Instrumental rationality without separability

Erkenntnis
https://doi.org/10.1007/s10670-018-0074-9

ORIGINAL RESEARCH

Instrumental Rationality Without Separability

Johanna Thoma[1]

Received: 20 February 2018 / Accepted: 4 October 2018

Abstract
This paper argues that instrumental rationality is more permissive than expected utility theory. The most compelling instrumentalist argument in favour of separability, its core requirement, is that agents with non-separable preferences end up badly off by their own lights in some dynamic choice problems. I argue that once we focus on the question of whether agents' attitudes to uncertain prospects help define their ends in their own right, or instead only assign instrumental value in virtue of the outcomes they may lead to, we see that the argument must fail. Either attitudes to prospects assign non-instrumental value in their own right, in which case we cannot establish the irrationality of the dynamic choice behaviour of agents with non-separable preferences. Or they don't, in which case agents with non-separable preferences can avoid the problematic choice behaviour without adopting separable preferences.

1 Introduction

We make most of our decisions in the context of uncertainty. That is, we don't know what the consequences of our actions are going to be. What does instrumental rationality require of us in the context of uncertainty? How do we act so as best to achieve our ends? The orthodox answer to this question is that we ought to be expected utility maximizers. Being an expected utility maximizer involves, amongst other things, having preferences over uncertain prospects that are separable: The evaluation of outcomes in distinct states of the world should make independent contributions to the overall assessment of an uncertain prospect. I here want to argue that instrumental rationality does not in fact require separability, and is thus more permissive than expected utility theory.

✉ Johanna Thoma
j.m.thoma@lse.ac.uk

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

Published online: 01 November 2018

There are various counterexamples to separability that show, at the very least, that preferences that violate separability are not obviously instrumentally irrational. I will here focus mostly on what is arguably the most famous, namely Allais's (1953) paradox. In the light of such counterexamples, we are in need of some compelling argument why it should be instrumentally irrational to have non-separable preferences. I take the best instrumentalist case that has been made, most notably by Hammond (1988), in favour of separability to consist in an appeal to how agents with non-separable preferences choose in some dynamic choice problems. These agents can be placed in choice situations, the argument goes, where they must choose in a way that is instrumentally criticizable. That is, they end up badly off by their own lights, by making a sure loss, or by behaving in a way that is at odds with their initial assessment of the best course of action.

While this argument has faced much criticism, there has been no agreement on what exactly is wrong with it. There are two common responses. On the one hand, there are those who think that, while agents with non-separable preferences will act in the allegedly instrumentally irrational ways, that choice behaviour is not actually irrational, because better courses of action are simply not available to those agents. This is the stance taken, for instance, by Seidenfeld (1988, 1994). On the other hand, there are those who agree that the alleged choice behaviour of agents with non-separable preferences is instrumentally irrational, but who think that those agents need not act in the alleged way, and could avoid instrumental irrationality without giving up their non-separable preferences. This is the argument made, for instance, by McClennen (1990). We are thus left with a kind of stalemate that implies that the jury is still out on the original argument: The success of the original argument depends on it both being the case that agents with non-separable preferences act in the alleged way, and that this is instrumentally irrational. There is in fact considerable support for each key ingredient of the argument, even amongst its critics.[1]

What I want to show here is that the key to resolving this stalemate is a more explicit discussion of what we take to be the standard of instrumental rationality. Any appeal to instrumental rationality must take at least some of the agent's conative attitudes to be beyond rational criticism, namely those attitudes we take to be picking out the agent's ends. Instrumental rationality is then about taking the best means to those ends. When we offer instrumentalist justifications of requirements on choice or preference, what we try to show is that agents must meet those requirements in order to best serve their ends, whatever they may be; that they might end up badly off by their own lights if they violate them. The question of the standard of instrumental rationality is the question of which of the agent's attitudes pick out her ends. Which ones are the attitudes that are beyond rational criticism, and that instrumental rationality aims at satisfying?

What turns out to be the crucial question for the success of the dynamic choice argument for separability is this: Are attitudes to uncertain prospects part of the standard of instrumental rationality, or do only attitudes to the possible outcomes of my actions count? In the first case, my attitudes to prospects assign non-

[1] For recent tentative support for the argument, see Steele (2010), and Briggs (2015).

instrumental value to prospects in their own right. That is, uncertain prospects are amongst the ends I want to achieve, and not mere means for achieving outcomes I like. In the second case, prospects are seen as having mere instrumental value, as being mere means for achieving desirable outcomes.

This paper argues that the dynamic choice argument is ultimately unsuccessful because it equivocates between different notions of the standard of instrumental rationality. Moreover, the different past criticisms of the argument rely on different understandings of the standard of instrumental rationality. If attitudes to the uncertain prospects open to the agent at the time of action are part of the standard of instrumental rationality, agents with non-separable preferences will act in the allegedly problematic ways, but we can't show that to be instrumentally irrational. If only attitudes to outcomes form the standard of instrumental rationality, we can show the allegedly problematic behaviour to indeed be instrumentally irrational, but agents with non-separable preferences need not act in that way. Either way, instrumental rationality turns out to be more permissive than expected utility theory claims.

My argument thus takes the form of a dilemma, and I am not ultimately endorsing any particular standard of instrumental rationality. Nevertheless, it follows from my argument that those who want to defend either expected utility theory or less demanding formal theories of choice under uncertainty as theories of *instrumental* rationality are well advised to defend the idea that only attitudes to outcomes form part of the standard of instrumental rationality. I will argue that we can't even defend the least controversial principles of choice under uncertainty when attitudes to prospects are part of the standard of instrumental rationality in their own right. Moreover, as far as the dynamic choice argument considered here is concerned, as long as attitudes to prospects are considered to assign only instrumental value, we can at least say that having separable preferences is *one* good way of avoiding instrumental irrationality, even if it is not the only way.

2 Expected Utility Theory and Separability

One standard way of representing the uncertainty we all face, going back to Savage (1972), supposes that there is a mutually exclusive and exhaustive set of states of the world $S_1 \ldots S_m$, which are assumed to be outside of the control of the agent. Each action $A_1 \ldots A_n$ open to an agent is then taken to lead to some assignment of outcomes $O_{11} \ldots O_{nm}$, that is, descriptions of everything the agent may care about in the consequences of her actions, to these states of the world. Each action is thus associated with an ordered n-tuple of outcomes, one for each state of the world. These ordered n-tuples are sometimes called *prospects*.[2] Since, on this picture, each action is associated with one prospect, we can alternately think of agents as choosing between actions or prospects.

[2] See, for instance, Broome (1991), p. 90 for this usage of the term 'prospect'. As I will be using the term, agents who face the same acts but assign different probabilities to states are facing the same prospects—they just evaluate them differently.

The orthodox theory of rational choice under uncertainty is expected utility theory. Expected utility theory comes in a variety of guises. But what versions of the theory tend to have in common is that agents are rationally required to maximize, or act *as if* they maximized the expectation resulting from some probability function p over states of the world, and some utility function u over outcomes. The expected utility of an action A_i (or the associated prospect) is then calculated as follows:

$$EU(A_i) = \sum_{j=1}^{m} p(S_j) \cdot u(O_{ij})$$

where do the utility and probability assignments come from? Most decision theorists, in particular those in economics, think that either just utility or both probability and utility are mere constructs that can be used to represent the agent's preferences over prospects as expected utility maximizing. Various representation theorems are supposed to show that such a representation is possible if the agent's preferences abide by a number of axioms. Under these interpretations, the rational requirements of expected utility theory are usually taken to be that one's preferences ought to abide by the axioms of one's favourite representation theorem. If one then acts in accordance with them, one behaves as if one were maximizing an expected utility function.

