Folk Psychology and the Interpretation of Decision Theory

January 1, 2020

Abstract

Most philosophical decision theorists and philosophers of the social sciences believe that decision theory is and should be in the business of providing folk psychological explanations of choice behaviour, and that it can only do so if we understand the preferences, utilities and probabilities that feature in decision-theoretic models as ascriptions of mental states not reducible to choice. The behavioural interpretation of preference and related concepts, still common in economics, is consequently cast as misguided. This paper argues that even those who strive to provide folk psychological explanations should side with the economists, and adopt a behavioural interpretation of the preferences featuring in decision-theoretic models. Under a mentalistic interpretation of preference, decision-theoretic models do not straightforwardly provide ordinary folk psychological explanations. Instead, they involve controversial enough commitments about the mental causes of choice to not only fail to adequately capture much unreflective decision-making, but also many intentional, reason-based and instrumentally rational choices. Satisfactory folk psychological explanation in fact only comes indirectly from inferring more fundamental conative attitudes from a pattern of decision-theoretic preferences. And the behavioural interpretation does a better job at facilitating such inferences. My argument extends to the related concepts of utility and probability.

1 Introduction

In an often cited passage from his 1974 paper on radical interpretation, David Lewis declared that "decision theory (at least if we omit the frills) is not an esoteric science, however unfamiliar it may seem to an outsider. Rather it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized." (Lewis (1974), p.338) It is still a near universal conviction amongst decision theorists and philosophers of the social sciences that decision theory, and in particular standard expected utility theory, provides us with a more refined version of common-sense,

or folk psychological explanation.[1] In fact, introductory expositions of decision theory often start out with such a claim.[2] And in its myriad applications in various branches of philosophy and social sciences other than economics, decision theory is usually introduced in this way. Its ability to provide folk psychological explanation is also taken to be a key explanatory strength of decision theory.

Folk psychology, as it is traditionally understood, takes actions to be caused by belief-desire pairs: A desire to drink some coffee, together with the belief that the drink in front of me is coffee, cause me to drink from the mug in front of me. Moreover, these belief-desire pairs are taken to be not merely causes of my behaviour, but also reasons for my behaviour: From a first person perspective, I can consult my desires and beliefs when deliberating about what to do. And from a third person perspective, the belief-desire pair can serve to rationalize my behaviour, and to make it intelligible to other people. Explanations that appeal to such belief-desire pairs are consequently taken to be deeper kinds of explanation than merely causal explanations. We not only learn what caused the agent's choice, we also come to understand the reasons why she acted as she did. Even if decision-theoretic explanations are ultimately taken to be more sophisticated than ordinary folk psychological explanations, the ambition for them is essentially the same: By citing an agent's relevant conative and cognitive mental states, they are meant to provide both the reasons and the causes of her choices.

Within expected utility theory, agents are modelled as maximizing the probability-weighted sum of the utilities of the various outcomes their choices might lead to. However, not utility but a binary preference relation is usually taken to be the more basic concept, with a variety of representation theorems showing that agents can be represented as expected utility maximizers if their preferences over outcomes and uncertain prospects abide by a variety of axioms.[3] To interpret expected utility theory as a refinement of folk psychology, the probabilities ascribed to agents are commonly assumed to play the counterpart of belief, and the utilities and/or preferences assigned to the agents are assumed to play the role of desire. Preferences and/or utilities combined with probabilities are thus meant to play the same role as belief-desire pairs traditionally do in folk psychology.[4] In the terms of expected utility theory, the explanation of my drinking from the mug in front of me just given could perhaps be expressed as follows: I drank from the mug in front of me

[1] While 'folk psychology' is often used as a disparaging term in other areas of philosophy, the term is used more approvingly by decision theorists and philosophers of the social sciences. I will follow this non-disparaging usage.

[2] See, for instance, Steele (2014).

[3] von Neumann and Morgenstern's (1944) is the one most commonly appealed to within economics, and takes probabilities to be independently given. Savage (1972) and Jeffrey (1965/1983) also derive probability from preference.

[4] Hausman (1998) takes belief-desire pairs and probability-utility pairs to play "virtually identical" functional roles. We also find this view expressed in, e.g., Pettit (1991), Gruene (2004), and more recently, Dietrich and List (2016) and Stefansson and Bradley (2019).

because I assign high probability to there being coffee in it, and I prefer (and thus assign higher utility to) drinking coffee to not drinking coffee (we shall see shortly that in fact things turn out to be more complicated).

The ambition to provide folk psychological explanations thus appears to necessitate a certain kind of interpretation of the probabilities, utilities and preferences featuring in decision-theoretic models: They must be ascribing mental states to agents. Moreover, they must be ascribing both conative and cognitive mental states, with probabilities assumed to ascribe the cognitive ones and utilities and preferences thought to ascribe the conative ones. This has come to be known as the 'mentalistic' interpretation of preference, utility and probability. To take a representative example, Hausman (2012) argues at length for an interpretation of preference as total comparative evaluation: Option a is preferred to option b just in case, taking into account all relevant considerations, the agent evaluates a to be better than b. Similarly, we often find them described as all-things-considered judgements of choice-worthiness or desirability by decision theorists.[5] Functionalists such as Dietrich and List (2016) think of preference, utility, and probability as mental states simply *in virtue of* playing the roles of desire and belief respectively in the folk psychological explanations allegedly furnished by decision theory. The mentalistic interpretation of preference, utility and probability stands in stark contrast to the behavioural interpretation of these concepts still common in economics, according to which the preferences featuring in decision-theoretic models are mere convenient re-descriptions of agents' choice behaviours, and utilities and probabilities, in turn, conveniently re-describe preference. The behavioural interpretation is accordingly criticized for diminishing the explanatory resources of decision theory, by taking away its ability to provide folk psychological, rationalizing explanations.

This paper argues that the ambition to provide folk psychological explanations does not justify adopting a mentalistic interpretation of preference, utility, and probability. In fact, even those who wish to provide folk psychological explanations should side with the economists, and adopt a behavioural interpretation of the preferences, utilities and probabilities featuring in decision-theoretic models. My argument will be the following: Satisfactory folk psychological explanations feature mental states that have coarse-grained states of affairs as their object, such as, for instance, the desire 'that I drink some coffee'. However, preferences, utilities and probabilities as they feature in decision theory have fine-grained states of affairs as their objects, such as a preference 'that I drink some coffee in the morning, on a day that I am not running late, when there is still enough left for my flatmate etc. ...'. The mentalistic interpretation of preference, utility and probability thus involves the ascription of fine-grained mental states, where fine-grained mental states are those that have fine-grained objects. But I argue that such fine-grained mental states cannot play the role mentalists would like mentalistic preference, utility and

[5] See, e.g., Bradley (2017) or Joyce (1999).

probability to play: Unlike the platitudes of folk psychology, ascription of fine-grained mental states is often intuitively implausible even in cases of intentional, reason-based and instrumentally rational choice; Best explanations of choice behaviour, be they folk psychological or by reference to theories from the behavioural sciences, reference coarser-grained attitudes and do not rely on the ascription of fine-grained mental states; And attempts at explanation by reference to fine-grained mental states are not satisfactory. Hence, the standard justification for the mentalist interpretation of preference, utility and probability does not work.

