Synthese (2022) 200:2
https://doi.org/10.1007/s11229-022-03468-6

ORIGINAL RESEARCH

Attitudes, deliberation and decisions

Richard Bradley[1]

Received: 18 August 2021 / Accepted: 20 October 2021 / Published online: 9 February 2022

Abstract
In this paper I discuss the challenges of several authors to the claims I make in Decision Theory with a Human Face regarding the relation between preference and choice, the nature of conditional desire, the semantics of conditionals, attitudes to chances and their role in individuating prospects, belief change under growing awareness and choice under ambiguity

Keywords Attitudes · Choices · Preferences · Rationality · Conditionals · Chance · Rational conditional desire · Belief and belief revision · Unawareness · Ambiguity

This volume of papers, written in response to my book *Decision Theory with a Human Face* (hereafter DTHF) presents a rich variety of both challenges to my work and new ideas. It is a true blessing to have one's work subject to serious and intelligent criticism of this sort and I am very grateful to the contributors. I cannot of course hope to do justice to all the interesting ideas that these papers contain and so shall focus my efforts on responding to direct criticisms of claims I make in the book. Two of the papers in the volume use my work to solve new problems and so I will have relatively little to say about them (though in both cases I very much endorse the project they undertake).

1 Preferences and the choice principle (Thoma)

In DTHF, I defend what I call the Choice Principle: That of the options available to an agent she should choose the one(s) that she prefers most. Thoma (2021) sees my commitment to this principle as deriving from my adherence to what she calls 'judgementalism', a doctrine made up of three claims:

This article belongs to the topical collection "Decision-Making and Hypothetical Reasoning: Themes in the Philosophy of Richard Bradley", edited by H. Orri Stefánsson.

✉ Richard Bradley
r.bradley@lse.ac.uk

[1] Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London, UK

1. Preferences are a type of judgement.
2. Qualitative attitudes have conceptual, methodological and explanatory priority over numerical ones.
3. Rational choice is constrained only by preference.

Judgmentalism, in turn, she takes to be motivated by an internalism about the requirements of rationality: that they concern only the relationship between an agent's attitudes and not their relationship to features of the external world. Internalism is a view she is willing to grant for the purposes of discussion, but she rejects my argument for it and, more importantly, regards it as an insufficient basis for judgementalism. Her main claim however is that judgementalism is false because, contrary to (3), attitudes other than preference rationally constrain choice. This leads her to reject the Choice Principle.

I agree with a great deal of what Thoma says: in particular that I am committed both to internalism and to the first and second of the above claims. But I don't make the third claim in my book and indeed consider it false. Thoma thinks that my statement of the Choice Principle commits me to it. But with this I disagree. I also disagree that the falsity of claim 3 means that the Choice Principle requires modification. I take up these issues below, but first let me respond to Thoma's criticism of my defence of internalism.

Thoma argues that internalism cannot be defended by appeal to the 'ought implies can' principle because it overshoots its target. In particular, she argues, it would imply that those who had false beliefs about the requirements of rationality would not be rationally required to obey them, something she takes to be false. Instead, she suggests, internalism should be defended on the grounds that norms of rationality must be action-guiding. This latter point I agree with, but I see it as being in tune with the application of the 'ought implies can' principle to rationality constraints. And I don't agree that the requirement to conform to the requirements of rationality is itself a requirement of rationality.

Let me spell this last point out a bit. It is true that you ought to conform to the requirements of rationality. But what kind of 'ought' is contained in this claim? It is not, I suggest, the ought of rationality, but some other kind of imperative (perhaps one rooted in the prudential advantages of being rational). We know that it can't be the rational ought precisely because someone who didn't respect it in virtue of not knowing what the requirements of rationality were would not be acting or judging in a manner that is inconsistent by their own lights.

Consider someone with intransitive preferences who believed that it was rationally required that they be so. Then it would both be the case that they have irrational preferences and so ought to revise them to make them transitive and that they ought not to both retain their belief about the requirement to have intransitive preferences and adopt transitive preferences. The second ought *is* an ought of rationality because it derives from a consistency requirement. Suppose that the first is as well (as Thoma claims and I deny). Then it would follow (assuming that rationality does not impose contradictory requirements) that the person rationally ought not to retain their belief that rationality requires intransitive preferences. But while they undoubtedly ought not to retain this belief, it is equally undoubtedly not an ought of rationality (on an internalist view). It follows that the requirement to conform to the requirement to have transitive preferences cannot be a rationality requirement.

Thoma's main claim is that judgementalism is untenable because there are norms governing the framing of decision problems that are not directed at preferences. Her argument for this is compelling. Some normative constraints on framing, such as the ideal of taking into account everything of relevance to the decision, are not internal norms of rationality. But some are: such as incorporating all contingencies into one's representation of a decision problem that one considers (more-than-marginally) relevant to the choice one will make. And one could conform to the Choice Principle even if one didn't respect these norms because the principle doesn't by itself constrain how alternatives are framed. So, she argues, the Choice Principle does not exhaust the rationality constraints on choice.

All this is true. But I do not claim otherwise in my book. On the contrary, I castigate behaviourism (p. 60) for its failure to recognise that the choice an agent makes depends on how the set of alternatives is conceived and, as Thoma herself points out, I argue against the view that 'anything goes' when it comes to framings (p. 13). Nonetheless, Thoma thinks that my endorsement of the Choice Principle commits me to falsely claim (3), that only preferences constrain choices. But the Choice Principle says that preference determines which of the available options are/should be chosen. For an internalist, 'available options' must be read subjectively, as the alternatives the agent believes are real options for her. So interpreted, the Choice Principle *does* allow that an agent's choice depends on what options she considers available to her, not just because she cannot choose what is not available but also because her preferences over these alternatives may depend on how they are conceived by her.