I will here focus on one central requirement that is characteristic of all versions of expected utility theory. And that is the requirement of *separability*. It finds expression in the axioms of various representation theorems. Roughly, the idea behind separability is as follows: Each prospect can be divided into sub-prospects, which assign outcomes to only a subset of the states that are part of the full prospects the agent faces. For agents with separable preferences, sub-prospects that don't overlap are not complementary. Rather, sub-prospects can be evaluated independently, in a way that is unaffected by what happens in all the other states that form part of the full prospects an agent faces. Moreover, they always make the same contribution to the overall assessment of the full prospects they form part of.

For instance, suppose I am thinking about whether to take the train or cycle to work today, and I think that it may or may not rain later. I can now consider the sub-prospects that cycling and taking the train lead to, respectively, in the event that it rains. That is, I am looking only at the outcomes cycling or taking the train may lead to in all of the states that involve rain today. Separability implies that my preference between these two sub-prospects is unaffected by what happens in the event that it doesn't rain. Moreover, suppose I have also considered the sub-prospects of cycling and taking the train in the event that it doesn't rain, and now make an overall assessment of whether to cycle or take the train given that I don't know whether it is going to rain or not. Separability now requires that my evaluations of the sub-prospects I face when it rains and when it doesn't, respectively, make independent contributions. For instance, if cycling and taking the train were to, implausibly, lead to the same prospect in the case it doesn't rain, and I prefer taking the train if it does, I should prefer taking the train given I don't know whether it will rain. Or, if I prefer cycling to taking the train in either event, then I should prefer to cycle given I don't know whether it will rain.

Separability is expressed in Savage's (1972) representation theorem by the sure-thing principle, and that is the version I will mostly focus on here.[3] To state the sure-thing principle, we need to define a set of events, which are disjunctions of states. Let $\succcurlyeq_E$ be weak preference conditional on event E occurring. We then require the following:

Sure-thing principle For any two actions A_i and A_j, and any mutually exclusive and exhaustive events E and F, if $A_i \succcurlyeq_E A_j$ and $A_i \succcurlyeq_F A_j$, then $A_i \succcurlyeq A_j$.

According to the sure-thing principle, a rational agent can determine her overall preferences over acts by event-wise comparison. She can partition the set of states into events, and then compare the prospects of each of her acts conditional on each event separately. If she prefers a particular act no matter which event occurs, then she should also prefer it when she does not know which event will occur.

Such a separability condition is in large part responsible for the possibility of an expected utility representation of an agent's preferences in the various representation theorems. And the expected utility representation itself has an important separability feature as well. As we just saw, in expected utility theory, the overall value of an action is represented as a probability-weighted sum of the utilities of the outcomes occurring in separate states. This means that the value contribution of an outcome in one state will be independent of the value contribution of an outcome in another state, holding the probabilities fixed. The same holds for sub-prospects.

Is separability a requirement of instrumental rationality? The next section introduces a famous apparent counter-example to expected utility theory that puts separability into question. We then consider what I take to be the most powerful defence of separability as a requirement of instrumental rationality.

3 The Allais Paradox, Static and Dynamic

There are a number of famous examples that motivate the view that violations of separability are not in fact irrational, as expected utility theory claims they are. One such example is the Allais Paradox, as first presented in Allais (1953). Would you rather have \$1 million for certain, or an 89% chance of winning \$1 million, a 10% chance of winning \$5 million, and a 1% chance of winning nothing, decided by a random draw from 100 lottery tickets? Many people choose the safe \$1 million. How about the choice between a 10% chance of \$5 million (and nothing otherwise) and an 11% chance of \$1 million? Here, most people go for the slightly lower chance of a much bigger win.[4]

Tables 1 and 2 represent the choices offered in the Allais problem. Agents with Allais preferences choose lottery B in the first choice, and lottery C in the second choice. This combination of preferences, henceforth 'Allais preferences', seems sensible. However, Allais preferences in fact violate the sure-thing principle, given a natural specification of the outcomes in terms of monetary gains.

[3] In von Neumann and Morgenstern's (1944) representation theorem, separability features in the form of the independence axiom, and in Jeffrey (1965/1983), it finds expression in the averaging axiom.

[4] See, for instance Morrison (1967) for experimental evidence that many people choose in this way.

Table 1 Allais paradox: first choice

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery A	$1 million	$5 million	$0
Lottery B	$1 million	$1 million	$1 million

Table 2 Allais paradox: second choice

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery C	$0	$5 million	$0
Lottery D	$0	$1 million	$1 million

In both choices, the two prospects to be chosen from are identical if tickets 1–89 are drawn. What matters, then, according to the sure-thing principle, is what happens if tickets 90–100 are drawn. But for these tickets, the first choice, between lottery A and lottery B, and the second choice, between lottery C and lottery D, are identical. And so, the agent should choose lottery B in the first choice if and only if she chooses lottery D in the second choice, implying that she shouldn't have Allais preferences.

One way of reconciling these preferences with expected utility theory may be to argue that the outcomes are under-described by merely the money amounts that the agent will win following some draw of the lottery. Perhaps, for instance, the agent cares about avoiding regret or disappointment, and this should be reflected in the description of the outcomes.[5] However, re-describing the outcomes to take account of disappointment and regret arguably cannot do away with the violation of separability in the Allais Paradox. Weber (1998) provides an extensive argument to that effect. In any case, even if these attitudes could explain why most people have Allais preferences, we can still conceive of an agent who cares about nothing but money in her evaluation of outcomes, and who still has the Allais preferences. Expected utility theory would declare such an agent irrational. But it is at least not immediately obvious that such an agent would be instrumentally irrational.

What can we say in favour of separability to such an agent, then? It has been pointed out, most notably by Hammond (1988), that agents with Allais preferences, or indeed any agents who violate separability, are prone to making choices in dynamic settings that leave them somehow worse off by their own lights, or otherwise prone to behaving in a way that is instrumentally criticizable. This may happen in choice settings where choices are made consecutively as uncertainty is gradually resolved. In such settings, sub-prospects that the sure-thing principle would require to be separable can be de facto separated in the dynamic structure of the decision problems, as agents decide about different sub-prospects gradually over time. And, for agents who violate separability, this can lead to patterns of choice that the agent can allegedly be instrumentally criticized for.

[5] Graham Loomes and Robert Sugden have explored both making regret and disappointment part of the description of outcomes—regret in Loomes and Sugden (1982), and disappointment in Loomes and Sugden (1986).

We can illustrate this with the following dynamic version of the Allais paradox, adapted from Machina (1989). In this dynamic version, agents only get to make a choice after they have found out whether one of tickets 1–89 has been drawn, or one of tickets 90–100 has been drawn. In the decision trees in Fig. 1, the square nodes are choice nodes, at which the agent decides whether to go 'down' or 'up'. The round nodes are chance nodes, at which chance 'decides' between the branches. Let t_1 be the time at which the uncertainty is resolved about whether one of tickets 1–89 has been drawn, or one of tickets 90–100 has been drawn. t_2, in turn, is the time at which the agent decides. At t_3, the agent finds out which ticket is drawn.