The upshot is this: Given that successful folk psychological explanation appeals to coarser-grained attitudes, but decision theory only deals in preferences, utilities and probabilities that have fine-grained objects, it cannot offer folk psychological explanations in any straightforward way. All it can hope to do is offer a framework that facilitates the ascription of the coarser-grained mental states that would feature in successful folk psychological explanation. By helping us recognize salient features that affect choice behaviour, decision theory can in fact serve a useful function here. However, it can serve this function better if we adopt a behavioural interpretation of preference, utility and probability. The behavioural interpretation makes decision-theoretic models more parsimonious by avoiding commitment to the ascription of fine-grained mental states.

Since preference is usually taken to be the more basic concept than utility and probability, the following discussion will mostly focus on preference. In the end, I will argue that, depending on one's view about the relation of utility and probability to preference, either the behavioural interpretation of preference I argue for directly implies a behavioural interpretation of utility and probability as well, or my argument can be made for those concepts *mutatis mutandis*.

The paper is structured as follows: Section 2 will outline the mentalist and behavioural interpretations of decision-theoretic preference, and the core advantages commonly attributed to them, namely that the behavioural interpretation allows for a greater range of applicability of the theory, while the mentalist interpretation guarantees greater explanatory power, in particular due to its alleged connection to folk psychological explanation. Section 3 lays the groundwork for the main argument of this paper by arguing that the attitudes featuring in ordinary folk psychological explanations have coarser-grained objects, while preferences in decision theory have fine-grained objects, rendering the mentalist about preference committed to the ascription of fine-grained mental states. Section 4 argues that the presupposition that mentalism makes possible folk psychological explanation in a way that behaviourism about preference does not allow for is thus mistaken. The way in which decision-theoretic models may help to provide folk psychological explanations is indirect at best, and not dependent on a mentalistic interpretation of preference. This undermines the claim that mentalism has an explanatory advantage over behaviourism.

Sections 5 and 6 argue that mentalism also leads to a significant loss in generality for decision theory, as there are no plausible grounds for ascribing fine-grained mentalistic preferences to agents who make decisions fairly unreflectively, even if, as argued in 6, they are intentional and instrumentally rational. Hence, there is a strong case to stick with a behavioural interpretation of preference, even if one wishes to provide folk psychological explanations. Section 7 extends my argument to utility and probability, and Section 8 concludes.

2 Behaviourism vs Mentalism about Preference

Standard decision theory assigns a binary preference relation to each agent. The relata of the preference relation are whatever the agent's options are taken to be by the specific decision theory. In the simple case of choice under conditions of certainty, we can think of these as the outcomes we are certain will arise from our choices. When there is uncertainty, the options will be descriptions of the uncertain prospects arising from our choices, such as probability distributions over outcomes known as 'lotteries' (as in von Neumann and Morgenstern's framework), or assignments of outcomes to states of the world (as in Savage's framework). For an agent to be representable as an expected utility maximizer, or be captured by the formalism of any other kind of decision theory, this preference relation has to fulfil various consistency conditions. Least controversially, strict preference has to be irreflexive and transitive. In fact, these conditions, applied in a context of certainty, will be enough to generate the problems I will discuss in the following. And so my argument applies to any formal decision theory that makes these minimal consistency assumptions about preference.

As I will be focusing on choice under certainty, the objects of choice and preference are outcomes. What kinds of things are outcomes? They are meant to capture things about the state of the world that the agent's choice brings about, and that are relevant to her choice, or that she cares about. Different decision theories differ by either taking outcomes to be primitive entities (as Savage does, and as is generally the case in economics), or by taking them to be propositions (as Jeffrey and many philosophical decision theorists do). When we think of outcomes as propositions, we think of them as descriptions of the state of the world that pick out things that are relevant to the agent's choice.

Proponents of mentalistic and behavioural interpretations of preference disagree about what it means to ascribe preferences to agents within a decision-theoretic framework. Mentalists think that preference ascriptions are ascriptions of conative mental states: To say that an agent prefers one outcome over another is to say that she likes it more, judges it to be better all-things-considered, or has a stronger desire for it. Proponents of behavioural

interpretations, on the other hand, take preferences to be mere re-descriptions of agents' choice behaviours. According to behavioural interpretations of preference, roughly, what it means to prefer outcome a to outcome b is just that one actually does, or hypothetically would choose outcome a rather than outcome b when faced with a choice between them.

Note that we usually do not take propositions to be the object of conative attitudes or of choice, yet many decision theorists think of outcomes as propositions. I do not desire or choose the proposition 'I drink tea', rather, I desire or choose to drink tea. Strictly speaking, whether they are mentalists or behaviourists about preference, what decision theorists who take outcomes to be propositions must be thinking of as the object of preference are the truth-values of the propositions. If preferences are conative attitudes, they express that an agent would rather one proposition be true than another. That is, she would rather that those states of affairs occur that make the proposition true than those that would make another proposition true.[6] If preferences are choices, then they are attempts to make one proposition true rather than another, by bringing about those states of affairs that would make it true. For simplicity, I will follow convention in the following and simply speak of outcomes as the object of choice and preference.

The behavioural understanding of preference is very popular in economics,[7] and has been bolstered by further representation theorems showing the representability of patterns of choices in terms of a binary relation fulfilling the basic formal requirements of a preference relation.[8] These representation theorems along with the behavioural interpretation of preference are known in economics under the heading of 'revealed preference theory'. What decision theory does, on the behavioural picture, is specify consistency conditions on choice behaviour that allow for a convenient representation of agents as, e.g., expected utility maximizers. This representation can be useful for various scientific purposes, not least the prediction of future choice behaviour. But, according to proponents of the behavioural interpretation, it does not involve the ascription of mental states to agents.[9]

[6] Also see Joyce (1999), pp. 50-51 on this point.

[7] For a widely discussed recent defence, see Gul and Pesendorfer (2008).

[8] See Houthakker (1950) and Afriat (1967).

