seems unacceptable for the simple reason that, *generally*, we lack a reliable method of finding out what the time limit on the situation is or what the objective probability is that some better act will occur to us before the time limit expires. Should an agent then really be called irrational because it seemed very unlikely to him that he could come up with anything better than some act a and therefore chose that act, whereas it was *really* quite probable that, had he thought a little longer about what to do, some alternative act b with greater cost adjusted expected utility than a would have occurred to him? Surely we must answer this question in the negative (see Carnap, 1962, p. 305).

[28] Here I am following the terminological usage of mathematicians and logicians, who identify theories with sets of axioms together with their consequences. However, nothing hangs on this usage, and, if one prefers, one may equally well think of the different interpretations of "**t**" and "$p_t(\varphi(a))$" as yielding new interpretations of DADT on which that theory has novel consequences.

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

This does not rule out all variants of DADT that assume interpretations of (some of) the formalism's key terms that are more objective than the purely subjective interpretation assumed throughout the previous section. Consider, for instance, the theory that is just like DADT but which takes the subjective probabilities appropriate for decision theory to be not mere subjective probabilities but what we may call *evidential probabilities*, that is, subjective probabilities based on, or informed by, the available (relevant) statistical data.[29] The requirement that our personal or subjective probabilities be evidential probabilities is not normally taken to be part of the Bayesian ideal of cognitive rationality, but there seems to be no good reason not to *make* it part of that ideal. Few (if any) of us will always base their personal probabilities on the available statistical data, but at least it is something we *could* all do if we wanted. So the proposal to add to DADT the requirement that our probabilities be evidential probabilities cannot be discounted for the same reason we discounted the purely objective decision theory considered above.[30]

Like the earlier objective decision theory, this new theory is a genuinely distinct theory[31] and not just a novel interpretation of DADT. To show this, let me just state one rather surprising consequence of this theory that does not follow from DADT.

Proposition 4.2 *If A believes that* $t \neq \mathbf{t}$*, then choosing rationally is easier to accomplish if he has generally been unsuccessful in coming up with better alternatives on earlier occasions of further deliberation than if he has been generally successful on such occasions.*

Proof: Suppose that Kate, who has generally been quite successful in thinking up better acts than the ones considered whenever she has deliberated further on what to do in a given situation, and Jim, who has been far less successful on such occasions, at t respectively t' consider

[29] See, e.g., Kyburg (1974, 1990) and Salmon (1990). On Kyburg's definition, evidential probabilities are required to exactly reflect relative frequencies and thus are not really probabilities (they are interval-valued instead of real-valued – see, for instance, Kyburg (1990, p. 43)). I see little harm in relaxing this requirement and taking evidential probabilities to be informed by statistical data *as much as possible* (where, of course, it will have to be precisely laid down what it is for probabilities to be informed by statistics "as much as possible"). Instead of evidential probability, one could consider taking other varieties of "more-than-merely-subjective" probability, like, e.g., Carnap's (1950) logical probability, which Carnap himself (1962) proposes as the appropriate notion of probability for the formulation of decision theory, or Gillies's (1991, 2000, Chapter 8) intersubjective probability (which he thinks is the key notion of probability involved in economics; see Gillies, 2000, Chapter 9 and Gillies and Ietto-Gillies, 1991), or Williamson's (1998) evidential probability, or Achinstein's (2001, Chapter 5) objective epistemic probability, or Swinburne's (2001) logical probability.

[30] In fact, the proposal could be regarded as a first step toward Maher's qualified Bayesianism; see note 12.