The interesting feature of this dynamic choice problem is that at the time when the agent gets to make a decision, at t_2, the rest of the tree, sometimes called the 'continuation tree', looks the same for the first and second choice, just as the sub-prospects involving tickets 1–89 we looked at in the static Allais problem. We might think that this means the agent should make the same choice in both cases. If the agent strictly prefers the prospect of receiving \$1 million for sure to the prospect that gives her a 1/11 chance of winning \$5 million, she should prefer to go 'down' in both cases. If she has the opposite strict preference, she should go 'up' in both cases. If she is indifferent between those prospects, we can add a sweetener to one of the

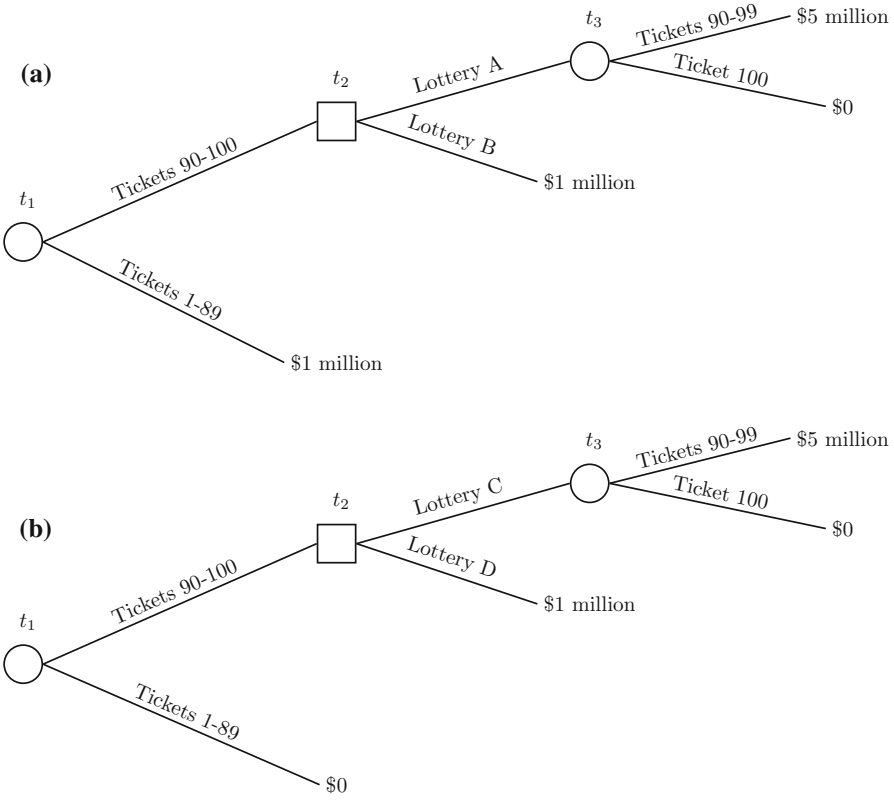

Fig. 1 Dynamic Allais problem. **a** First choice, **b** second choice

options to create a strict preference. Adding such a sweetener would presumably not alter the fact that she has the Allais preferences over the full gambles.

Suppose that in the continuation trees, an agent, call her Frieda, strictly prefers to go 'down' and get \$1 million for certain. She chooses in accordance with that preference when she gets to t_2. In that case, she will have chosen, over the course of the dynamic choice problems, to undergo lotteries B and D from the original problem. In the first choice, she receives \$1 million for certain in the course of the dynamic choice problem. In the second choice, she will run an 89% chance of receiving nothing, and an 11% chance of receiving \$1 million. But note that, in the second case, this is not a lottery Frieda would have chosen at the beginning of the decision problem, were she to make a choice upfront.

More generally, we can say that if an agent must treat like continuation trees alike, then she will end up choosing in accordance either with lotteries A and C respectively, or with lotteries B and D respectively, but not according to the Allais preferences. Similar dynamic choice problems can be constructed for any preference relation over prospects that violates separability. In all of these decision problems, sub-prospects over which the agent has non-separable preferences are de facto separated in the dynamic structure of the decision problem, by resolving some of the uncertainty involved in the original problem before the agent gets to make a choice.

In such dynamic choice problems, agents like Frieda can end up acting against their preferences over the prospects available initially. This has been held to be rationally problematic. We can distinguish two major accounts of what is supposed to be instrumentally irrational about Frieda's choice behaviour. According to the first, presented in Hammond (1988) and reconstructed in McClennen (1990), both the requirement to act in accordance with one's preferences over the sub-prospects one faces, as well as the requirement to behave in dynamic choice problems like one would were one to settle on a course of action in advance are requirements of instrumental rationality. Frieda violates the latter. And if she were to conform with the latter, she would violate the former. In fact, it can be shown that if we add a number of more technical assumptions, agents who abide by both of these requirements must be expected utility maximizers. Following Hammond, I will call this the 'consequentialist' argument. The next section argues that the consequentialist argument fails to establish that Frieda is instrumentally irrational, because it is confused about what attitudes form the standard of instrumental rationality.

The second account of what is instrumentally irrational about Frieda points out that, if we give Frieda the chance to make a costly pre-commitment to act in accordance with her Allais preferences, she will take it, and thereby run a sure loss.[6] I will argue that this is in fact the more promising strategy for establishing that Frieda is instrumentally irrational. However, the notion of the standard of instrumental rationality it commits us to also implies that Frieda could have rationally avoided running a sure loss while keeping her non-separable preferences.

[6] See, for instance, Machina (1989), Rabinowicz (1995), or Steele (2010).

4 Consequentialism

As we have seen, for Frieda the dynamic structure of the decision problem clearly makes a difference to what she will choose. If Frieda were able to make a choice and stick to it before any of the uncertainty is resolved, she would choose in accordance with her Allais preferences. But in the second dynamic choice problem, the prospect she ends up with is not endorsed by her Allais preferences. For instrumentally rational agents, we might think, the dynamic structure of a decision problem should not make a difference in this way.

The dynamic structure of a decision problem not making a difference is part of what Hammond (1988) considers to be *consequentialist* decision-making—decision-making with an eye to the consequences of one's actions only. The thought is that if an agent's choice is changed by the dynamic structure of a decision problem in cases where the attainable consequences are the same, the agent's choice must have been influenced by something other than the consequences of her actions. In his reconstruction of Hammond's argument, McClennen (1990) calls this the requirement of normal-form/extensive-form coincidence: In dynamic choice problems, the agent should choose the same as she would, were she to simply choose one course of action at the beginning of the decision problem. In the following, when I refer to 'consequentialism', I will be concerned with this requirement.

McClennen and Hammond show that this requirement, together with the assumption that the agent is 'sophisticated' in dynamic choice contexts and given some technical assumptions, implies that the agent must be an expected utility maximizer, and thus can't have Allais preferences. Sophisticated agents solve dynamic choice problems by a process of backward induction. They make a prediction of their choice at all terminal nodes, assuming that they will pick one of their most preferred prospects then. They then similarly make predictions of their own behaviour at all future choice nodes upstream from the terminal ones, each time assuming choice in accordance with their preferences over the prospects predicted to be associated with the sub-branch chosen. Sophisticated agents then go on to in fact choose, at each choice node, in accordance with their preferences over the prospects still open to them given their prediction of future choice behaviour. In our example, Frieda chooses in such a sophisticated way. And we have seen that she ends up violating consequentialism.