If this is correct, then the Choice Principle should not be abandoned, but supplemented with an account of how decision problems should be framed.[1] Thoma makes two interesting suggestions in this regard. Firstly, that formulating decision problems requires reference to non-preference attitudes such as beliefs *and* to attitudes to objects that are neither actions nor outcomes: in particular, to properties. And, secondly, that there are rationality conditions relating attitudes to properties to attitudes to prospects. I agree with both claims: in many cases it is our preferences for properties that explains our preferences for fully described outcomes. It is, for example, my preference for sweet fruits (over sour ones) and my belief that this pear before me is sweet that explains my desire for it. And this explanation implicitly depends on it being the case that my preferences for properties impose rationality constraints on my all-things-considered preferences (this point is familiar from Pettit (1991)).

This being said, since an attitude to a property can be captured by an attitude to a proposition, it seems to me that all the ingredients for such an account of the role of property preferences are already in place in my book. When I explain my desire for a pear in terms of my attitude to the property of sweetness, it is the sweetness of pears (or, more broadly, of fruits, or even of edibles) that explains my desire, not an 'unattached' property of sweetness. Pears being sweet is something that is captured by a proposition: the set of worlds in which the pears are sweet. The property attitudes that explain my desire for the pear are attitudes to propositions like this one. And it

[1] Thoma herself recognises that the judgementalist could respond in this way.

will follow from the treatment of attitudes to propositions that I give that my attitudes to properties so-conceived will constrain my attitudes to alternatives that instantiate them. Note that it does not follow on this account that I will always desire the pear in front of me just because I prefer my pears sweet, even when I believe this pear to be a sweet one. The pear may also be discoloured and I may also dislike the property of discolouration in fruit. So the relevant attitudes to look to for explanations in this case are those directed at the proposition that is the intersection of the pears-being-sweet propositions and the fruits-being-discoloured propositions.

In conclusion, I fully concur with the importance of exploring how all-things-considered preferences over options depend on attitudes to the properties that they instantiate. But such an account is best viewed as an enrichment of the Rationality Hypothesis and so would supplement the Choice Principle, rather than replace it.[2] Once an agent has formed all-things-considered preferences over a set of options in the light of the properties of their outcomes that she considers relevant, she *should* choose the option that she most prefers (as required by the Choice Principle). If this option is not in fact best, given her beliefs and her attitudes to properties, then her *preferences* must fail to reflect these attitudes and so she must have violated one of the requirements of rationality.

2 Conditional desire (Joyce)

Over the years since the publication of his ground-breaking book *The Foundations of Causal Decision Theory* and my review of it, Joyce and I have exchanged views on numerous occasions on the nature of supposition and its role in reasoning and decision making.[3] There is, as he says, much that we agree on; perhaps most significantly he has persuaded me of the correctness of causal decision theory. But there is also some residual disagreement relating mainly to the role of suppositions in evaluating actions and about the kinds of judgements that can provide foundations for a decision theory. Since both issues draw on disagreement about the best expression for desirability under the supposition that some proposition is true, let me start by saying something about the motivation for my approach to this topic.

To keep things simple, let's focus on the case of evidential supposition. When we suppose that, as a matter of fact, it will rain tomorrow, we put ourselves in a judgemental state in which we look at the various possibilities in the light of this fact. The supposed truth of it raining provides, as it were, the backdrop for judgements of both the credibility and the desirability of other propositions. For this reason, I use measures of conditional credibility and desirability that factor out the credibility/desirability of the condition assumed true. This feature is exemplified by the normalisation of them with respect to the tautology, with the conditional probability of the tautology always

[2] The Rationality Hypothesis says that that agents are rationally required to choose the option(s) that maximise the expectation of benefit, conditional on the choice.

[3] Joyce (1999), Bradley (2001).

equalling one on my account and its conditional desirability always equalling zero.[4] Together with the usual consistency requirements this normalisation yields, in the case of evidential supposition, the following expressions for conditional desirability ($V(\cdot|A)$) and probability ($P(\cdot|A)$), given that A:

$$V(B|A) = V(AB) - V(A)$$

$$P(B|A) = P(AB)/P(A)$$

Both expressions define what Joyce calls a *relative* measure of an attitude. The conditional probability of getting wet given that it will rain, for instance, doesn't express anything about the prior credibility of rain—only about the probability of getting wet from the rain relative to the probability of it raining. Analogously, my proposed expression for the conditional desirability of getting wet, given that it will rain, doesn't express anything about the desirability of rain. Rather it measures the desirability difference that getting wet makes when it rains.

Joyce (2021) argues, and I agree, that the notions of suppositional belief and desire expressed by relative measures are not the only ones of potential interest. For one can also define *absolute* notions of suppositional belief and desire such that the probability or desirability of a proposition under the supposition of A measures not its probability or desirability in the light of, or relative to, the supposed truth of A, but simply the probability or desirability of the state of affairs that holds if both the proposition and A is true, or would hold if both were. Joyce opts for just such an absolute measure of suppositional desirability, but retains a relative measure of suppositional probability (conditional probability)—a combination that I find odd.

What is at stake in the choice between absolute and relative measures of suppositional desirability? Joyce argues that adopting an absolute measure allows him to make sense of the kind of deliberation involved in choice and in particular the kind of cross-suppositional comparisons that he takes to be essential to it. Consider his example of the choice between going birding and going surfing in the light of one's uncertainty about whether one's friend will bring binoculars or a wetsuit. He and I are agreed that birding should be chosen over surfing just in case the expectation of desirability on the supposition of birding exceeds the expectation of desirability on the supposition of surfing, where these expectations are calculated relative to a suppositional probability P^*_A measuring the probability of propositions on the supposition that A is or were true. Where we disagree is over how to interpret this quantity and what sort of reasoning is required to determine its value.

