DECISION THEORY WITH A HUMAN FACE

RICHARD BRADLEY

Reviewed by Seamus Bradley

Decision Theory with a Human Face
Richard Bradley
Cambridge: Cambridge University Press, 2017, £75.00
ISBN 9781107003217

This is an ambitious book on an interesting topic. Bradley's goal is to extend decision theory in several ways, by accommodating agents who respond to a richer theory of uncertainty, agents with attitudes towards conditionals, and agents who are bounded or imperfect in various ways.

The aim of this book is to develop a formal theory that allows the representation of rational agents with fairly sophisticated attitudes (encompassing several kinds of uncertainty, attitudes towards conditionals, and so on), and to discuss decision-making and change in attitudes in that framework. The book consists of four parts. Part I lays out the basic framework of Bayesian decision theory, Parts II and III detail Bradley's sophisticated version of that theory, and Part IV extends the theory to bounded agents.

The first five chapters of the book—Part I and the first chapter of Part II—provide a careful and thorough recapitulation of Jeffrey's approach to decision-making (covering the first nine chapters of *The Logic of Decision*, more or less). That description undersells the book up to this point. These chapters of *Decision Theory with a Human Face* cover some difficult material and do an admirable job of making it clear. Bradley is explicit and

careful about the interpretation of his basic terms (preference, choice, and so on) and about the methodology he takes himself to be using. For example, he is interested in Bayesianism as a normative theory, but that does not mean that he thinks psychological evidence has no role to play. Representation theorems are 'moves within a search for a reflective equilibrium', where psychological evidence and intuitive judgements about toy cases also have an input. These methodological questions concern matters in need of greater clarity. If the book ended here, after Chapter 5, I would already be happy to recommend it heartily to anyone who wanted to learn about decision theory.

Chapter 6 is where the book really gets out of the shadow of Richard Jeffrey and radically extends his theory. Bradley discusses conditional attitudes, being careful to distinguish conditional attitudes from attitudes to conditional prospects (though the relationship between the two is the topic of the next chapter). He puts it this way:

> I would prefer that were I confronted by a bully that I would act bravely rather than run away. If I were confronted by a bully, however, I would (I predict) prefer to run.

The former is a preference about conditionals, the latter a conditional preference. Bradley represents conditional attitudes in terms of 'suppositional preferences', 'suppositional probability', and so on. Supposition is not a univocal notion and one important contribution of this chapter is to make clear what makes a particular sort of supposition evidential rather than counterfactual; he provides axioms of suppositional preference that are characteristic of evidential supposition.

Throughout the book, there are a great many definitions, axioms, and principles of rationality. Up until Chapter 5, many of the definitions will be somewhat familiar to those working in the field. This changes in Chapter 6, with the introduction of many new axioms—for example, the axiom of suppositional preference—that will not be familiar to many. I found that I just didn't have intuitions about the rationality or otherwise of many of these principles. For example, Bradley suggests that agents can have preferences that take into account not just the actual outcomes, but the chances with which those outcomes occur. (I prefer the outcome where Alice gets the sweets as the result of a fair coin toss where I could have won, to the outcome where Alice just gets the sweets.) While this example of valuing fairness makes sense, it is much less clear what rational requirements there are on valuing the chances of unactualized alternatives.

One might respond that the proof of pudding is in the eating: that these axioms will show themselves useful in being part of an interesting and useful representation theorem. This is, perhaps, fair enough; but Savage's axioms came in for a deserved beating in an earlier chapter despite their being the foundation of an extremely powerful and interesting theorem. Even if I had had intuitions about the status of these principles as rational axioms, I would have been hesitant to accept them: many apparently reasonable principles only become suspect once their unintuitive consequences are carefully spelled out. For example, at first blush the sure thing principle seems like a sensible thing to expect an agent to conform to; and representing acts as functions from states to outcomes seems like a harmless representational device until the constraint such a modelling practice entails is highlighted. So whether Bradley's various conditions for rational belief in conditionals will continue to seem acceptable in the light of sustained scrutiny, for example, is not something I would like to predict here. (It should be noted that this book builds on several papers published by Bradley, some as long ago as 1998, so at least some of this work has been subject to long scrutiny.)

The last chapter of Part II is on the Ramsey test and, more generally, on the relationship between conditional attitudes and attitudes to conditionals. Along the way Bradley discusses Adams's thesis (that the probability of a conditional is the conditional probability) and the attendant triviality results. Evidential supposition on a factual

prospect amounts to conditionalizing on the prospect, so a restricted version of Adams's thesis emerges as a consequence of Bradley's framework. Bradley shows that the restricted version of the thesis is not subject to the triviality worries (although triviality is still something of a looming danger when dealing with probabilities of conditionals). This is also where Bradley discusses rational attitudes to chances and their relationship with credences in counterfactuals.