Hammond's proof shows that if both sophistication and consequentialism are requirements of instrumental rationality, agents like Frieda should adopt separable preferences. But are both sophistication and consequentialism requirements of instrumental rationality? I will argue that each of these principles is attractive only under conflicting understandings of what the ultimate standard of instrumental rationality is. That is, each is only attractive under different accounts of which of the agent's conative attitudes pick out her ends, which instrumental rationality requires her to serve optimally. Insofar as we find both requirements attractive, we thus equivocate between these two notions of the standard of instrumental rationality. I

conclude that we can at most justify one of these principles as a requirement of instrumental rationality.

Sophistication requires Frieda to predict what she will choose at future choice nodes, and make any present choices taking this prediction for granted. I take no issue with his part of sophistication. However, sophistication, as we characterized it here, also requires her to, at each point in time, choose the prospect that she most prefers out of the prospects available to her then. Consequentialism, on the other hand, requires her to engage in a series of choices such that the prospect she faces at the outset of the decision problem is the one, or one of the ones she prefers then, out of all the prospects she could possibly end up with through a series of choices in the dynamic choice problem. Both consequentialism and sophistication thus require Frieda to choose in accordance with some preference she has over prospects. We might think that we can thus easily defend them as requirements of instrumental rationality if we take agents' preferences over prospects to be the standard of instrumental rationality: Instrumentally rational agents are those that do well by their preferences over prospects, that bring about the prospects they most prefer. On closer inspection, however, sophistication and consequentialism are only plausible requirements of instrumental rationality on different conceptions of *which* of the agent's preferences over prospects are the standard of instrumental rationality.

Sophistication would be a straightforward requirement of instrumental rationality if, at each point in time, instrumental rationality required the agent to do well by her preferences over the prospects open to her at that point in time. Call this account of the standard of instrumental rationality *open prospects*. *Open prospects* says that at each point in time, the agent's end is to bring about the prospect she most prefers then. The best way to serve this end is to choose such that, given her prediction of her future choice behaviour, the agent expects her most preferred prospect to come about. And thus, *open prospects* supports sophistication.

The problem with *open prospects*, if we want to make a consequentialist argument in favour of separability, is that it apparently cannot explain why anything is wrong with Frieda for failing to abide by consequentialism. At the time when she gets to make a choice, she chooses the prospect she most prefers, out of the ones available to her then. If those preferences define her ends then, then that's just what is instrumentally rational for her to do. To say that Frieda is rationally criticizable for failing to abide by consequentialism would be to say that Frieda is irrational for failing to make it the case that in the beginning of the decision problem, she faced the prospect she most prefers then. It is true that according to *open prospects*, in the beginning of the decision problem, Frieda's preferences over all the prospects open to her then would form the standard of instrumental rationality for any choices she makes then. And so, making it the case that she faced her most preferred prospect at the beginning of the decision problem would serve Frieda's past ends. But at t_2, according to *open prospects*, the standard is a different one. Frieda's choices are then judged against her preferences over the prospects available to her then, and no longer against whether they make it the case that she previously faced her most preferred prospect.

Given Frieda's preferences over prospects, *open prospects* implies that which of the agent's attitudes pick out the agent's ends changes throughout the dynamic

choice problem. In Frieda's case, the attitudes that pick out the agent's ends do not stably support the same course of action. Consequently, what it is instrumentally rational for Frieda to do at t_2 changes. At t_1, *open prospects* endorses taking the gamble and going 'up' at t_2, and at t_2, it endorses playing it safe and going 'down' at t_2. Granted, Frieda retains her preferences over the prospects available to her at t_1 throughout. Even as she chooses the safe option at t_2, she at the same time keeps her Allais preferences over the original gambles. While this is true, and is part of what makes this case so puzzling, this stable preference does not help us in justifying consequentialism as a requirement of instrumental rationality, given *open prospects*. According to *open prospects*, it is simply not instrumentally relevant that Frieda retains her Allais preferences over the prospects available at t_1. This stable preference only picks out the agent's ends at t_1, but not at t_2. What matters at t_2 is that Frieda does well by her preferences over the prospects open to her then. And the full Allais prospects are no longer open to her at t_2.

The requirement of consequentialism thus appears to be either redundant or at odds with instrumental rationality. If the agent's later preferences over sub-prospects agree with her earlier preferences over the entire prospects open to her in the whole dynamic choice problem—as would be the case for agents with separable preferences— then the agent already abides by consequentialism simply by being sophisticated. But if she has preferences like Frieda's instead, then consequentialism would require her to choose against her preferences over the prospects open to her at t_2. And that would be instrumentally irrational according to *open prospects*. She would not be taking the best means to her ends at t_2. Hence, according to the conception of the standard of instrumental rationality that makes sophistication a plausible principle of instrumental rationality, namely *open prospects*, consequentialism turns out not to be a requirement of instrumental rationality. Is there a different way of thinking about the standard of instrumental rationality that would make consequentialism a plausible requirement of instrumental rationality? To avoid the problems we just pointed out, this would have to be a standard that does not imply that Frieda's ends shift in the course of the dynamic choice problem, as *open prospects* did.

Consequentialism would be a requirement of instrumental rationality if the agent's preferences over the prospects available to her at the outset of the decision problem remained the standard against which her later choices are judged. As we already noted, the agent in fact retains those preferences throughout. According to what I want to call *initial prospects*, agents' preferences over the prospects initially open to them define their ends, and instrumental rationality requires them to do well by those preferences. Note that *initial prospects* does not require agents to act by their *initial* preferences over the prospects they face initially. This would hardly be defensible as a notion of instrumental rationality, since it would take something other than agents' current conative attitudes to be the standard of instrumental rationality. Instead, *initial prospects* requires agents to act well according to their current preferences over the initial prospects the actions currently open to them would help bring about. According to *initial prospects*, for decisions that happen after some uncertainty has been resolved, this bygone uncertainty remains relevant.

In our example, *initial prospects* would require Frieda to serve her Allais preferences, even after the uncertainty has been partially resolved at t_2.

Initial prospects seems to offer an instrumental justification for consequentialism, at least assuming stable preferences. In our example, Frieda's Allais preferences over the prospects initially open to her form a stable standard against which to evaluate her actions. However, even disregarding its intuitive implausibility,[7] *initial prospects* will not do for those who want to make a consequentialist argument in favour of separability. And that is because *initial prospects* also implies that sophistication is not a requirement of instrumental rationality. The example of Frieda brings this out. By choosing to go 'down' and play it safe at t_2, Frieda acts in a sophisticated manner. But this choice is not endorsed by *initial prospects*. To best serve her preferences over the initial prospects open to her, she would have to choose 'up'. Sophistication seems instrumentally rational when the agent's preferences over the prospects open to her at the time of action are the standard of instrumental rationality. But if preferences over prospects including bygone uncertainty are the standard of instrumental rationality, the agent may be required to violate sophistication.