Joyce identifies the choice-worthiness of A with the absolute suppositional desirability of the status quo induced by A-ing, denoted in his paper in this volume by J^*_A (T^*). He claims moreover that cross-suppositional comparisons are essential to making choices. The choice between birding and surfing for example requires a comparison, he says, between the desirability of your friend bringing binoculars (or a wetsuit) on

[4] This normalisation just follows from the fact that the *unconditional* probability/desirability of a tautology is one/zero and the usual requirement that conditional probability/desirability functions satisfy the axioms of unconditional probability/desirability.

the supposition that you choose birding and on the supposition that you choose surfing. But this is demonstrably not the case because, on Joyce's theory, for any action A, J^*_A (T*) $= \sum_i V(w_i).P^*_A(w_i)$, where the w_i are the different possible final outcomes. So we can express the relevant quantities in his example as follows:

$$J^*_{bird}(\mathrm{T}^*) = V(bird, binocs).P^*_{bird}(binocs) + V(bird, wetsuit).P^*_{bird}(wetsuit)$$

$$J^*_{surf}(\mathrm{T}^*) = V(surf, binocs).P^*_{surf}(binocs) + V(surf, wetsuit).P^*_{surf}(wetsuit)$$

Notice that no cross-suppositional comparisons are involved in determining the values on the right-hand sides of these equations. So one can assess whether or not $J^*_{bird}(\mathrm{T}^*) > J^*_{surf}(\mathrm{T}^*)$ without them.

This example brings out our broader disagreement about how to think of choices between acts. For Joyce when we compare act A to act A^*, we calculate the desirability of the status quo on the supposition that we perform A and compare that to the desirability of the status quo on the supposition that we perform A^*. So we make a comparison of one object (T*) across two different suppositions. In contrast, I claim that we compare the desirability of two different objects: that of the situation obtained by A-ing to that of the situation obtained by A^*-ing. Contrary to what Joyce claims, no cross-suppositional desirability comparisons are required for this because we simply compare the desirability, *from our current perspective*, of the two different situations that we (expect) would be achieved by A-ing or A^*-ing. And so the quantity measured by $J^*_A(\mathrm{T}^*)$, as he defines it, is best thought of as the desirability of the state of affairs we expect to achieve by A-ing (or the expected gain in desirability in virtue of A-ing) and not as a suppositional desirability.

The claim that evaluation of action requires no cross-suppositional comparisons of desirability does not depend on rejecting the possibility of making these kinds of comparisons. But in my book I do argue against seeking foundations for decision theory in cross-suppositional *preference* comparisons of the form $B|A \succsim D|C$ (typically glossed as a preference for B given that A over D given that C). My objection was essentially that A and C serve in this expression to determine different standpoints from which the prospects B and D are to be evaluated, not as objects of preference, and that an evaluation cannot be made, at a single moment, from two different standpoints.

My rejection of cross-suppositional comparisons of this sort puzzles Joyce for a couple of reasons. Firstly, he thinks we are in fact capable of making them. And, secondly, he thinks forbidding such comparisons makes it difficult to explain the use of suppositions in the assessment of actions. I have already shown that the second concern is misguided, but I grant his first point. I should clarify however. I did not (intend to) claim that we cannot make cross-suppositional comparisons of probability and desirability. We can. For instance, we can compare the evidential conditional desirability of B given that A to that of D given that C by computing $V(B|A)$ from $V(AB) - V(A)$ and $V(D|C)$ from $V(CD) - V(C)$ and seeing which is greater. What I don't think we should do is take such comparisons as primitive and use them to derive numerical desirabilities, because I don't see that we can make the comparisons without already having at least rough numerical desirabilities to hand. So it is only in the context of the project of supplying foundations for cardinal measures of both desirability and choice-worthiness that I reject use of cross-suppositional comparisons.

3 Conditionals (Huttegger & Rothfus; Hájek)

Conditionals and suppositional reasoning of the kind discussed in the previous section are closely related. Suppose I want to be first in line for a new show but would rather not leave before finishing my dinner. In thinking about whether to leave right away (without dinner) I consider whether, were I not to, I would still arrive in time for the show. The opinions I form can be reported using conditionals like 'If I were to leave later, I would be late for the show' and 'If I were to leave now, I will be first in line'. Since it matters to me whether or not I can both eat my dinner and make the show, I want to get the right answer to the question about what would happen if I were to leave later. It is difficult to explain why I should be interested in settling the question if there is no truth to the matter.

The semantics for conditionals developed in Bradley (2012) takes this truth-aptness of conditionals at face-value and embeds it within a broadly suppositional theory of conditionals in the tradition of Adams (1975), Stalnaker (1968) and McGee (1989). The crucial feature of this account is that the role of the truth makers for conditionals are played by counterfacts: facts about the worlds picked out by the antecedent of the conditional being evaluated and not (typically) facts about the actual world. Propositions are therefore modelled not as sets of possible worlds, but as sets of ordered sets of possible worlds, where the latter specify the facts under all relevant suppositions.

At the most general level this semantics imposes no constraints on how the facts and the counterfacts may combine. But it is possible to distinguish the truth conditions for indicative and subjunctive/counterfactual conditionals by imposing constraints on the relationship between them that are characteristic of the different kinds of supposition associated with these two types of conditional. From this one can derive a number of their distinguishing properties; properties that explain the different roles that they play in our thinking and in a discourse. These include Adams' famous thesis that rational degrees of belief in conditionals are conditional degrees of belief, something widely reckoned to be true but which has proven very difficult to accommodate within standard semantic theories.

It is baked into this account of conditionals that there is a close relationship between their semantics and the use to which they are put in the kind of suppositional reasoning that we employ in deliberating about what to do. So unsurprisingly they serve in my book to support an account of deliberation aimed at choice of action. But conditionals also play a number of other roles in DTHF. They are used to (re-)individuate the outcomes of actions to capture what would have happened if uncertainty had been resolved differently, something that Allais' paradox (Allais, 1953) suggests is a matter of concern for agents. And they provide for a Savage-style representation of actions as conjunctions of indicative conditionals (that specify the outcome of the performance of the action in each state of the world), thereby allowing for a derivation of Savage's theory within the extended Jeffrey-framework developed in my book.