[9] There is one sense in which this last claim needs to be qualified. In order for decision theory to have any hope of being empirically adequate, the description of agents' options in decision-theoretic models (so the specification of outcomes in the case of certainty) must be consistent with what agents believe and/or perceive about their options. If they are not, changes in choice behaviour resulting from a change in belief state will be taken to result in an inconsistency that undermines representability in terms of a decision-theoretic model. Elsewhere (see redacted), I argue in more detail that revealed preference theorists should adopt such a restricted mentalism about the description of options in decision-theoretic models, but that this is consistent with economic practice, and that this mentalism is moreover fairly minimal. For one, even with this concession, revealed preference theorists can hold on to the core behavioural idea that preference is choice: It is just choice between options described in a way that needs to be consistent with the agent's beliefs (or perceptive states). But for our purposes, it suffices to note that this concession does not undermine the argument against mentalism given in the following. The kind of mentalism defended by appeal to folk psychology is significantly stronger, as it involves the ascription not only of cognitive states, but also of conative states. And moreover, the case in favour of the behavioural interpretation of

Why might the behavioural interpretation be an attractive way to think about preference? Like several other recent commentators,[10] I take the core motivation of revealed preference theory not to be a general and outdated behaviourism, but rather appeal to the potential advantages of black-boxing the mental causes of choice. Common sense tells us that motivation is multifarious: Some of our choices are unreflected and habitual. Others are made after a great deal of deliberation. Some are motivated by brute impulse, others are motivated by a sense of obligation, yet others are calculated to get the most of something we value. A look at the psychological literature on choice confirms this common sense observation: While there is clearly much we don't know about the psychology of choice, and there is a great deal of controversy surrounding different theories of the psychology of choice, the one claim that does seem to be uncontroversial in the psychological literature is that agents make choices in different ways under different circumstances.[11] Moreover, while my argument in the following does not hang on incorporating the non-human case, note that decision theory is not only applied to humans. In fact, it has been found to fruitfully apply to the behaviour of, amongst others, rats, pigeons, and primates.[12] Presumably, the psychology of choice of these non-human animals differs in important ways from that of humans.

Provided we have a model that has a good fit with the choices agents (including non-human agents) ultimately end up making, these seem to be precisely the kinds of circumstances under which it might make good sense to black-box underlying psychological processes. Decision theory under a behavioural interpretation of preference allows us to analyze choice behaviours resulting from any psychological mechanisms, known or unknown, with the same formal apparatus, provided the resulting choices abide by the standard axioms of revealed preference theory. All that needs to be true of agents for the

preference I shall make in the following is not undermined by the concession of a limited mentalism about the description of options. And that is because this case does not rely on any strong commitment to a more general behaviourism. See Clarke (2016) for a similar argument in the context of economics. Moreover, this concession to mentalism does not require the ascription of fine-grained mental states, as the agent's relevant beliefs could have coarse-grained objects.

[10] See Clarke (2016), Ross (2011), Gul and Pesendorfer (2008), redacted.

[11] Bettman et al. (1998), for instance, provide an overview of different choice strategies consumers may use to make consumption choices. These vary in cognitive demandingness and in how well suited they are to different choice situations. Together, they are claimed to form an "adaptive toolbox" from which agents can choose depending on the circumstances. Even those presenting theories of cognitively more involved, "reason-based choice", such as Shafir et al. (1993), admit that many ordinary choices are affect-based. In the philosophical literature, Gibbard (1998) has expressed concern that mentalistic preference might not be able to do justice to the variety of human motivation: "Motivations are of diverse kinds. They can be grounded in emotions, in craving and appetites, in the maintenance of self-esteem, in the social pressures of one's circumstances, and in the acceptance of norms. A good psychology of human motivation would presumably extend this list and revise it." (p. 250)

[12] Kagel et al. (1995) discuss primarily rats and pigeons, and see Santos and Chen (2009) for a study involving primates, and Angner (forthcoming) for another discussion of the implications of such work for the interpretation of preference in economic theory. Interestingly, in a review article, Kalenscher and van Wingerden (2011) find that non-human animals and humans even tend to violate standard decision theory under similar kinds of circumstances.

theory to find fruitful application is that they consistently respond to what they believe or register about their choice environments. On the behavioural interpretation, decision theory is a theory of consistent choice, and can be applied to all consistent choosers, no matter what their psychology. Mentalist interpretations of preference, on the other hand, seem to risk arbitrarily restricting the range of applicability of the theory to agents who can plausibly be ascribed mentalistic preferences, and thus lose out on the potential generality of the theory.

In the light of this, why would the majority of philosophers writing on decision theory insist on a mentalistic interpretation of preference? The answer lies, I think, in the desire to provide folk psychological explanations of choice behaviour, or even insight into an agent's reasoning processes, and the apparent ability of a mentalistic decision theory, and only a mentalistic decision theory, to do so. There are various predictive and explanatory goals we might pursue with decision theory. Decision theory under a behavioural interpretation of preference seems to serve predictive purposes well in a wide range of circumstances. It might also help us come up with a more systematic and parsimonious representation of agents' choice behaviours, and thus even allow for a thin kind of explanation through unification, by showing how a single choice fits into a pattern of choices.[13] But what it can't seem to give us is a causal and rationalizing explanation of agents' choices. If preferring a to b just means that I choose a rather than b whenever I am given a choice between the two, then my preference does not causally explain or rationalize my choice.[14] Only an interpretation of preference as a mental state seems to allow decision theory to provide such explanation. And moreover, if the mental state in question is conative, this explanation will have the advantage of closely mirroring folk psychological explanation.

It is also undeniable that understanding preference as a mental attitude akin to desire that can cause and rationalize action closely matches the concept of preference as it is usually understood in ordinary language. When I say I prefer drinking coffee to tea, this will usually be understood as an expression of a more positive attitude to coffee than to tea, which can then help to explain and rationalize my behaviour in hot beverage contexts. The mentalistic interpretation of preference within decision theory thus seems to create continuity with this ordinary usage, and with ordinary folk psychological explanation.