Hence, *open prospects* can justify sophistication, and *initial prospects* can justify consequentialism, but neither account of the standard of instrumental rationality can justify both principles. And then neither *open prospects* nor *initial prospects* condemn an agent with Frieda's preferences as irrational. Agents with Frieda's preferences can at most abide by one of sophistication and consequentialism. But according to each standard, that would be enough. According to *open prospects*, there is nothing wrong with Frieda if she is sophisticated and violates consequentialism. And according to *initial prospects*, there is nothing wrong with Frieda if she violates sophistication and abides by consequentialism. Of course, if Frieda adopted separable preferences, she would abide by both requirements at the same time. The problem, however, is that we cannot explain what the instrumental appeal is of abiding by both principles simultaneously, since they only seem attractive on distinct notions of what the standard of instrumental rationality is. And so an agent who doesn't already have separable preferences can't be given an instrumental reason to abide by both, and adopt separable preferences.

Now one might think that both *open prospects* and *initial prospects* are quite narrow notions of the standard of instrumental rationality, and that in fact the true standard of instrumental rationality should include attitudes to both initial and to open prospects. However, conceding that would not help us make the consequentialist case for separability. As we have seen, if she has Allais preferences, Frieda's preferences over initial prospects and her preferences over open prospects are in conflict with each other regarding the question of how to choose at t_2. lf both sets of preferences are part of the standard of instrumental rationality, that is, if both combine to define her ends, then instrumental rationality requires her to find some

[7] However, Machina (1989), at least, seems to think that something like *initial prospects* correctly captures what agents with non-separable preferences ultimately care about: "The key thing is to remember that an agent with non-expected utility/nonseparable preferences feels (*both ex ante and ex post*) that risk which is borne but not realized ... is gone in the sense of having been consumed (or "borne"), rather than gone in the sense of irrelevant" (p. 1647, my emphasis).

compromise between them, and to serve both to some extent. But it does not require her to change her preferences such that the conflict disappears.

The proof presented by Hammond and reconstructed by McClennen should therefore do nothing to convince us that agents like Frieda are instrumentally irrational. We can justify at most one of the two crucial rationality requirements that Hammond presupposes. We can justify consequentialism under *initial prospects*, and we can justify sophistication under *open prospects*. In the following, I want to argue that both of these conceptions of the standard of instrumental rationality share a common feature that in fact makes them of little use for those who want to defend expected utility theory or common alternatives instrumentally. And that is that both proposals assume that it is preferences over prospects—be it preferences over prospects including, or not including bygone risks—that form the standard of instrumental rationality. The next section argues that if we allow such preferences to form part of the standard of instrumental rationality, we cannot justify any general principles of choice under uncertainty instrumentally. In particular, the best account of what is instrumentally irrational about Frieda's choice behaviour, which appeals to her propensity to make a sure loss, can only be made once we abandon this assumption.

5 Prospects and the Standard of Instrumental Rationality

The requirement of state-wise dominance is much less controversial than separability, and accepted even by most rivals of expected utility theory.[8] It is in fact implied by the sure-thing principle.

State-Wise Dominance For any two actions A_i and A_j, if for every state of the world S_k, $O_{ik} \succcurlyeq O_{jk}$, then $A_i \succcurlyeq A_j$. If, in addition, $O_{ik} \succ O_{jk}$ for at least one state of the world S_k, then $A_i \succ A_j$.

This requirement states that if an action leads to an outcome that is weakly preferred to the outcome brought about by another available action in every state of the world, then that action ought to be weakly preferred. Moreover, if it is strictly preferred in at least one state of the world, then the action ought to be strictly preferred.

State-wise dominance seems like a fairly uncontroversial requirement of instrumental rationality. However, we cannot justify even this principle if we take attitudes to prospects to be part of the standard of instrumental rationality, as *open prospects* and *initial prospects* do. Suppose, for instance, that I have a strong desire for secure prospects. This desire is satisfied whenever I choose a prospect that leads to the same outcome in every state of the world. If I have such a desire, that desire is strong enough, and we take it to be part of the standard of instrumental rationality, instrumental rationality does not prohibit me from violating state-wise dominance. I may prefer a safe prospect that leads to a worse outcome no matter what happens, because at least I know in advance what to expect.

[8] I follow McClennen's (1990) formulation of what he calls 'dominance in terms of sure (riskless) outcomes', or DSO (p. 50). See also Buchak (2013), p. 94, who requires state-wise dominance in her rival theory to expected utility theory, just as Quiggin (1982) does.

Similarly, in distribution decisions, a strong desire for giving every potential beneficiary a chance at an equally good outcome might lead to violations of state-wise dominance. If I flip a coin to decide who of my two friends will get some candy, and I have the option of throwing in some extra, at no cost to me, only if it's the one to my left, I may well decide not to do so. The resulting prospect would be less equal, in that it gives the one to my left a chance at a better outcome. Aversion to such inequality, again, is an attitude that has uncertain prospects as its object. If such attitudes are allowed as part of the standard of instrumental rationality, and if they are strong enough, they may well result in permissible violations of state-wise dominance.

If attitudes to prospects such as these are instrumentally relevant, state-wise dominance is not a general requirement of instrumental rationality. In fact, if we admit these attitudes as part of the standard of instrumental rationality, we cannot formulate any general principles about how our preferences over uncertain prospects should relate to our preferences over outcomes. Whatever those proposed principles are, we can imagine an agent who has a strong desire for prospects the choice of which would violate the principle. If this desire forms part of the standard of instrumental rationality, then it helps to define the agent's ends, and it cannot itself be rationally criticized. And then instrumental rationality may require an agent to violate the principle.

We are here interested in whether any instrumentalist justification can be given for principles of how preferences over uncertain prospects ought to relate to preferences over outcomes. For this project to succeed, and for something even as uncontroversial as state-wise dominance to be justified instrumentally, we need to exclude attitudes to uncertain prospects from the standard of instrumental rationality. Instead, we should consider only attitudes to outcomes—be they preferences or more basic attitudes to features of outcomes—to form part of the standard of instrumental rationality. Call this family of standards *outcomes only*. The rest of the paper will argue that instrumentalist arguments in favour of separability fail on *outcomes only*, too.

Outcomes only is in fact a popular position in the debate on the normative status of expected utility theory. It is commonly held that reasons for action must ultimately derive from what things will be like in some state of the world. Broome (1991) appeals to such a claim in his defence of separability. Buchak (2013) is committed to the claim that one prospect can be rationally preferred to another only if it is better in some state. She calls this claim 'betterness-for-reasons' (p. 75), and appeals to it in order to justify her version of state-wise dominance. She thus also takes reasons for action to derive from our evaluations of outcomes in states. In the case of instrumental rationality, where our reasons for action derive from our own conative attitudes, the claim is that the attitudes relevant for instrumental rationality only concern outcomes.

For our purposes, what is important is that *outcomes only* appears to be well suited to the project of justifying expected utility theory instrumentally. And that is because, once we take only attitudes to outcomes to define the agent's ends, the agent's attitudes to prospects are no longer beyond the scope of rational criticism. We can now ask of them whether they help serve the agent's ends, as they are

picked out by the agent's attitudes to outcomes. State-wise dominance, at least, can now apparently be given a straightforward justification. An agent who violates it will do worse by her preferences over outcomes no matter what happens. If those preferences themselves define the standard of instrumental rationality, or if they are correct and complete all-things-considered evaluations based on all the agent's relevant conative attitudes to features of the outcomes, then this is clearly instrumentally irrational.