Huttegger and Rothfus' (2021) paper provides a fourth application, to the modelling of planning and sequential choice within this extended Jeffrey-framework. In their model of sequential choice, conditionals are used to express the content of plans at natural nodes in a decision tree, by specifying for each eventuality open at that node what will be chosen if that eventuality transpires. A plan is dynamically consistent only if it can permissibly be chosen at all nodes reached in the course of its implementation. Remarkably Rothfus and Huttegger are able to show that choice of plan in accordance with desirability maximisation is dynamically consistent if the conditionals that express plans have just the semantic properties that, in my theory, characterise indicative conditionals. Furthermore, although satisfaction of these properties is not necessary for base-line dynamic consistency, it is so for their stronger condition of preferential stability: that a preference for one plan over another at some node should not be reversed at any downstream node obtained under both plans.

The question that naturally arises is: what properties of conditionals are necessary and sufficient for the dynamic consistency and preferential stability of choice in accordance with causal decision theory? Addressing it is not straightforward since we lack an account of the causal expected utility of a plan, qua conjunction of conditionals, at a natural node. My suggestion would be that it is the average of the causal expected utilities of the possible actions that could be performed at that node with the weight on an action being given by the probability of the eventuality upon which its performance is contingent. But clearly more work needs to be done here.

These applications provide explanations of the role of conditionals in decision making and, more generally, in thought and talk, thereby giving indirect support for my account of conditionals. Hájek (2020) grants this, but thinks that the metaphysical cost of postulating the existence of counterfacts is too high a price to pay for these explanatory benefits. It isn't that Hájek believes that there are no counterfacts—just that there aren't that many of them. In particular, he denies that to every conditional there corresponds a counterfact that makes it true or false (a thesis he dubs Counterfactual Plenitude). Nor, it should be emphasised, does he deny that conditionals have truth-values. Rather, he thinks most conditionals are false; something that follows in his view from the fact that there is typically no way of settling the question of whether the consequent of the conditional would have been true had its antecedent been.

Now Hajek's main target in his paper is Stefánsson's (2018) version of this view and his use of my semantics to underwrite his rejection of the claim that most conditionals are false. And I think he incorrectly attributes some of Stefánsson's views to me. Although I take counterfacts to be the truth-makers for conditionals, I am not committed to what he calls Primitive Counterfacts Realism: neither to counterfacts being primitive nor to realism about them. In the first place, the semantic role that counterfacts play in my theory doesn't determine any particular view about their metaphysical status (and I largely stayed clear of expressing one). Secondly counterfacts are not even *semantically* primitive in my theory. The only primitives are possible worlds, from which the ordered sets of worlds are constructed for the purposes of defining truth-conditions. Consequently, my theory has no more metaphysical commitments than standard possible world semantics. If there are other possible worlds than the actual one, then counterfacts exists. If they are real then so are the counterfacts, etc.

I am of course committed to the existence of counterfacts and Hájek thinks that this is a mistake. I don't, on the whole, see much value in directly engaging in arguments for and against their existence however (any more, say, than arguing about whether or not imaginary numbers exist). The primary question surely is whether in postulating their existence we are able to explain and/or rationalise the role conditionals play in thought, talk and choice in a way that we cannot by denying their existence. Or whether, on balance, the explanation afforded by so-doing is sufficiently better (in terms of strength, simplicity and fit perhaps) than the alternatives on offer. DTHF shows what you can do with a theory which embraces counterfacts. What theory that does without them is even in the same ballpark in terms of strength, simplicity and fit?[5]

Hajek's main reason for rejecting Counterfactual Plenitude is that we typically have no way of determining what the counterfacts are. He is surely right about this. If we had tossed a coin yesterday evening to settle what to have for dinner, would it have landed heads or tails? We usually cannot say. But questions about what exists cannot be settled by what we are able to discern. It is often difficult or even physically impossible to discern what is occurring very far away or what will occur far into the future, but this not in itself a decisive reason for doubting that there is some fact (some future or distant fact) of the matter as to what is or will occur.

So why does Hájek think that these epistemic considerations are decisive in the case of conditionals? He asks whether it's true that 'If Sophie had gone to the parade, she would have seen Pedro dance', and concludes that it is not, because there are "so many possible relevant ways for Sophie to have gone to the parade and not seen Pedro dance: by standing at the wrong place, getting there at the wrong time, looking away at the crucial time, …" (Hájek, this volume). That there are many possible ways in which Sophie could fail to see Pedro dance is reason, Hájek thinks, for saying it must be false that she would have seen him dance. But equally there are many possible relevant ways for Sophie to have gone to the parade and seen him dance, so it must also be false that 'If Sophie had gone to the parade, she would not have seen Pedro dance'. We are led to conclude that most counterfactuals are false.

It seems clear to me that this argument is mistaken. If Sophie had gone to the dance she would either have seen Pedro dance or she would not have. We don't know which is true—that she would have seen him dance or that she would not have—but we do know that one of them is. What I think is correct is that there is no fact about the *actual* world which determines which it is. From this I draw support for my claim that counterfactuals are not true at stand-alone possible worlds but at ordered sets of them. 'If Sophie had gone to the parade, she would have seen Pedro dance' is true at those ordered sets of possible worlds in which the possible counteractual world under the supposition that Sophie was at the parade is one in which she saw Pedro dance. Its false at those in which the counteractual world is one in which she failed to see him dance. No ordered set contains both or neither, since at every possible world in which Sophie goes to the parade she either sees Pedro dance or she does not, and not both.

[5] A referee offered a possible answer: an error theory that takes users to behave as if counterfacts existed, event though they don't. Such a theory would indeed fit the evidence and arguably gains metaphysical simplicity at the expense of introducing psychological complexity.