There are two importantly different ways of making this mentalist rejoinder, and they

[13]See Vredenburgh (2019) for a defence of this idea. Early proponents of revealed preference theory also invoked this sense of explanation. Little (1949) writes: "If an individual's behaviour is consistent, then it must be possible to explain that behaviour without reference to anything other than behaviour. Someone, on the other hand, might object that market behaviour cannot be really explained by means of a map which is constructed out of nothing but that behaviour. The metaphor I have used to state this objection provides the answer. The terrain of England really is explained by a map of England. The map is constructed only by reference to this terrain." (pp. 97-98)

[14]See, for instance, Joyce (1999), pp. 21-22 for a representative statement of this case against the behavioural interpretation pf preference.

correspond to different strands of mentalism. On the one hand, there are functionalists such as Dietrich and List (2016), who believe that what makes it the case that we can ascribe a conative mental state corresponding to the preferences featuring in a decision-theoretic model is simply that the preferences play a desire-like functional role in the folk psychological explanation of choice behaviour allegedly furnished by decision-theoretic models. The fact that there is a robust decision-theoretic model of an agent's choice behaviours would then be enough to ascribe the relevant mental states to her. If this is so, the behaviourist's case for black-boxing the mental causes of choice is moot. Mentalistic preferences can be ascribed to any kind of agent that is well captured by a decision-theoretic model (including the non-human ones just mentioned), and there is thus no loss of generality in ascribing mentalistic preferences. But the plausibility of the view crucially depends on preferences in decision-theoretic models in fact playing a desire-like role in folk psychological explanations.

On the other hand, there are those mentalists who believe that decision theory captures something substantive about the mechanisms of decision-making, be it an agent's actual conscious reasoning process leading to a decision, or potentially sub-conscious mental processes – processes that can potentially be traced independently of decision-theoretic models (be it through conscious access by the agent, brain data or otherwise). Assuming that conative and cognitive mental states play a role in these processes, citing them in an explanation would give insight into the processes that bring about a choice and thus significantly illuminate, or even rationalize it. On this view, too, then, the key advantage of mentalism is that it allows for a kind of folk psychological explanation, albeit one that is assumed to track actual mental processes that can be traced independently of the folk psychological explanation.

One significant type of such a more substantive mentalism involves viewing decision theory as a theory of conscious reasoning or deliberation. This is arguably the standard view of decision theory amongst philosophical decision theorists, and it usually comes with an understanding of preference as an 'all-things-considered evaluative judgement' (see, e.g., Bradley 2017 and Joyce 1999) or a 'total comparative evaluation' (see Hausman 2012). On this view, while such judgements or evaluations perhaps don't need to be mental states that are occurrent at the time of choice, an agent at least needs to have conscious access to them, as they were the results of an earlier conscious reasoning process resulting in that judgement. This is expressed by the standard interpretation of the completeness axiom as being cognitively very demanding by requiring agents to have actively considered all potential outcomes or uncertain prospects they might ever choose between.[15]

[15]Hausman (2012), for instance, treats completeness as a boundary condition on rational choice, while acknowledging that abiding by it would be a "remarkable intellectual achievement" resulting from "an unmodeled process of exhaustive comparative evaluation" (p.18).

The potential loss in generality pointed out by the behaviourist is a greater worry for substantive kinds of mentalism about preference, as not all types of agents whose choice behaviours can be captured with decision-theoretic models need to undergo conscious or unconscious mental processes that are well-tracked by decision-theoretic models in the substantive way just described. Certainly, not all agents whose choices have a robust decision-theoretic representation are reasoning agents. But at least, by standardly understanding preference to be a kind of summary attitude (total, or all-things-considered), this kind of mentalism can subsume many different kinds of motivations. And moreover, the hope is that the gain in explanatoriness through insight into the mental causes of choice can make up for the potential loss in generality.

In the following, I will argue against both functionalist and more substantive types of mentalism about preference. I will not dispute that providing folk psychological explanations of choice behaviours is a legitimate goal for decision theory. Instead, I will argue that even if we are interested in providing folk psychological explanations, the behavioural interpretation of the preferences featuring in decision-theoretic models is preferable. Interpreting preference to be a conative mental attitude does not actually furnish us with satisfactory folk psychological explanations in any straightforward way; It does not, like the functionalist claims, play a desire-like role in folk psychological explanation. And moreover, no matter what kind of mentalism we subscribe to, interpreting preferences as mental states comes with a more significant loss of generality than one might think. To the extent that decision theory can play a useful role in providing folk psychological explanations or offer insight into cognitive processes at all, I will show that the behavioural interpretation of preference fits the bill better.

3 On the Objects of Preference

My argument against mentalism and its appeal to folk psychology will be based on an observation about the objects of preference in decision theory and the objects of preference and desire in ordinary folk psychological explanation. This section will elaborate on that observation, before I turn to my core argument in the next section. Take my choice of whether to have coffee or tea in the morning. My flatmate observes me drinking coffee one morning and tea on the next, despite both being available on both days. He knows me pretty well, and comes up with the following folk psychological explanations: X [author's name] drank coffee on the first day because she prefers tasting coffee to tasting tea, and she knew that she would taste coffee if she drank coffee. X drank tea on the second day, because she wanted to keep her nerves down for her important meeting, and believed the tea would keep her less nervous than the coffee.

How could my flatmate use a decision-theoretic model to capture this choice behaviour? Suppose he modelled the outcomes I am choosing between simply as "X drinks coffee" and "X drinks tea". In that case, unless my flatmate wanted to stipulate that my preferences have changed or that I acted counter-preferentially on some occasions, he could only capture my choices with a consistent preference relation if I was indifferent between the outcomes. But suppose my flatmate also has reason to believe that I am not indifferent. He knows that on each day, I would have chosen as I did even if my chosen option was made slightly more inconvenient, for instance by him having misplaced my beverage of choice. And so if the outcomes are described as "X drinks coffee" and "X drinks tea" the preferences I exhibit violate irreflexivity of strict preference.

To capture my choice behaviour with a consistent preference relation, it seems my flatmate has to specify the outcomes in a more fine-grained, that is, more detailed, way. In particular, at a minimum, he should include in the specification of the outcomes not only descriptions of the taste of the beverage involved, the beverage's caffeine content, and whether or not I have an important meeting that day. To consistently capture my behaviour on other days, he would likely also have to include as part of the outcome descriptions of just how nice the coffee or tea is that we have in the house right now, whether we are running low on coffee, tea or milk, if I am pressed for time, and whether there is a tea drinker present whom I am trying to impress. The preferences the decision-theoretic model would be stipulating in order to make sense of this case would be preferences over those more fine-grained outcomes, and thus descriptions of the various combinations of circumstances that might affect my choice in the contexts we are interested in capturing.[16]

The more fine-grained nature of the outcomes that feature in decision-theoretic models distinguishes the preferences decision theory invokes from preference as it is typically appealed to in ordinary language and explanation. Agreeing with my flatmate's original folk psychological assessment of me, I might say that I prefer tasting coffee to tasting tea. While this preference featured in my flatmate's folk psychological explanation of my choice behaviour above, it will not feature as such in any useful decision-theoretic model of my choice situation. Useful decision-theoretic models will feature only preferences over more specific outcomes involving me tasting coffee or tea, combined with descriptions of various other features of the consequences of the actions available to me that may affect my choices in contexts of interest. At best, we can read off a general preference for tasting