[31] In the mathematical sense of that word; see note 28.

exactly the same acts to choose from, or briefly, $CA_t^{Kate} = CA_{t'}^{Jim}$. Suppose further that each act in this set has the same expected utility given Kate's probability and utility functions as it has given Jim's, and also that $\mathbf{t}-t = \mathbf{t}' - t'$ (i.e., the amount of time Kate at t believes to have available for further deliberation is the same amount of time Jim at t' believes to have available for further deliberation). Then, given what we said about their past success rates for further deliberation, for each act a in the collection of acts considered by Kate at t and Jim at t', it holds that Jim's evidential probability that he can come up with something better than a before $\mathbf{t}'$ is less than or equal to Kate's evidential probability that she can come up with something better than a before $\mathbf{t}$. Thus, for each of the acts they both consider, the value of further deliberation will be at least as high for Kate as it is for Jim. And so, for each act $a \in CA_{t'}^{Jim}$ $(= CA_t^{Kate})$, Jim's choosing it at t' will always be at least as rational as Kate's choosing it at t. The converse does not hold: again, given our assumption about their past performances, for any act a with maximum expected utility relative to those Kate and Jim consider, Kate's evidential probability at t that she can come up with something better before $\mathbf{t}$ will be higher than Jim's evidential probability at t' that he can come up with something better than a before $\mathbf{t}'$. ■

To see that this proposition does not hold on DADT, suppose that Jim is mostly over-confident about his own ingenuity. Totally neglecting his past failures in decision situations to come up with better acts than the ones already considered, he, in the above situation, believes that he can think up some better act to a degree that is not at all warranted by the statistical data. So, it can happen that for Jim the value of further deliberation for the best act(s) in $CA_{t'}^{Jim}$ is higher than the value of further deliberation this or these acts have for Kate at t. Thus, it may be that if Kate chooses that act (respectively one of those acts) at t, her choice is more rational than Jim's if he chooses the same act (respectively one of the same acts) at t'.

5. SUMMARY AND PROJECTS

We started by showing that standard Bayesian decision theory may lead to incorrect results in a large class of decision-making situations. Owing to the fact that the theory is abstracted from the *process* of decision making, it systematically neglects factors related to the search for options that from an intuitive standpoint are crucially important to the issue of rational choice. It was then exhibited how these factors can be incorporated in an intuitively correct way into the formal structure of a decision theoretic model. The central new notion needed for carrying out the project was that of the value of further deliberation for a given act,

https://doi.org/10.1017/S0266267102002079 Published online by Cambridge University Press

which we defined as a function that weighs against each other the chief factors determinative of how valuable further deliberation about other possible acts is as opposed to choosing the given act right away. The decision theory formulated by means of this function was seen to be in partial agreement with standard decision theory. More importantly, however, we saw that the new theory is in agreement with intuition in those decision-making situations in which BDT is not. Finally, it was shown that the formalism developed has more than one plausible interpretation, and that distinct interpretations of it may lead to interestingly distinct decision theories.

At the end of Section 2 it was remarked that, although DADT is a less idealized theory than BDT, it too neglects certain of the factors BDT neglects but which intuitively matter to the rationality of action. As we saw, many aspects of a decision-making situation may call for further deliberation, not just the aspect of possibly unconsidered acts that are available for choice. One obvious avenue for further research, then, is to see whether, with some (preferably minor) modifications, DADT can account for the rationality of further deliberation in a more general sense. A second issue that I have not taken up but which is certainly worth pursuing is whether, of the decision theories presented in this paper, there is one that is to be preferred to any of the others. I have yet to see a decisive philosophical argument favouring any particular one; I am not even sure that this matter can be decided on purely philosophical grounds. Empirical work on DADT and its variants, if that is feasible, might be of help here.