In summary, Hajek's mistake is to infer from the true claim that, for most ordinary counterfactuals, no set of facts about the actual world suffices to determine whether the counterfactual is true or not, to the false conclusion that there are no counterfacts. The choice we face, rather, is to infer *either* that there are no counterfacts and that most counterfactuals are not truth-apt *or* that the truth or falsity of counterfactuals is not determined by the facts but by the counterfacts. It is the latter path that I chose.

4 Individuation and chance (Mongin & Baccelli; Goldschmidt & Nissan-Rozen)

In their wide-ranging paper, Mongin and Baccelli (2020), (1) contest that the Bolker-Jeffrey theory of decision has the advantages I claim for it; (2) argue against certain uses of the 'redescription strategy' to defend expected utility (EU) theory, and (3) cast doubt on the version of this strategy that I use to defuse the Ellsberg paradox (Ellsberg, 1961). I will not say much about the first, as their discussion of the merits of using the Bolker-Jeffrey theory as the basis for reorganising normative decision theory is nuanced and insightful and I agree with many of their points. In particular, I agree that the use of an atomless algebra of propositions seems to be in tension with the project of developing a theory of rationality for bounded agents, especially when combined with the assumption of complete preferences. Indeed, this explains (in part) why I think that it is so important to develop a theory that allows for incomplete preferences, a project developed in later parts of the book.

One point of disagreement however about the first issue. Mongin and Baccelli approve, I think, of the ambition of recovering Savage's expected utility theory from the Bolker-Jeffrey one. But they doubt that I have succeeded in doing so, because Bolker's uniqueness theorem is not strong enough for an SEU representation. In the book I rely primarily on Joyce's representation theorem to overcome this problem, but Mongin and Baccelli dislike the reliance on a second primitive in Joyce's framework, something which they point out is a departure from the main tradition in decision theory. This is all true. But I have proved a 'traditional' representation theorem for an enriched version of the Bolker-Jeffrey theory elsewhere (Bradley, 2007) and point out in the book that one could equally well work with it rather than Joyce's. So I think that the derivation of Savage's theory within the enriched Bolker-Jeffrey framework is solid.

Let's turn to the second issue. Defendants of expected utility theory often deal with putative counterexamples to their theory by recourse to the 'redescription strategy', i.e., by arguing that if the elements of a decision problem are properly described, taking into account all that is relevant to the agent, then the counterexample is revealed to be no such thing. Mongin and Baccelli argue that such use of the redescription strategy should be neither indiscriminate nor lazy. Merely refining the outcomes in the Allais paradox to include feelings of regret or disappointment is an example, they claim, of such laziness since it serves merely to disarm the particular counterexample without offering any resources for predicting choices in similar circumstances. On these methodological grounds they prefer the route taken by non-EU theories, of proposing models which explain why we observe the patterns of choices that we do.

As Mongin and Baccelli acknowledge, I don't actually endorse the instances of the redescription strategy that they dislike and indeed draw on their point (often ignored by philosophers) that the Allais preferences are primarily a problem for the von Neumann and Morgenstern theory and only indirectly (i.e., in combination with some other assumptions) for Bayesian versions of EU theory. More importantly, I think that they are entirely correct in arguing against indiscriminate uses of the redescription strategy. But under what conditions is such a strategy acceptable? Redescriptions should be based on generalisable claims about what factors matters to an agent's decisions that can be fruitfully employed to support predictions and explanations of their choices in a variety of situations. Redescriptions of states of the world will be useful, for instance, if they pick out features about which agents are uncertain and which they believe affect the outcomes of their choices, while redescriptions of consequences must pick out features that agents plausibly care about. These are not terribly precise criteria and it might take a while to determine whether a proposed redescription meets them or not. But the claim that individuals care not just about intrinsic properties of their circumstances (e.g. their income level) but also relational ones (e.g. how well off they are compared to others around them) is the sort of hypothesis that does meet the criteria, whereas I think the jury is still out as to whether the claim we experience feelings of regret does (the challenge being that of saying how such feelings depend on the formulation of the decision problem).

Let me turn to the third issue raised by them: my approach to the Ellsberg paradox. In the book I show that the pattern of preferences exhibited in the Ellsberg paradox are justifiable within Bayesian decision theory, provided that we accept that agents are concerned not just about final monetary outcomes but also their chances of obtaining these outcomes. For if they are, then the outcomes of the different prospects in Ellsberg's set-up should be re-individuated in terms of the chances of monetary outcomes. This redescription suffices to show why no violation of the axioms of Bayesian decision theory (and in particular of the Sure-thing principle) is implied by the pattern of preferences identified by Ellsberg.

Goldschmidt and Nissan-Rozen (2021) correctly interpret my project here of enriching Jeffrey's framework by taking chances of goods and bads themselves to be objects of agents' attitudes "as a way to pursue the re-individuation strategy while avoiding the threat of triviality" (this volume). Is this threat successfully avoided? I think that it quite clearly is. Firstly, the hypothesis that some agents do in fact value the chances of outcomes is empirically fruitful, as evidenced by the explanation it provides of a number of phenomena unconnected to those exhibited by Ellsberg's paradox. One is our concern for the fairness of procedures which, I argue, is in part a matter of sensitivity to the chances that these procedures confer. A second is the value we attribute to succeeding at tasks that are less-than-sure to succeed, something that is hard to explain if the relatively low chance of success at these tasks did not enhance their value.

Secondly, the hypothesis is theoretically fruitful. This is amply illustrated by Goldschmidt and Nissan-Rozen's paper in which they show that it implies interesting constraints on value: in particular, that the desirability value of a risky prospect or lottery can be represented as the sum two of expectations, respectively of the prospect's intrinsic value and of its instrumental value. That these two kinds of value can be so neatly separated is both surprising and significant. As is a corollary of this, that both the expected intrinsic value and the expected instrumental value of a set of mutually exclusive and exhaustive prospects must equal zero. The non-triviality of the framework is thus clearly established by them.