[16]One might think that a simple description of the outcome as simply "X drinks coffee" will be enough in the standard case, and more fine-grained outcome specifications, such as "X drinks coffee and has an important meeting" are only necessary when unusual things, such as an important meeting, happen. However, in the Savage and von Neumann-Morgenstern frameworks, at least, outcomes need to be mutually exclusive – and these are the main frameworks used in the social sciences. In Jeffrey-style decision theory, both could potentially be treated as outcomes in different models using one consistent utility representation. However, as I argue in [redacted], preferences over simple propositions such as "X drinks some coffee" in the Jeffrey-style framework are quite different from what I call the 'coarser-grained attitudes' featuring in ordinary folk psychological explanation in the following.

coffee from the fact that I tend to prefer outcomes that involve me drinking coffee to ones that involve me drinking tea, unless other more important considerations lead me to prefer tea.[17]

More generally, decision theorists usually assume that outcomes have to be described in enough detail that everything that might affect an agent's choice in the contexts of interest is included. Economists typically take the objects of preference to be consumption bundles capturing the various different goods an agent consumes. Why is that? Because it is only at that level of description that we are likely to find the consistency we need in order to allow for representation with a decision-theoretic model. Even when we are only interested in capturing a specific subset of an agent's choices, the description of outcomes needs to include at least those circumstances and goods that are important complements to the main object of choice we are interested in. This will typically render the object of preference even in simple applications, e.g. in the analysis of consumer demand, more complex than the object of preference in ordinary preference-talk. Economists can't just model agents as choosing between 'apples''and 'oranges' if they want to fruitfully apply a decision-theoretic model. If they don't model outcomes as, e.g., bundles of fruit, other consumption goods and time of year, they will encounter inconsistency as soon as the season and prices of other goods change. We can of course come up with very simple decision-theoretic models that feature outcomes as coarse as the objects of preference and desire in ordinary folk psychological explanation. However, these would need to have only very narrow applicability – in our coffee case the model featuring such simple outcomes would cease to apply as soon as an important meeting comes up. And in any case, all I need to make my case is that the objects of decision-theoretic models are *typically* more fine-grained than the objects of preference and desire in folk psychological explanation.

Along similar lines, Pettit (1991) noted that preference and desire can in fact have two sorts of objects, what he calls 'prospects' and 'properties'. Prospects are descriptions of states of the world, ways the world may turn out to be, essentially what we have been calling 'outcomes'. According to Pettit, each such prospect (or outcome) will instantiate various properties. One such property could be the property of involving me tasting coffee. Pettit notes that these properties could also be described as kinds of prospects, albeit coarser grained ones than the ones decision theory typically deals with – ones that

[17]It might be responded that a fuller folk psychological explanation of my choice behaviour would not only state that I prefer tasting coffee to tasting tea, but also that this desire was not outweighed by any other relevant factors. And it might be thought that the addition of this clause amounts to reference to a fine-grained mental state. However, note that decision-theoretic preferences do not contain information, considered individually, about how the preference relates to underlying coarser-grained attitudes, which this folk psychological explanation does. We will see in the following section that this the crucial reason why mentalist decision-theoretic preferences do not, in their own right, furnish us with satisfactory folk psychological explanations. Moreover, I will argue in sections 5 and 6 that we have no reason to suppose that the weighing of coarser-grained attitudes we no doubt often engage in results in a fine-grained mental state rather than simply a choice. I thank an anonymous referee for pushing me on this point.

will be instantiated by many different finer grained prospects. I prefer to refer to both as 'outcomes', distinguished by being coarser-grained and finer-grained. Finer-grained outcomes describe or capture more specific states of affairs, coarser-grained less specific ones. Both are potential objects of our conative attitudes. I will in the following speak of coarser-grained and finer-grained attitudes when speaking of attitudes with coarser or finer outcomes as their object.[18]

My core point here is the following: The preferences and desires that feature in standard folk psychological explanations are typically very coarse-grained. They have as their objects outcomes such as 'I taste some coffee'. The preferences that feature in decision theory, on the other hand, are preferences over more fine-grained outcomes. In particular, they are outcomes specific enough to capture all factors that can make a difference to choice in the decision contexts we wish to model with our decision-theoretic model. If they were not, we would regularly find choice inconsistencies in the decision contexts we aimed to capture, undermining the use of the theory. In the following, this is the level of grain I have in mind when speaking simply of 'fine-grained' outcomes and attitudes. If the preferences that feature in decision theoretic models are understood to be conative mental attitudes, they are thus what I will call fine-grained attitudes.

I would like to make one claim about fine-grained attitudes that I will take for granted in the following. If we do form attitudes to fine-grained outcomes, then these are at least partly explained by our coarser-grained attitudes.[19] For instance, if I prefer the fine-grained outcome involving me drinking coffee from my rainbow streetcar mug this morning to the fine-grained outcome involving me drinking tea from a plain grey mug this morning, this is explained by my more coarse-grained preferences of coffee over tea, and of the rainbow streetcar mug over the plain grey one, and the fact that this is not outweighed by any countervailing considerations. More generally, attitudes to fine-grained outcomes, if we form them at all, are at least partly the result of weighing various coarser-grained attitudes to the different coarser-grained outcomes that comprise the fine-grained outcomes.

Still, in the following, I do not offer a full account of how choice results from underlying coarse-grained attitudes. My concern is with the interpretation of preference and related concepts in decision-theoretic models, and as I argued here, coarse-grained attitudes do not feature in standard decision-theoretic models. My point is not that decision theory

[18]Note that there are other philosophical debates, in particular in formal epistemology, where coarseness of grain of attitudes distinguishes the type of attitude – e.g. whether an agent can assign precise probabilities or only coarser categories of confidence in propositions. My usage here is different, in that coarseness of grain of an attitude is determined by the coarseness of its object.

[19]Pettit (1991) in fact claims that attitudes to fine-grained outcomes are fully explained by attitudes to the properties of the fine-grained outcomes (or coarser-grained outcomes). He calls this a 'platitude of desiderative structure'.

needs to be extended to model how fine-grained preferences are formed from coarse-grained attitudes, although such an extension would surely be fruitful. Rather, my point will be that the observations about the objects of preference I have made here ultimately count against a mentalistic and in favour of a in favour of a behavioural interpretation of the decision-theoretic models we do have.