I will thus adopt *outcomes only* for the sake of argument. Can we justify any further requirements on preferences over prospects, such as separability, on this notion of the standard of instrumental rationality? The next section will argue that those who accept *outcomes only* should also accept what I will call the 'presumption of permissiveness'. It is due to this presumption that instrumentalist arguments for separability ultimately fail.

6 *Outcomes Only* and the Presumption of Permissiveness

Critics of expected utility theory often point out that agents do seem to be sensitive to features of prospects—for instance in the ways I described in the last section. For instance, Lopes (1981, 1996) argues that next to certainty, mean, mode, variance, skewness and probability of loss are further 'global' features of gambles agents may care about. Buchak (2013) calls agents who are sensitive to values that are only achieved though a combination of outcomes across different states (other than expected utility itself) 'globally sensitive'. However, as we saw in the last section, if we were to take this sensitivity to imply that agents have attitudes to prospects that are not merely instrumental, and then included them in the standard of instrumental rationality in their own right we could not justify any general principles of how preferences over prospects ought to relate to preferences over outcomes. We cannot even justify state-wise dominance. But state-wise dominance is accepted by most critics of expected utility theory.

Still, some level of, at least apparent, global sensitivity is generally accepted as perfectly rational. Does this render *outcomes only* immediately implausible? This section argues that this is not so, but that those defending *outcomes only* should adopt what I will call a 'presumption of permissiveness' about attitudes to uncertain prospects. There are two common ways of accommodating (apparent) global sensitivity that are consistent with *outcomes only* and a commitment to state-wise dominance. On the one hand, we could reinterpret any attitude that seems to be directly and non-instrumentally about features of prospects as an attitude to features of outcomes. On the other hand, we could take such attitudes to be expressions of a sensitivity to risk that is compatible with attitudes to prospects being merely instrumental. The first strategy seems to be the orthodox response, while the second strategy is often advocated by those arguing for alternatives to expected utility theory.

To start with an example of the first strategy, that it is part of a certain prospect also seems to be a feature of each of the outcomes of a certain prospect. My desire for certain prospects may then be fully accounted for by my preferences over

outcomes thus described. And then we can justify state-wise dominance instrumentally after all, as well as perhaps other principles, such as separability. Similarly, that it was obtained by means of a lottery that gave my friends chances of unequal amounts of candy could be thought of as a feature of the outcome where my left-hand friend receives candy. And then perhaps my desire for equality could be accounted for by my preferences over outcomes.

At the limit, and in order to capture any kind of apparent global sensitivity, we could include the precise structure of the prospect an outcome is part of in the description of an outcome. Buchak (2013) calls this 'global individuation'. Global individuation is often appealed to in order to defend expected utility theory against apparent counterexamples, like the Allais preferences.[9] Note, however, that this move only helps expected utility theory if we can give some positive defence of separability as a principle of instrumental rationality in the first place. But this is precisely what is at stake here, and the following shows that this is harder than it may seem. In fact the global individuation strategy comes at the cost of both an unnatural description of what the objects of agents' attitudes are, and a proliferation of outcomes that ultimately makes it difficult to give any structure at all to choice under uncertainty.[10] We are thus well advised to put limits on the extent to which outcomes can be redescribed to capture apparent global sensitivity. It is beyond the scope of this paper to discuss precisely how this should be done.

The second strategy for allowing for global sensitivity that is consistent with *outcomes only* is often adopted by proponents of alternatives to expected utility theory, but is open to expected utility theorists as well. According to Buchak, in line with *outcomes only*, agents view prospects as instrumental for doing well by their attitudes to outcomes. Given this is so, state-wise dominance is a minimal requirement for instrumental rationality.[11] But different agents may structure the attainment of their ends differently, and it is here where global sensitivity may play a role. Some agents may be more, and others less risk-averse in the attainment of their goals, and this is consistent with their respective attitudes to uncertain prospects being merely instrumental. Within bounds, which Buchak's risk-weighted expected utility theory aims to capture, instrumental rationality is permissive about how agents pursue their ends, as picked out by their attitudes to outcomes.

In fact, I take the kind of permissiveness about attitudes to prospects that Buchak argues for to be the default position once we accept *outcomes only*. Suppose the Cookie Monster desires only cookies, and it likes them all the same. Everything that the Cookie Monster genuinely cares about in the outcomes of its potential actions is

[9] See, for instance, Weirich (1986), Pettigrew (2015).

[10] See, for instance, Buchak (2013), pp. 139–145. For further arguments against global individuation, see Stefansson and Bradley (2015, forthcoming). They instead defend the view that chances (probabilities of achieving an outcome) can have non-instrumental value, resulting in their value being non-linear in probabilities. My worry is that once we allow for chances to have non-instrumental value, we can no longer require that the value of chances should even be increasing in probabilities, at least not while appealing to instrumental rationality alone. And then not even state-wise dominance seems justifiable.

[11] And so, on this picture, the extreme kind of desire for safe prospects I described earlier can only be considered rational if we can plausibly recast it as an attitude to outcomes. Whether this is so will depend on the case and the precise rule for the individuation of outcomes we adopt.

captured by a description of the number of cookies it will eat. It then considers the question of whether it would be willing to forego 40 cookies for the chance to win 100 cookies in a fair coin toss. It now appears like either answer is compatible with the Cookie Monster being instrumentally rational. And that is because it is not clear which option, the more or the less risky one, serves its desire for cookies better. Sure, we may not allow just any preference over prospects that does not violate state-wise dominance to be instrumentally rational. For instance, in normal circumstances, it may not be instrumentally rational for the Cookie Monster to forego 99 cookies for the chance to win 100 in a fair coin toss. But as long as its preferences over prospects are not this extreme, why shouldn't instrumental rationality be fairly permissive, given only attitudes to outcomes are the standard of instrumental rationality?[12]

In the light of such examples, proponents of *outcomes only* should adopt what I will call a 'presumption of permissiveness', that is, a presumption that beyond state-wise dominance, and the exclusion of extreme forms of risk-aversion and risk-taking, instrumental rationality is permissive about what attitudes to uncertain prospects agents may have given their attitudes to outcomes. The burden of proof lies with those who want to justify requirements on preferences over prospects that go beyond state-wise dominance. This brings us back to the dynamic choice argument in favour of separability, which we might think will provide us with such a justification. Assuming *outcomes only*, does the argument go through?

Unfortunately, the consequentialist version of the argument considered in Sect. 4 still does not go through, since neither consequentialism nor sophistication come out as requirements of instrumental rationality anymore under *outcomes only*. The problem is that if instrumental rationality is permissive about how agents choose between prospects given their preferences over outcomes, then instrumental rationality cannot require agents to choose in accordance with the preferences over prospects they happen to have—be it the prospects open to them at the time of decision, or the initial prospects open to them. As long as the agent ends up choosing some prospect that is permissible given her attitudes to outcomes, then she is not instrumentally criticizable. But sophistication and consequentialism rely on such requirements to be guided by one's preferences.