Mongin and Baccelli don't say otherwise directly. Instead, they claim that the redescription supported by my proposal is not the natural one and propose instead that outcomes be individuated in terms of the chances of drawing a ball of a given colour from an urn of a given composition. But why should agents care about this? Ball colours are (typically) of no concern to us in themselves; it is only because in the Ellsberg choice problem ball colours are associated with monetary amounts that we track them. So re-individuation of the *states* in terms of the chances of colour draws has some merit, but not re-individuation of the *outcomes*. Because they now omit from the description of the outcomes precisely what it is that agents might plausibly care about, they recover the violation of the Sure-thing principle. But this simply serves to confirm my original claim: that the paradox arises only if we don't take into account everything that agents care about.

5 Unawareness (Mahtani)

Decision Theory with a Human Face attempts to develop a theory of rational decision making tailored to agents that are limited in their cognitive resources. Such agents face decision problems without full awareness of all relevant considerations and without having a settled opinion on all those that they are aware of. And the sorts of attitude changes they can undergo are not restricted to updates in the light of newly acquired information: they also include suspension of opinion, formation of opinion through deliberation and inference (rather than information acquisition), and the gain in (and sometimes loss of) awareness of the possibilities they face. The standard Bayesian theory of conditionalization is not adequate as a model of such changes and so there is a need to supplement it with further principles.

Mahtani's (2020) paper is concerned with the principles applying to situations in which an agent is initially unaware of one or more possibilities. In my book I define such a situation as one in which judgements about certain possibilities are not available to an agent when deliberating about what to think or do. Mahtani is critical of this characterisation on the grounds that one may be aware of, and have attitudes, to possibilities at some moment of time, even if these are not available to one's consciousness at that time. I don't deny this. On the contrary, 'availability' is something that comes in grades, spanning from cases in which a possibility is immediately available to judgement, through those in which they are accessible given enough time and effort, all the way to those in which possibilities are essentially inaccessible without some external intervention. Perhaps the word 'unawareness' should only apply to the latter cases and the others are better characterised as states of inattention. In any case the differences are important and Mahtani is right to ask for greater clarity.

What I propose as a starting point is a four-tiered model. In the inner core are one's attitudes to the prospects to which one is paying attention. In one layer out from them are the prospects to which one has an attitude but to which one is not paying attention (the objects of implicit attitudes). In the next, those of which one is aware but to which one is not paying attention and towards which one has not formed an attitude. Finally in the outer layer are those prospects of which one is strictly unaware. Right now, for instance, I am paying attention to the question of what rationality requires of us, but not as to whether there is milk in the fridge. On this question however I do have a belief (now brought into my consciousness by the process of writing), namely that there is. On other questions, such as whether most residents in Kathmandu keep milk in their fridges, I have no opinion (though this entirely conceivable possibility, now having been brought to my attention, is one on which I could form an opinion with a little effort). Finally, there are no doubt all sorts of possibilities that I cannot at present conceive of, but of which I cannot of course provide an example! (Examples applying to others are easily found however: surely, for instance, ancient Egyptians had no opinion on which of the current smartphones has the longest battery life and could not have formed an opinion because they lacked the conceptual resources required to conceive of smartphones.)

Mahtani focuses her discussion on the intermediate layers, consisting of possibilities outside of my immediate attention, but firmly within the realm of conceivability. These are just the sorts of things that one gives attention to if prompted by the circumstances and about which one has little difficulty in forming an opinion. I agree with her that cases like these are often discussed in the literature on unawareness. But that doesn't mean that there aren't cases in which new possibilities are brought to one's attention about which one has no information at all. Moreover, it seems to me that it is better to start with such cases (i.e. those involving no new information), so that one can better separate the effects of a change in the domain of one's awareness from changes that result from the gain of new information and from making new inferences.

Suppose for instance that you are not aware that more than two horses can be entered for a race and that you have invested a good deal of effort in gathering information and forming a view as to how probable it is that either of the two entrants, Speedy and Steady, will win. Suppose that at the last minute you are told that there is another horse in the race about which you know nothing. How should this discovery affect your credences? In my book I apply three principles to address this question.

1. *Success*: Your new domain of awareness should contain the possibility of this 3rd horse running (and winning).
2. *Consistency*: Your new credal state should be consistent.
3. *Conservatism*: If what you learn does not give you reason to revise some aspect of your credal state, then you shouldn't revise it.

It is the last of these conditions that does most of the work. For when conjoined with the thought that merely learning of new possibilities does not change the balance of reasons for and against old possibilities it implies that the relative probability of Speedy and Steady should not change just because another horse has been entered.[6] This is not to deny that the entertaining of new possibilities can trigger new enquiry or inference that ultimately leads to revisions of ratios of old probabilities (as the examples of Steele and Stefansson 2020 show). But this should be modelled separately, downstream from the initial response to the growth in awareness.

[6] Unless, of course, the performance of the horses depends on how many are running, in which case we need to model the situation as one in which there is new awareness *and* new information.

The view that a rational agent will, in situations in which they are made aware of a new possibility but acquire no further information about it, adopt new credences that preserve the ratios between credences in all propositions in her old algebra is what Mahtani calls Reverse Bayesianism (hereafter RB).[7] In her paper she presents an objection to it based on a pair of contrasting examples, one involving an extension to the set of possibilities and one involving a refinement of it. She also argues that the problem of awareness is misconstrued by me and that a solution to it is available within a broadly Bayesian framework by adopting a dispositionalist account of credence.

Mahtani's objection to RB is completely sound, but her examples raise complicated issues about how propositions are to be individuated when we shift from one algebra of possibilities to another that are not immediately apparent. So let's put her objection in more abstract form, which will also serve to show how general it is. Without loss of generality consider a 2-element partition $\{L, R\}$ of the state space e.g., the partition {Landlord, Tenant}. Suppose that you become aware of a new possibility M (e.g., Other}, distinct from both, so that the $\{L, R\}$ partition must be extended to the $\{L, R, M\}$ one. Then RB requires that the relative probability of L to that of R stays the same. But an extension of the $\{L, R\}$ partition by M is also a refinement of the $\{L, \neg L\}$ partition by M and $\neg M$, since $R = \neg L \wedge \neg M$, $L = L \wedge \neg M$ and $M = \neg L \wedge M$. So RB also requires that the probabilities of L and $\neg L$ stay the same. These two constraints are consistent only if we assign probability zero to M, which trivialises the whole thing.