There are two important lessons to be drawn from the observations I have made here for the mentalist appeal to folk psychology, which the next three sections will, in turn, elaborate on. Firstly, as we have seen, even if we adopt a mentalistic interpretation of preference, standard decision-theoretic models typically do not ascribe the coarser-grained mental attitudes that feature in ordinary folk psychological explanations. I will argue in the next section that this means that decision-theoretic models do not directly provide us with satisfactory folk psychological explanations no matter how we interpret preference, removing the mentalist's alleged edge in explanation. To the extent that decision-theoretic models can play an indirect role in folk psychological explanation, the behavioural interpretation of preference does just as well. And secondly, as just argued, if the preferences featuring in decision-theoretic models ascribe conative mental attitudes, then these are fine-grained mental attitudes. I will argue in sections

4 Against the Argument from Folk Psychology

As we have seen above, the case for mentalism relies crucially on the idea that decision theory offers folk psychological explanation, and does so only when preferences are interpreted mentalistically. Functionalists interpret preferences mentalistically *in virtue* of them supposedly playing a desire-like role in folk psychological explanation. Other mentalists claim that decision-theoretic models under a mentalist interpretation provide insight into an agent's actual reasoning or other independently measurable mental processes leading up to a choice, involving desire-like and belief-like attitudes, and are thus more explanatory than they would be under a behavioural interpretation. This section argues that the presupposition that mentalism makes possible a folk psychological explanation behaviourism about preference cannot afford is mistaken. Insofar as decision theory can help us arrive at folk psychological explanation, it can do so equally well under a behavioural conception of preference. Where folk psychological explanation is successful, this explanation relies on our hypothesizing coarser-grained attitudes that explain preferences over fine-grained outcomes. Whether fine-grained preferences themselves are understood mentalistically is irrelevant to this explanation.

To see this, let's go back to the coffee-drinking example. Suppose you ask on the second day, on which I drink tea: Why did you drink tea this morning? My flatmate's

ordinary folk psychological explanation of my choice says that I drank tea on the second day, because I wanted to keep my nerves down for my important meeting, and believed the tea would keep me less nervous than the coffee. For the reasons given above, a decision-theoretic explanation of my choice behaviour would involve making a model of my choice situation that specifies the outcomes I am choosing between in a fine-grained way: The outcome involving tea-drinking will also include a description of the tea's caffeine content, the fact that I have an important meeting later that day, that this particular tea is not horrible, that I am not using up the last tea bag, and so on. All of these things can potentially affect my choice behaviour in similar circumstances of interest. Next, the model attributes preferences to me: Most importantly, I prefer the fine-grained outcome involving drinking tea that day to the other outcomes available to me. In citing these preferences and interpreting them mentalistically, have you provided a folk psychological explanation of my choosing tea?

One way in which one might think you have is the following: In coming up with the model of the choice situation and specifying the outcomes I am choosing between, you had to think about all the things that could possibly be relevant for my choice. To do so, you probably considered the coarser-grained preferences and desires that you think I might have and that you would also standardly appeal to in ordinary folk psychological explanations. Moreover, when you then look at the differences between the outcome I chose and preferred on this occasion, the other outcomes available now, and the outcomes I chose on other occasions, two salient features of today's choice stand out: That drinking tea meant drinking the beverage with the lower caffeine content, and that I had an important meeting that day. In those circumstances, we can easily infer a desire of the type that standardly features in folk psychological explanation – a desire to keep one's nerves down for the meeting – from the preferences over fine-grained outcomes that feature in the decision-theoretic model. This desire, perhaps along with a clause that it was not outweighed by other factors, explains my choice in the way a standard folk psychological explanation does. The exercise as a whole could then be regarded as just an elaborate way of having provided the folk psychological explanation my flatmate gave in the first place: I drank tea on the second day, because I wanted to keep my nerves down for my important meeting, and believed the tea would keep me less nervous than the coffee. While the exercise might look pointless in this simple example, we can imagine it being genuinely enlightening, e.g. when the analysis of consumption data using decision-theoretic models in economics brings to the fore certain salient features that affect demand and make plausible an inference to a desire for such features amongst consumers.

Crucially, however, the process just described does not actually depend on interpreting preferences mentalistically. It works just as well if we think of preferences behaviourally: Having specified the outcomes I can choose between in a way that captures everything relevant to my choice, you ascribe to me a behavioural preference of the fine-grained

outcome involving drinking tea over each of the other available outcomes – capturing simply that I choose this outcome over the others. Looking at the salient differences between the outcomes available to me, the opportunity to keep my caffeine consumption low on the day of an important meeting stands out, and you infer a coarse-grained desire to keep my nerves down, which explains my choice. More generally, decision-theoretic models under a behavioural interpretation can help us identify systematic patterns in choice behaviour, which serves as a good basis to make inferences about coarser-grained attitudes that we can then appeal to in folk psychological explanations.

One might respond here that the inference from a pattern of preferences understood as fine-grained mental attitudes to a coarser-grained desire is safer than the inference from a pattern of choice behaviours to a coarser-grained desire. However, this point is moot when we consider that mentalistic preference and choice only come apart when the agent acts counter-preferentially. If she does act counter-preferentially, then appeal to her mentalistic preference and the underlying coarser-grained attitudes will not serve as a correct folk psychological explanation of her choices anyway, even if we do identify them correctly. And if she doesn't act counter-preferentially, then the inference from the choice behaviour to the underlying coarser-grained attitude is just as safe as the inference from the mentalistic attitude. And so either the mentalistic preference is not a guide to a correct folk psychological explanation at all, or it is just as good a guide as choice behaviour.

But perhaps fine-grained mentalistic preferences could contribute to folk psychological explanations beyond the inferences to coarser-grained attitudes they facilitate. A good way to test this is to imagine a case where the preferences over fine-grained outcomes we attribute to an agent are such that we cannot easily infer any coarser-grained desires or preferences, say, because the choice context is completely alien to us. For instance, suppose that we observe a child swapping cards with pictures and descriptions of mythical creatures on them with another child. We can offer very fine-grained descriptions of everything the child registers about the options open to her, and then hypothesize a preference for the fine-grained outcome involving owning the new card over the fine-grained outcome involving keeping her old card. Does citing merely this preference, without appeal to underlying coarser-grained attitudes, provide us with a folk psychological explanation?

If it is explanatory at all, this explanation is extremely thin. In fact, were we to ask the child, "Why did you pick that card?" and she answered "because I preferred it," we would assume she was just mocking the out-of-touch adults. What we really wanted to know is what it is about the card that the child likes (Does the mythical creature confer some advantage in a later game? Does it complete a set? Was this just a bluff? Was the card just prettier?). Ordinary folk psychological explanations would appeal to positive conative attitudes to such features of the card, or, in the terms of our previous discussion, coarser-grained attitudes. If we replace "it" in the child's explanation above with reference

to a fine-grained outcome describing all potentially relevant features of the card and choice situation, then we have learned no more. For any individual feature of the card or of the choice situation, that is, for any coarser-grained outcome comprising the fine-grained one, we will not know whether the child chose the card because of it, or in spite of it. We might glean such information from detecting a pattern of choice in favour of fine-grained outcomes that comprise a particular coarser-grained one. But, as just argued, we can also do that under a behavioural interpretation of preference.