REFERENCES

Achinstein, P. 2001. *The Book of Evidence*. Oxford University Press

Byron, M. 1998. Satisficing and optimality. *Ethics*, 109:67–93

Carnap, R. 1950. *Logical Foundations of Probability*. University of Chicago Press

Carnap, R. 1962. The aim of inductive logic. In *Logic, Methodology and Philosophy of Science*, pp. 303–18. E. Nagel, P. Suppes and A. Tarski (eds.). Stanford University Press

Cartwright, N. 1983. *How the Laws of Physics Lie*. Clarendon Press

Cartwright, N. 1999. *The Dappled World*. Cambridge University Press

Chernoff, H. and L. Moses. 1959. *Elementary Decision Theory*. Wiley

Douven, I. 1999. Inference to the best explanation made coherent. *Philosophy of Science*, 66:S424–35

Douven, I. 2002a. Testing inference to the best explanation. *Synthese*, 130:355–77

Douven, I. 2002b. Empirical equivalence, explanatory force, and the inference to the best theory. In *Logics of Scientific Cognition: Essays in Debate with Theo Kuipers*. A. Aliseda, R. Festa and J. Peijnenburg (eds.). Rodopi, forthcoming

Douven, I. 2002c. A new solution to the paradoxes of rational acceptability. *British Journal for the Philosophy of Science*, 53:391–410

Dreier, J. 1996. Rational preference: decision theory as a theory of practical rationality. *Theory and Decision*, 40:249–76

Earman, J. 1992. *Bayes or Bust?* MIT Press

Eells, E. 1982. *Rational Decision and Causality*. Cambridge University Press

Elster, J. 1983. *Sour Grapes*. Cambridge University Press
Ferguson, T. 1967. *Mathematical Statistics*. Academic Press
Foley, R. and R. Fumerton. 1982. Epistemic indolence. *Mind*, 91:38–56
Giere, R. 1988. *Explaining Science*. University of Chicago Press
Giere, R. 1993. In *Taking the Naturalistic Turn*. W. Callebaut (ed). University of Chicago Press
Giere, R. 1999. *Science Without Laws*. University of Chicago Press
Gillies, D. 1991. Intersubjective probability and confirmation theory. *British Journal for the Philosophy of Science*, 42:513–33
Gillies, D. 2000. *Philosophical Theories of Probability*. Routledge
Gillies, D. and G. Ietto-Gillies. 1991. Intersubjective probability and economics. *Review of Political Economy*, 3:393–417
Hacking, I. 1975. *The Emergence of Probability*. Cambridge University Press
Hacking, I. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge University Press
Hamminga, B. and N. De Marchi (eds.). 1994. *Idealization VI: Idealization in Economics* (*Poznań Studies in the Philosophy of the Sciences and the Humanities, Vol. 38*). Rodopi
Hampton, J. 1994. The failure of expected-utility theory as a theory of reason. *Economics and Philosophy*, 10:195–242
Hansson, B. 1988. Risk aversion as a problem of conjoint measurement. In *Decision, Probability, and Utility*, pp. 136–58. P. Gärdernfors and N.-E. Sahlin (eds.). Cambridge University Press
Herrnstein, R. 1988. A behavioural alternative to utility maximization. In *Applied Behavioural Economics*, Vol. 1, pp. 3–60. S. Maital (ed.). Wheatsheaf
Johansen, L. 1977. *Lectures on Macroeconomic Planning*, Part 1. North-Holland
Jeffrey, R. 1983. *The Logic of Decision* (2nd edn.). University of Chicago Press
Kaplan, M. 1996. *Decision Theory as Philosophy*. Cambridge University Press
Krajewski, W. 1977. *Correspondence Principle and Growth of Science*. Reidel
Kuipers, T. 2000. *From Instrumentalism to Constructive Realism*. Kluwer
Kuipers, T. 2001. *Structures in Science*. Kluwer
Kusser, A. and W. Spohn. 1992. The utility of pleasure is a pain for decision theory. *Journal of Philosophy*, 89:10–29
Kyburg, Jr., H. 1974. *The Logical Foundations of Statistical Inference*. Reidel
Kyburg, Jr., H. 1990. *Science and Reason*. Oxford University Press
Laville, F. 2000. Foundations of procedural rationality: cognitive limits and decision processes. *Economics and Philosophy*, 16:117–38
Lewis, D. 1980. A subjectivists guide to objective chance. In *Studies in Inductive Logic and Probability*, pp. 263–93. R. Jeffrey (ed.). University of California Press
Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy*, 59:5–30
Luce, R. D. and H. Raiffa. 1957. *Games and Decisions*. Wiley
Maher, P. 1993. *Betting on Theories*. Cambridge University Press
Mäki, U. 1992. On the method of isolation in economics. In *Idealization IV: Intelligibility in Science* (*Poznań Studies in the Philosophy of the Sciences and the Humanities, Vol. 26*), pp. 319–54. C. Dilworth (ed.). Rodopi
Mäki, U. 1994. Isolation, idealization and truth in economics. In Hamminga and De Marchi (eds.). (1994), pp. 147–68
March, J. 1978. Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics*, 9:587–608
March, J. and H. Simon. 1958. *Organizations*. Wiley
Mongin, P. and B. Walliser. 1988. Infinite regression in the optimizing theory of decision. In *Risk, Decision and Rationality*, pp. 435–57. B. Munier (ed.). Reidel
Moser, P. 1990. Rationality in action. In *Rationality in Action*, pp. 1–16. P. Moser (ed.). Cambridge University Press