Mahtani argues that extending attitudes to possibilities of which one was previously unaware requires more information than Reverse Bayesianism draws on. This information can come, she contends, from our prior unconscious attitudes to the propositions that enter into our awareness. Indeed she suggests that the whole problem of belief change under growing awareness as I have posed it should be rejected because it presupposes that one cannot possess an attitude to propositions of which one is not conscious. In contrast, on the dispositionalist view of attitudes (which she recommends), attitudes are constituted by behavioural dispositions of one kind or another: to assent to an utterance, to accept a proposition or to bet on its truth, when prompted to do so. Since one may be in possession of the relevant disposition without being aware of its object, the dispositionalist doesn't think the question of how to extend one's attitudes to new objects ever arises.

Mahtani is right in saying that the problem of awareness growth cannot be solved without more information than RB uses, but I don't think that what she proposes constitutes an adequate solution to it. I cannot survey all possible dispositionalist accounts that might support her account, but I think they will all fail for much the same reason. So let's just focus on the version that says that to believe X to degree x is to be disposed to bet on X at odds x when offered the opportunity to do so. Clearly X does not have to be available to consciousness for this criterion to be applied and so a dispositionalist can argue that the problem of what belief to adopt to a proposition when one first becomes aware of it doesn't arise: one already has a belief! Only if its coming into consciousness brings new relevant information in its wake is anything required

[7] This term was introduced by Karni and Viero (2013). Their proposal and mine are very similar in spirit, but developed in rather different frameworks.

of one. And in this case the appropriate response (says Mahtani) is to conditionalize one's degrees of belief on the new evidence. So no modification of the Bayesian model is required in order to handle cases of new awareness.

But how can a disposition to bet at some odds serve as a justification for having the corresponding degree of belief? Imagine that I have never heard of a quetzal. Suppose also that I am disposed to accept bets at even odds on the truth of any proposition concerning objects of whose existence I had been previously unaware that have a name beginning with the letter 'q'. So I am disposed to bet at even odds, for instance, on the proposition that quetzals can fly. But what of it? It is neither correct to say that in virtue of this disposition I have a credence greater than 0.5 in quetzals being able to fly, nor that I should adopt this credence when I become aware of their existence. Such betting dispositions only serve as plausible markers for credences in circumstances in which someone is able to rationally assess the expected value of the bet—which in these cases they cannot.

So Mahtani's solution is not satisfactory. I prefer the one proposed by Roussos (2020), which starts with the observation that refinement and extension of the set of possibilities one conceives of involve different embeddings of one's old algebra into the new one. To illustrate consider the two lattices respectively based on the two sets of basic possibilities $\{L, R\}$ and $\{l, m, r\}$. There are several ways in which the first 4-element lattice can be mapped onto the second 8-element one, each of which determines a consistent application of RB. There is, for example, the 'extension' mapping $\mathcal{E}$ such that $\mathcal{E}(L) = l, \mathcal{E}(R) = r, \mathcal{E}(L \vee R) = l \vee r$ and $\mathcal{E}(L \wedge R) = l \wedge m \wedge r$. And there is the 'refinement' mapping $\mathcal{R}$ such that $\mathcal{R}(L) = l$, $\mathcal{R}(R) = m \vee r$, $\mathcal{R}$ $(L \vee R) = l \vee m \vee r$ and $\mathcal{R}(L \wedge R) = l \wedge m \wedge r$.

Given the extension mapping, RB says that relative probabilities of l and r should equal those of L and R. Given the refinement mapping, on the other hand, it says that relative probabilities of l and $m \vee r$ should equal those of L and R. It should be clear therefore why RB generates an inconsistency if we apply it simultaneously to multiple embeddings. In the counterexample to RB presented before for instance I applied *both* the extension and refinement mappings that have just been characterised. But this was a mistake: in adopting a new algebra of possibilities following the growth of awareness, I should not treat both r and $m \vee r$ as the counterparts to the element R of the old algebra.

To apply RB one must first choose an embedding. Sometimes an extension embedding seems more sensible, sometimes the refinement one. In Mathani's first case, involving the landlord and tenant, new awareness of the possibility that some other person might be around is intuitively a case of extension. On the other hand, her second 'coin' case, is intuitively one in which we refine the tail-possibility. So the two situations are quite different and this fact should be represented in a satisfactory account of awareness growth.

There is clearly still much work to be done in providing one. First, RB needs to be reformulated so that it takes as the input to the attitude revision procedure not just an old and new algebra, but also an embedding of the former into the latter. Second, RB needs a more substantial account of the grounds for embedding in one way rather than another. And finally, there is the task of showing how this account of pure awareness growth combines with other mechanisms for attitude change and in particular those based on acquisition of evidence and on inference.

6 Ambiguity (Steele)

In DTHF I argued that in circumstances of informational poverty agents may reasonably eschew adoption of precise degrees of belief for all relevant contingencies and yet nonetheless act as if they had precise beliefs when making a choice. The idea was that she could in effect temporarily adopt a probability for decision purposes but without long-term commitment to it, so that the adopted probability did not constrain subsequent belief revision. Steele (2020) argues that such 'as-if' Bayesianism is unsatisfactory because it doesn't protect the agent from the possibility of sure losses in diachronic decision problems.