We have found, thus, that if we want to justify even the most uncontroversial principle of choice under uncertainty instrumentally, we have to allow only for

[12] Those who want to resist this apparent permissiveness of *outcomes only* will either have to insist that there is something else that the Cookie Monster cares about in the outcomes after all, such as how likely the cookies' attainment was, despite the Cookie Monster's insistence that it doesn't, or abandon *outcomes only* and count the Cookie Monster's attitudes to cookie lotteries as non-instrumental. The first response not only amounts to questioning an intuitive description of the case, it also leads to the problematic proliferation of outcomes just mentioned. We should thus only adopt this response if we have very good independent reason to do so. The burden of proof lies with those who want to resist permissiveness to provide such reason. And note that this burden of proof is more demanding than merely providing an argument in favour of separability. Expected utility theory is in fact compatible with permissiveness of the kind described here, as long as we don't interpret the utility function as providing a cardinal measure of the agent's degrees of desire for outcomes. The problem, as the rest of the paper argues, is just that given the presumption of permissiveness, we cannot offer an instrumentalist defence of expected utility theory in the first place. The second response, as previously argued, implies that the instrumentalist argument for separability does not go through.

attitudes to outcomes to form part of the standard of instrumental rationality, rather than also for attitudes to prospects. But if that is so, neither of the two principles Hammond uses to derive a requirement to have separable preferences comes out as a straightforward principle of instrumental rationality. We are hence in need of a different defence of separability. In the following, I argue that a better version of the dynamic choice argument in favour of separability points out that agents with non-separable preferences may end up violating state-wise dominance *over time*.

7 Dynamic Dominance Violations

It can be shown that an agent like Frieda may end up choosing a course of action that leaves her with a worse outcome, no matter what happens, than another course of action she could have engaged in—even though no individual choice she makes is a sure loss choice. Suppose Frieda is offered the opportunity to pay some small cost ϵ,[13] at the beginning of the decision problem, in order to bind herself to the choice she prefers then. This alters the second decision problem to the one pictured in Fig. 2, where t_0 is the point in time at which Frieda can bind herself to lottery C at cost ϵ. As a sophisticated agent, she should choose to in fact bind herself in this way. She knows that if she does not do so, she will choose in accordance with lottery D. If she has a strict preference, at the outset, for C over D, there will be a small enough ϵ such that she prefers to go 'down' at t_0 and bind herself.

This result is worrying because no matter what happens, Frieda will end up with an outcome she strictly dis-prefers to the outcome she would have had, had she taken another course of action that was available to her. The course of action where she chooses not to bind herself, and then chooses to go 'up' at t_2 is also available to her. If she took it, she would also end up choosing in accordance with lottery C, but would avoid paying ϵ. She thus violates state-wise dominance, if we understand it as a principle about entire courses of action. As also noted by Steele (2010), this is a stronger result than the one Hammond (1988) appeals to. It is not only that Frieda ends up with a prospect that she dis-prefers at the outset of the decision problem. She in fact ends up with a worse *outcome* no matter what happens. She is thus instrumentally criticizable even if we take the standard of instrumental rationality to be attitudes to outcomes only.

We may be worried about requiring the agent's entire course of action to abide by state-wise dominance. Diachronic requirements of choice seem especially problematic when they require an agent to choose in a way that seems to be itself instrumentally criticizable at the time of action. For instance, consider what are sometimes called 'temptation cases':[14] Under conditions of certainty, an agent's preferences over outcomes shift, such that the most preferred outcome at the time of 'temptation' is different from what it otherwise is. In such cases, avoiding the sure loss of costly pre-commitment not to give into temptation would require an agent to

[13] This cost need not be monetary. Perhaps pre-commitment involves a social interaction that Frieda is anxious about. Perhaps it involves wasting some precious ink.

[14] See, for instance, Gauthier (1996) and McClennen (1998).

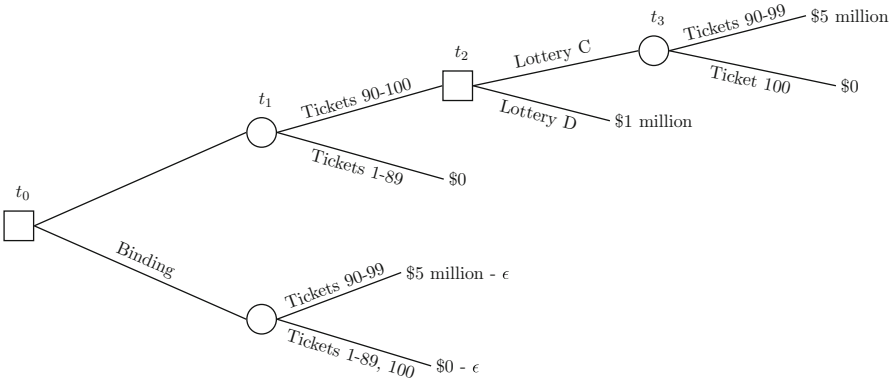

Fig. 2 Alternative second choice

freely choose to act against her preferences over outcomes at the time of temptation. But this seems to be instrumentally irrational in its own right.[15]

However, these worries do not apply in Frieda's case. For her, avoiding taking a dominated course of action would not involve engaging in any actions that are themselves problematic in terms of instrumental rationality, at least not if we are permissive in the way we argued for in the last section. If instrumental rationality is permissive about what attitudes to prospects Frieda may have given her preferences over outcomes, acting against the preference she happens to have over the prospects available at t_2 need not be instrumentally irrational. In fact, if a counter-preferential choice would allow her to abide by state-wise dominance over time, and does not involve extreme risk, it seems like instrumental rationality requires her to make it.

Let's now consider Seidenfeld's (1988, 1994) argument that agents like Frieda are not actually instrumentally irrational. He claims that the alternative course of action whereby agents like Frieda get lottery C without costs of commitment is in some sense not available to them. Since Frieda is in fact sophisticated, it does not seem to be open to her to choose to go 'up' at t_2. And then Allais preferences will not lead her to choose a course of action that is dominated by another one that would have been available to her. Along with McClennen (1990) and Steele (2010), I do not find this response satisfying. *Outcomes only*, which we must adopt in order to make a successful sure loss argument, can help us explain why it is not. It is because she is sophisticated that the course of action of going 'up' at t_2 is not available to Frieda. But sophistication is itself in need of instrumental defence. As we have seen, it is only defensible as such a requirement on an understanding of the standard of instrumental rationality that does not allow us to make an instrumentalist argument in favour of separability—namely *open prospects*. If only attitudes to outcomes are the standard of instrumental rationality, agents are not required to be sophisticated. And so Seidenfeld's response does not work.

[15] See [redacted] for an argument that standard instrumentalist arguments in favour of resisting temptation fail.

I therefore think that, provided *outcomes only* is defensible, this is a convincing argument showing that something is wrong with Frieda: She chooses a course of action over time that is strictly worse with respect to her preferences over outcomes than another available course of action. And she could have avoided this without acting in a way that is itself instrumentally criticizable. Her course of action over time is thus clearly instrumentally deficient. Does this mean she is instrumentally required to adopt separable preferences instead, as the dynamic arguments intended to show? Unfortunately, this is not so, given *outcomes only*. As we have just seen, according to this standard, she is not required to be sophisticated, and can avoid instrumental irrationality by failing to be sophisticated instead.