There are two reasons why an explanation appealing only to fine-grained mentalistic preferences is much thinner than ordinary folk psychological explanations. The first is that, as we have noted above, fine-grained attitudes are at least partly explained by coarser-grained attitudes. By only citing a fine-grained attitude, we have thus at best cited the immediate mental cause of choice. By citing coarser-grained attitudes, ordinary folk psychological explanations, on the other hand, have the potential to offer deeper insight into the mental causes of choice.

Secondly, what makes the ordinary folk psychological explanations offered above superior is that, in virtue of citing coarser-grained attitudes, they cite more *general* attitudes: I generally prefer tasting coffee to tasting tea, in a wide variety of circumstances. Likewise, I generally prefer to be less nervous for important meetings. These attitudes have coarser-grained states of affairs as their object. These coarser-grained states of affairs can form part of many different fine-grained outcomes: I face different kinds of potential outcomes involving me tasting coffee all the time. Consequently, the coarser-grained desires and preferences apply in many different choice situations, and can contribute to the folk psychological explanation of many different choices. Explanatory force, in that case, comes not only from citing reasons and causes of choices, but from subsumption of the reasons and causes of one particular choice under general reasons and causes of the agent's choices.

Moreover, to the extent that citing a fine-grained mentalistic preference on its own does provide a very thin kind of folk psychological explanation, it would be very easy for the proponent of the behavioural interpretation to offer it as well. Granted, the revealed preference theorist is not offering this folk psychological explanation in virtue of the decision-theoretic model, while the mentalist about preferences is. But those proposing the mentalistic interpretation of preference must think that we can, in the circumstances where the model is to be useful, infer fine-grained mentalistic preferences from observing choice behaviour. If that is so, then offering the thin folk psychological explanation comes cheap for the proponent of the behavioural interpretation. Like the child in our example, she could just add to the presentation of her choice model: "... and the agent chose this option because she (mentalistically) preferred it."

Defenders of the mentalistic conception might object here that a stronger kind of explanatory force comes from citing not just the preferences immediately relevant to the choice to be explained, but from citing the agent's wider preference pattern of which those preferences form part. I think this is true, but doesn't count in favour of the mentalistic interpretation. One reason why citing the wider pattern is helpful is because often, it helps us infer more coarser-grained attitudes. As just argued, this is also possible under the behavioural interpretation. Another reason why citing a wider preference pattern can be helpful is because it helps us see the preference/choice to be explained as part of a systematic pattern of preferences/choices. However, it is not clear why such a unificationist kind of explanation shouldn't be just as strong if it cites a systematic pattern of choice behaviour, rather than a systematic set of mentalistic preferences.

I thus conclude that the mentalistic conception of preference does not actually hold an advantage over the behavioural one when it comes to offering folk psychological explanations. Folk psychological explanation gets its explanatory force from citing coarser-grained, and thus more general and more fundamental preferences and desires. These coarser-grained attitudes do not explicitly feature in decision-theoretic models, whether we think of preference mentalistically or not. Enthusiasts for folk psychological explanation thus need to lower their ambitions. We can potentially infer coarser-grainer attitudes from decision-theoretic models in order to offer a folk psychological explanation. But doing so does not require a mentalistic interpretation of preference. It is open to us under the behavioural interpretation just as much.

My argument here directly undermines the functionalist kind of mentalism about preference described above. Decision-theoretic preferences do not play a desire-like role in folk psychological explanations. Rather, they are better viewed as means for systematizing choice behaviour, which then, amongst other things, often helps us infer desire-like, coarser-grained attitudes that feature in folk psychological explanations. What about the more substantive kinds of mentalism considered above? The fact that providing ordinary folk psychological explanations does not necessitate a mentalist interpretation of preference is problematic for these kinds of mentalism, too, as it makes it harder to justify the potential loss of generality of mentalist decision-theoretic models. However, those mentalists may still insist that interpreting preferences mentalistically tells us something important about the actual mental processes that bring about choice, namely, that agents form fine-grained attitudes on the basis of, amongst other things, coarser-grained attitudes, and that choice is then based on these fine-grained attitudes. Even if citing the fine-grained attitude does not constitute, and is not necessary for a satisfactory folk psychological explanation, one might think that this window into the processes that bring about choice is an important strength of mentalism. The next two sections will argue, however, that the assumption that the mental processes that lead to choice typically involve the formation of fine-grained, all-things-considered attitudes can't be maintained.

Mentalism about preference would thus involve a serious restriction of the domain of applicability of standard decision theory, which ultimately counts against the view and in favour of the behavioural interpretation of preference.

5 Choice Without Fine-Grained Mental Attitude

Imagine an agent of the following kind in our coffee vs. tea scenario. It is undeniable that our agent has various coarser-grained attitudes to the states of affairs that comprise the outcomes she faces: These are consciously accessible to her, and they explain her choices. For instance, she prefers tasting coffee to tasting tea, and she desires not to be nervous at important meetings, or to be late to work. One morning, when she is neither running late nor has an important meeting that day, she goes and makes herself some coffee. Still quite tired, nothing much crosses her mind but the thought of some delicious coffee. She neither then, nor at any previous point in time ever consciously forms a fine-grained mental attitude to the fully specified outcome of "drinking a cup of coffee on a day when I don't have an important meeting, and I am not running late, and...". On introspection, it seems to her that she was motivated simply by her coarser-grained attitude of preferring tasting coffee to tasting tea, and that no other strong desires interfered at the time of choice. Perhaps, when she is asked about her choice, she then consciously forms a fine-grained mental attitude. But she does not think she held a fine-grained mental state at the time of choice.[20] Let's also suppose that her choices over time abide by the consistency conditions of standard revealed preference theory. But they only do so at the fine-grained level of specification of outcomes, since, had there been an important meeting, she would not have made herself coffee.