The dilemma Steele sets up is the following. Consider a decision problem, such as the diachronic version of the Ellsberg paradox, that requires an agent with imprecise credences to make a first decision at time t_0 and then a second one at later time t_1 when she has received some new evidence. Suppose that at t_0 the agent selects a precise 'prior' probability for decision purposes in accordance with her adopted rule of choice for proxy beliefs and maximises expected utility relative to this probability. Suppose that she does this again at t_1, by first revising her initial imprecise set of credences by point-by-point conditionalization, selecting one from the updated set in accordance with her choice rule, and then maximising expected utility relative to the chosen probability. By standard arguments if the probability function chosen at t_1 is not equal to the prior chosen at t_0, conditioned on the newly acquired evidence, then she will be vulnerable to sure loss.

Steele shows that neither of the two rules I consider—MaxEnt and linear averaging—satisfy this requirement. In general, as she points out, only a rule of proxy choice that satisfies the External Bayesianity condition will ensure that it is respected.[8] So her argument leaves us with three options.

1. We can adopt a rule of proxy selection that satisfies External Bayesianity (geometric averaging being the most salient example).
2. We can deny that the possibility of sure losses counts decisively against a decision method.
3. We can deny that an agent should handle diachronic decision problems in the manner sketched above.

All three routes seem open to me. Geometric averaging is a form of averaging with many advantages and perhaps the guarantee of immunity to sure loss is sufficient reason to use it for proxy selection. But I don't personally regard such immunity as decisive: it is just one consideration to be weighed against others. If one's preferences change over time, this can make one vulnerable to exploitation by someone who anticipated

[8] External Bayesianity says that, given some set of priors P, the precise posterior probability adopted on the basis of the set of posteriors obtained by conditionalization of member of P by the same evidence E should be the same as the posterior probability obtained by conditionalization on this evidence of the precise prior adopted on the basis of P.

how they change (see DTHF, p. 286). But that doesn't make it rational to stick with one's preferences just in order to avoid such a possibility.

Finally, even if avoiding sure loss is a decisive consideration, an agent can still assure it by being resolute in her choice of proxy. In particular, I think there is more to the second resolute strategy identified by Steele than she grants. When the less-than-fully opinionated agent adopts precise degrees of belief for the purposes of decision making she can reasonably do so for the entirety of the decision problem at hand, without needing to re-apply the selection rule at every node. For the idea is simply to adopt a set of beliefs to work with, recognising that although the choice of this set is not determined by the evidence she holds, it is not totally arbitrary either. That this entails that her precise credence at t_1 is different from the one that would be selected from the set of probabilities obtained by conditioning her initial imprecise credal state on any new evidence, is arguably neither here nor there.

Acknowledgements I am grateful to H. Orri Stéfansson, both for editing the symposium on *Decision Theory with a Human Face* and for his comments on this reply to the papers appearing in it, and to the two referees for their helpful comments on the submission.

References

Adams, E. (1975). *The logic of conditionals*. Reidel.

Allais, M. (1953). Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica, 21*(4), 503–546.

Bradley, R. (2001). Foundations of causal decision theory, In: Joyce, J. M. (ed) Cambridge University Press, 1999, (pp. xii+ 268). *Economics and Philosophy*, 17(2), 275.

Bradley, R. (2007). A unified Bayesian decision theory. *Theory and Decision, 63*, 233–263. https://doi.org/10.1007/s11238-007-9029-3

Bradley, R. (2012). Multidimensional semantics for conditionals. *Philosophical Review, 121*(4), 539–571. https://doi.org/10.1215/00318108-1630921

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*(4), 643–669. https://doi.org/10.2307/1884324

Goldschmidt, Z., & Nissan-Rozen, I. (2021). The intrinsic value of risky prospects. *Synthese, 198*, 7553–7575. https://doi.org/10.1007/s11229-020-02532-3

Hájek, A. (2020). Contra counterfactism. *Synthese*. https://doi.org/10.1007/s11229-020-02643-x

Huttegger, S. M., & Rothfus, G. J. (2021). Bradley conditionals and dynamic choice. *Synthese*. https://doi.org/10.1007/s11229-021-03082-y

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. M. (2021). Conditional desirability: Comments on Richard Bradley's decision theory with a human face. *Synthese*. https://doi.org/10.1007/s11229-020-02579-2

Karni, E., & Vierø, M. L. (2013). "Reverse Bayesianism": A choice-based theory of growing awareness. *American Economic Review, 103*(7), 2790–2810. https://doi.org/10.1257/aer.103.7.2790

Mahtani, A. (2020). Awareness growth and dispositional attitudes. *Synthese*. https://doi.org/10.1007/s11229-020-02611-5

McGee, V. (1989). Conditional probabilities and compounds of conditionals. *The Philosophical Review, 98*(4), 485–541.

Mongin, P., & Baccelli, J. (2020). Expected utility theory, Jeffrey's decision theory, and the paradoxes. *Synthese*. https://doi.org/10.1007/s11229-020-02691-3

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory: Issues and advances* (pp. 147–175). Blackwells.

Roussos, J. (2020) Awareness Growth and Belief Revision downloaded at: http://joeroussos.org/wp-content/uploads/2020/06/Roussos-Awareness-growth-and-belief-revision-Jun2020.pdf

Stalnaker, R. (1968). A theory of conditionals: Studies in logical theory. In N. Rescher (Ed.), *American Philosophical Quarterly (Monograph Series, 2)* (pp. 98–112). Blackwell.

Steele, K. (2020). How to be imprecise and yet immune to sure loss. *Synthese*. https://doi.org/10.1007/s11229-020-02665-5

Steele, K., & Stefánsson, H. O. (2020). Belief revision for growing awareness. *Mind, 120*, 520–1207. https://doi.org/10.1093/mind/fzaa056

Stefánsson, H. O. (2018). Counterfactual skepticism and multidimensional semantics. *Erkenntnis, 83*, 875–898.

Thoma, J. (2021). Judgementalism about normative decision theory. *Synthese, 198*(7), 6767–6787. https://doi.org/10.1007/s11229-019-02487-0

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.