Frieda could choose, at t_2, to go 'up', against her preference for the safe prospect available at that point in time, and, at t_0, choose to forego paying ϵ to bind herself, for instance because she anticipates her later counter-preferential choice. Or she could go 'down' at t_2, in accordance with her preference then, and nevertheless choose to forego paying ϵ at t_0, against her preference at that point in time. Any dynamic choice strategy that allows Frieda to engage in one of these series of choices would enable her to avoid sure loss. For instance, McClennen's (1990) 'resolute choice' would allow Frieda to choose in the first way: Resolute agents make a plan at the beginning of a dynamic choice problem to act in accordance with their most preferred course of action then, and then simply go through with it. But other dynamic choice strategies may have the same effect.[16]

Many authors have been sceptical of the claim that counter-preferential choice can be instrumentally rational.[17] I take this to be motivated by the idea that the agent's preferences over prospects define her ends in action, and are themselves the standard of instrumental rationality. But we have already argued that we should take only the agent's attitudes to outcomes to be the standard of instrumental rationality, if we want there to be any hope of justifying principles of choice under uncertainty instrumentally. As we argued before, if the agent's preferences over prospects are not themselves part of the standard of instrumental rationality, there seems to be no reason why we shouldn't be permissive with regard to how agents may choose between different prospects. And then we cannot justify a requirement for agents to act in accordance with the preferences over prospects they happen to have.

Frieda's choices may be instrumentally irrational. But to do better, she need not adopt separable preferences—as long as she acts counter-preferentially at the right point in time, she can avoid sure loss. And so we cannot offer a justification for separability as a general principle of instrumental rationality.

8 Conclusions

Expected utility theory is often treated as the correct theory of instrumental rationality under uncertainty. When defending proposed principles of rationality as requirements of instrumental rationality, we usually try to show that agents will do

[16] See, for instance, Rabinowicz's (2014) 'unified' choice.

[17] See, next to Seidenfeld, Levi (1991), Maher (1992), Steele (2010), and, to some extent, Rabinowicz (1995).

badly by their own lights if they violate the principle. In the case of the core requirement of expected utility theory, separability, too, such instrumentalist arguments have been made. My discussion here showed that, to evaluate them, we have to be more explicit about the basis of our instrumentalist argument. When we note that an agent does badly by her own lights, which of her attitudes are we appealing to? That is, what are we treating to be the standard of instrumental rationality?

The most compelling arguments in favour of separability appeal to the way in which agents with non-separable preferences behave in dynamic choice contexts. The success of this kind of argument depends on us being able to establish that (1) agents like Frieda are instrumentally irrational, and (2) to rationally avoid this instrumental irrationality, agents like Frieda need to adopt separable preferences. I argued here that there is no conception of the standard of rationality according to which we can establish both. If we take attitudes to prospects to be part of the standard of instrumental rationality in their own right, we cannot establish (1), though we might be able to establish (2). Indeed, we cannot even establish the much less controversial requirement of state-wise dominance. If we take only attitudes to outcomes to form the standard of instrumental rationality, we can establish (1), given Frieda's propensity to make a sure loss. But we cannot establish (2): There are alternative admissible ways to avoid instrumental irrationality. And so, supporting both (1) and (2) at the same time seems to involve an equivocation about the standard of instrumental rationality.

If we want to defend expected utility theory, or any weaker formal theory of choice under uncertainty, as a theory of instrumental rationality, we are nevertheless well advised to try and defend the idea that only attitudes to outcomes are part of the standard of instrumental rationality. This allows us to, at least, justify the weak requirement of state-wise dominance. Moreover, regarding the dynamic choice argument discussed here, we can then at least say that adopting separable preferences is one good way for Frieda to remain instrumentally rational, even if it is not the only way. Some agents may indeed have desires that make this the uniquely rational response to such choice situations. Still, what we can't establish is that separability is a general requirement of instrumental rationality. Instrumental rationality is more permissive than expected utility theory.

Acknowledgements I am very grateful to Sergio Tenenbaum, Jonathan Weisberg, Joseph Heath, Julia Nefsky, Michael Bratman, and an anonymous referee for very helpful feedback on earlier drafts of this paper. The paper benefitted greatly from a research stay at Stanford University in early 2016, funded by the Balzan Foundation and hosted by Michael Bratman, and from discussion with audiences at the Workshop on Attitudes to Risk at the Hebrew University Jerusalem, the University of Konstanz, the Canadian Philosophical Association Annual Congress at the Université du Québec à Montréal, the Foundations of Normative Decision Theory Workshop at the University of Oxford, and the Foundations of Utility and Risk Conference at the University of York, all in 2018.

References

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.

Briggs, R. A. (2015). Costs of abandoning the sure-thing principle. *Canadian Journal of Philosophy*, *45*(5–6), 827–840.

Broome, J. (1991). *Weighing goods*. England: Blackwell.

Buchak, L. (2013). *Risk and rationality*. Oxford: Oxford University Press.

Gauthier, D. (1996). Commitment and choice. In F. Farina, S. Vannucci, & F. Hahn (Eds.), *Ethics, rationality, and economic behaviour* (pp. 217–243). Oxford: Oxford University Press.

Hammond, P. (1988). Consequentialist foundations for expected utility. *Theory and Decision*, *25*, 25–78.

Jeffrey, R. (1965/1983). *The logic of decision* (2nd ed.). University of Chicago Press.

Levi, I. (1991). Consequentialism and sequential choice. In S. Hurley & M. Bacharach (Eds.), *Foundations of decision theory*. England: Blackwell.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824.

Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, *53*(2), 271–282.

Lopes, L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 377–385.

Lopes, L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Journal of Experimental Psychology*, *65*(3), 179–189.

Machina, M. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, *27*(4), 1622–1668.

Maher, P. (1992). Diachronic rationality. *Philosophy of Science*, *59*(1), 120–141.

McClennen, E. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

McClennen, E. (1998). Rationality and rules. In P. Danielson (Ed.), *Modeling rationality, morality, and evolution* (pp. 13–40). Oxford: Oxford University Press.

Morrison, D. (1967). On the consistency of preferences in Allais' paradox. *Behavioral Science*, *12*(5), 373–383.

Pettigrew, R. (2015). Risk, rationality, and expected utility theory. *Canadian Journal of Philosophy*, *45*(5–6), 798–826.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Rabinowicz, W. (1995). To have one's cake and eat it, too: Sequential choice and expected utility violations. *Journal of Philosophy*, *92*(11), 586–620.

Rabinowicz, W. (2014). Safeguards of a disunified mind. *Inquiry*, *57*(3), 356–383.

Savage, L. (1972). *The foundations of statistics*. New York: Wiley. second revised edition.

Seidenfeld, T. (1988). Decision theory without independence or without ordering. *Economics and Philosophy*, *4*, 267–290.

Seidenfeld, T. (1994). When normal and extensive form decisions differ. *Logic, Methodology and Philosophy of Science*, *9*, 451–463.

Steele, K. (2010). What are the minimal requirements of rational choice? Arguments from the sequential-decision setting. *Theory and Decision*, *68*(4), 463–487.

Stefansson, H. O., & Bradley, R. (2015). How valuable are chances? *Philosophy of Science*, *82*(4), 602–625.

Stefansson, H. O., & Bradley, R. (forthcoming). What is risk aversion? British Journal for the Philosophy of Science. https://doi-org.gate2.library.lse.ac.uk/10.1093/bjps/axx035

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Weber, M. (1998). The resilience of the Allais paradox. *Ethics*, *109*(1), 94–118.

Weirich, P. (1986). Expected utility and risk. *British Journal for the Philosophy of Science*, *37*, 419–442.