Nelson, A. 2001. Two models of idealization in economics. In *The Economic World View*, pp. 359–68. U. Mäki (ed.). Cambridge University Press
Niiniluoto, I. 1999. *Critical Scientific Realism*. Clarendon Press
Nowak, L. 1980. *The Structure of Idealization*. Reidel
Pettit, P. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society* (Supplement), 58:165–76
Psillos, S. 1999. *Scientific Realism*. Routledge
Rabin, M. 2000. Risk aversion and expected-utility theory: a calibration theorem. *Econometrica*, 68:1281–92
Raiffa, H. 1968. *Decision Analysis*. Addison-Wesley
Ramsey, F. P. 1926. Truth and probability. In his *The Foundations of Mathematics*, pp. 156–98. Routledge and Kegan Paul, 1931
Resnik, M. 1987. *Choices*. University of Minnesota Press
Romer, P. 2000. Thinking and feeling. *American Economic Review*, 90:439–43
Salmon, W. 1990. Rationality and objectivity *or* Tom Kuhn meets Tom Bayes. In *Scientific Theories*, pp. 175–204. C. Wade Savage (ed.). University of Minnesota Press. (Reprinted in *The Philosophy of Science*, pp. 256–89. D. Papineau (ed.). Oxford University Press, 1996; the page reference is to the reprint.)
Satz, D. and J. Ferejohn. 1994. Rational choice and social theory. *Journal of Philosophy*, 91:71–87
Savage, L. 1954. *The Foundations of Statistics*. Wiley
Schick, F. 1997. *Making Choices*. Cambridge University Press
Schmidtz, D. 1992. Rationality within reason. *Journal of Philosophy*, 89:445–66
Simon, H. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–118. (Reprinted in his (1979), pp. 7–19; the page reference is to the reprint.)
Simon, H. 1979. *Models of Thought*. Yale University Press
Simon, H. 1983. *Reason in Human Affairs*. Blackwell
Slote, M. 1989. *Beyond Optimizing*. Harvard University Press
Smith, H. 1991. Deciding how to decide: is there a regress problem?. In *Foundations of Decision Theory*, pp. 194–219. M. Bacharach and S. Hurley (eds.). Basil Blackwell
Sunstein, C. and E. Ullmann-Margalit. 1999. Second-order decisions. *Ethics*, 110:5–31
Swinburne, R. 2001. *Epistemic Justification*. Clarendon Press
van Fraassen, B. 1989. *Laws and Symmetry*. Clarendon Press
Weirich, P. 1986. Expected utility and risk. *British Journal for the Philosophy of Science*, 37:419–42
Weirich, P. 2001. Risk's place in decision rules. *Synthese*, 126:427–41
Williamson, T. 1998. Conditionalizing on knowledge. *British Journal for the Philosophy of Science*, 49:89–121.