Part II provides, among other things, a syntactic characterization of a general theory of prospects. Chapter 8—the first chapter of Part III—provides the semantics. This 'multidimensional possible worlds' semantics is, in some respects, similar to standard possible-worlds semantics. The innovation of Bradley's theory here is that as well as fixing the truth or falsity of all factual prospects, a multidimensional possible world also fixes the truth of what would have been the case under some supposition. Two multidimensional worlds can be the same in terms of what they say about what is actually the case, but differ in what they say about what would be the case. This allows Bradley to describe agents who have modal uncertainty—uncertainty about what would be the case—in the standard way: by having a probability function over these fancy possible worlds. So formally speaking, the interpretation of a prospect is a set of sequences of states. A sequence of states consists of a state that tells us what is actually the case, a state that tells us what would be the case if *A* were the case, a state that tells us what would be the case if *B* were the case, and so on for each possible supposition that you want to model.

Discussions about modality or counterfactuals often presume that there is a unique accessibility relation or selection function, or at least that the actual state fixes the selection function. This means that fixing the actual world determines the counterfacts. Bradley's somewhat cumbersome formalism emphasizes that this needn't be the case. I have a suspicion that one could reformulate the theory in terms of pairs of states and selection functions, which might be a neater formulation.

Chapter 9 turns to the topic of rational choice. Bradley describes a very general Jeffrey-style theory of decision, and then shows how various classic decision theories—Savage's, von Neumann and Morgenstern's, causal decision theory—can be accommodated as special cases of the theory. For example, a Savage act is simply a conjunction of conditionals: 'if this state, then this consequence'. Acts of these forms are evaluated in terms of their expected desirability only when certain conditions on preference hold.

Having covered rational belief and rational choice, Bradley naturally turns to rational change in belief in Chapter 10. Conditional belief and supposition have been discussed in earlier chapters, but Chapter 9 tackles the issue of change in belief directly. As well as standard Bayesian conditioning, he looks at Jeffrey conditioning and Adams conditioning (which is like Jeffrey conditioning, except that it is your conditional probabilities in pairs of propositions from two partitions that are set, rather than your unconditional probabilities in some partition). All of these kinds of updates are shown to follow from general principles of change in attitude: you ought to be responsive to your evidence, but in other respects your change in attitude should be conservative. As well as belief change, Bradley also discusses preference change.

In Part IV of the book, Bradley turns to the question of how to accommodate agents who are somehow bounded or imperfect in his framework. As with the last part, there are three tasks here: the basic structure of rational belief, rational decision-making, and rational change in belief. In Chapter 11, he discusses agents who are not fully opinionated or not fully aware. He discusses several distinct kinds of inconsistency: an agent can believe inconsistent things (inconsistency); an agent can believe things that entail an inconsistency, while not believing those consequences (implicit inconsistency); an agent can fail to believe the consequences of her

beliefs (non-omniscience); and an agent can be such that her beliefs can't be extended without implicit inconsistency (non-extendability). These are distinct ways to fail to be rational, and they differ in how severe they are. Bradley also outlines a version of imprecise probabilism here, namely, the view that your degrees of belief are represented by a set of probability functions.

The boundedness of agents under discussion opens up space for new kinds of rational change in belief. How to rationally become aware of something you were previously unaware? How to become opinionated about something about which you previously had no opinion? These are among the issues discussed in Chapter 12. Bradley develops a theory of attitude change that has some similarities to AGM belief revision theory, except that the attitudes are sets of probability functions, rather than sets of propositions.

Chapter 13 turns to decision-making in this more permissive setting. In particular, how should agents make decisions in circumstances of severe uncertainty? Bradley canvasses various decision rules for imprecise probabilities. What's interesting is that he spends most of his time discussing decision rules that induce a complete order on the prospects. There seems to be something of a mismatch here: surely cases of severe uncertainty are cases where it's permissible to have your preference order be incomplete? Indeed, imprecise probabilities were shown to be a natural model of incomplete preferences in Chapter 11. So to now have the decision rules ride roughshod over that incompleteness seems odd. To my mind, the best we can hope for is fairly permissive decision rules that rule out some obviously bad options, but that don't necessarily induce a complete order on the options. Of course, at the end of the day you still have to choose, but the same is true of options judged equal in the preference ranking: one needn't think that choice between equally good options is determined by some principle of rationality, so why assume that in the case of incomparable options?

The final chapter—Chapter 14—outlines a way to go beyond a 'sets of probabilities' model of belief. This new representation of belief uses a set of sets of probability functions, and a relation of 'confidence' between those sets. A set of probability functions represents a probabilistic judgement, and the relation reflects which judgements you are more confident in.

There are two reasons why I would like this book to be widely read. The first is because Bradley's framework is general enough to bring under one roof both parties to a dispute: evidential and causal decision theorists can reframe their disagreement within this general framework as one about the reasonableness of a particular conceptualization of what an act is. The other reason I'd like to see this book read by many people is because of the attention Bradley pays to the methodological underpinnings of his project: such focus on methodology is rare, but it is necessary to move the subject forward.

This is not an easy book. Some sections are mathematically demanding, and—especially in the middle chapters—there are many discussions of non-standard axioms or principles of rationality that are hard to keep track of and about which one doesn't have strong intuitions. But what that mathematical sophistication and those non-standard axioms buys is worth the cost: what Richard Bradley has developed here is an extremely general and extremely powerful theory of rational belief and decision, and he has done so with admirable clarity, thoughtfulness, and rigour.

Seamus Bradley
University of Tilburg
s.bradley@uvt.nl