I think the most natural analysis of this case is that our agent really does never form a fine-grained mental attitude (at least not until she is asked about her choice later). But if that is so, decision theory under a mentalistic interpretation of preference cannot apply to this case, as the agent simply does not have the fine-grained mental states required. On a behavioural interpretation of preference, a decision-theoretic model can apply, as the agent's choice behaviour is in fact consistent at a fine-grained level of specification of outcomes. Moreover, this case does not seem outlandish, but is a rather familiar description of my early morning decision-making. In fact, much of the behaviour social scientists try to predict and explain with the use of decision-theoretic models, such as consumption behaviour, appears to be similarly unreflective. At first pass, then, it seems to be a significant restriction if standard decision theory could not apply to this case. In

[20] There is fairly solid evidence that when mentalistic preferences over anything but very coarse-grained outcomes are elicited, these are typically constructed "on the fly". See Lichtenstein and Slovic (2006) for an overview and Bettman (1979) for an early proponent.

the following I will argue that there are indeed no plausible grounds for the ascription of a fine-grained mental attitude in such a case, no matter what kind of mentalism we subscribe to.

For those mentalist decision theorists who hold that decision theory models conscious reasoning and deliberation (such as, at least in parts, Bradley (2017)), the case described above would indeed not warrant ascription of a mentalistic preference and thus could not be modelled decision-theoretically (or at best with a model featuring incomplete preference). While those decision theorists may not require that mentalistic preferences are occurrent, that is, consciously held at the time of choice, as we have said above, it is usually held that the standard completeness condition is only satisfied when agents have previously actively considered, and made their mind up about all options, and their preference remains at least consciously accessible. We have stipulated in our case above that the agent has not done so for the fine-grained outcomes open to her, at least not prior to her choice.

There is of course room for a kind of mentalism that, while aiming to provide insight into the mental processes that bring about choice, does not require the mental states involved to be actively formed or introspectively accessible. We have already encountered one type of mentalism that does not require this, namely the kind of functionalism which holds that decision-theoretic preferences are mental states just in virtue of playing a desire-like role in folk psychologically explaining an agent's choice behaviours – which does not require those states to be introspectively accessible. Regarding that view, we have already argued that, since fine-grained mental states don't feature in satisfactory folk psychological explanations, functionalist mentalism about preference that appeals to folk psychology is unpersuasive. But perhaps fine-grained mental states are an essential feature of our best scientific theories about the actual mental processes – conscious or not – that bring about choice. This would give us a solid basis for the ascription of such mental states even to unreflective agents like the one in our example.

However, theories of choice from the behavioural and cognitive sciences do not hold out much hope for the mentalist about decision-theoretic preferences. While there are some theories describing choice mechanisms that seem to presuppose that fine-grained mental states are formed,[21] most psychologists appear to agree that choice need not always involve the formation of a fine-grained attitude. For instance, it is now uncontroversial that choice is at least sometimes, and at least in part *affect-based*,[22] with affect-based decision-making being an important component of most dual process theories. Affect is a type of mental pro-attitude that is both taken to be coarse-grained, and to motivate fairly

[21] For instance, Goldstein and Einhorn (1987) suppose that choice is made in three stages, an encoding stage, where attributes of options are assigned some value, an evaluation stage, where these are integrated into an overall assessment of the option, which we can understand as a fine-grained mental attitude, and an expression stage, where the evaluations translate to responses.

[22] Zajonc (1980) was a famous early proponent of the importance of affect in decision-making.

directly, without much reflection.[23] We can potentially think of my choice of coffee in the morning as a choice motivated by a coarse-grained affective response to coffee. And even most of the more cognitively involved choice mechanisms described in the psychological literature involve some degree of 'selective processing', resulting in agents stopping short of forming attitudes as fine-grained as the mentalistic interpretation of preference in decision theory would require it.[24]

Now the mentalist might respond that, while individual choices can be explained without presupposing fine-grained mental attitudes, perhaps the consistency of choice behaviour (which we have granted in our example) cannot be explained without appealing to fine-grained mental states. In line with the idea that preferences are total or all-things-considered comparative evaluations, the thought could be that in environments where many factors are potentially relevant for agents' choices, some kind of mental weighing exercise is necessary in order for an agent to choose consistently. The fine-grained mental attitude would be the result of this weighing exercise.

My response here depends on whether the mentalist requires the weighing and resulting fine-grained attitude to be conscious or not. If it is to be conscious, then the rejoinder clearly fails, as there are mechanisms other than any conscious mental weighing conducted by the agent that could lead to consistency in choice behaviour. Binmore (2008) gives an evolutionary rationale for why and how agents' choices could end up consistent that does not presuppose conscious weighing. As money pump arguments and other pragmatic arguments for expected utility theory show, agents who violate the standard axioms of expected utility theory and revealed preference theory are at risk of exploitation and making sure losses – unless they use fairly sophisticated strategies for avoiding sure loss and exploitation. It is thus not implausible that even fairly unreflective agents could have learned to replicate only consistent choice behaviours, or evolved sub-conscious mechanisms to avoid inconsistent choice behaviour. This also provides a more plausible explanation of cases where decision theory has a good fit, but the ascription of conscious fine-grained mental states is otherwise unnatural. Consumer choice is arguably often affect-based, and decision theory is successfully applied here. And as mentioned above, there is even good

[23] For instance, Slovic et al. (2002) define affect as "the specific quality of goodness or badness (a) experienced as a feeling state (with or without consciousness) and (b) demarcating a positive or negative quality of a stimulus" (p. 397) and claim that affect can lead to rapid or automatic choice. Peters (2006) claims that affect can focus attention on specific features of an object of choice, but can also serve as a kind of common currency for the aggregation of different considerations. However, like Slovic et al. (2002), she also holds that affect can motivate fairly directly, and even if such aggregation has not taken place.

[24] For instance, Busemeyer et al. (2006) describe a choice mechanism they call 'decision field theory', whereby an agent only ever evaluates one aspect of an option at any one moment in time. Over time, attention shifts stochastically to other aspects of the option, and the evaluation is integrated into the previous evaluation. Once some threshold is reached, a decision is announced. This decision procedure is obviously consistent with the agent never forming an attitude with regard to all the aspects of the option that are potentially relevant. Several of the choice mechanisms discussed by Bettman et al. (1998) involve even more selective processing.

evidence that some non-human animals exhibit choice patterns consistent with decision theory.

One might worry that any of the affect-based choice mechanisms or choice strategies involving selective processing of only some aspects of outcomes described in the psychological literature can lead to decision-theoretic inconsistency if used exclusively, and that thus only conscious weighing can avoid it. However, it is generally acknowledged that agents use different choice strategies in different contexts. Which strategy is used in a particular context is to some extent a learned response to which strategy tends to lead to good decision-making in that context — and good decision-making surely includes avoiding exploitation. For instance, if I am a well-adapted agent, then on a day where I know I have an important meeting, my affect-based motivation to go make myself some coffee will hopefully be blocked, and I will think twice about my caffeine intake. Of course, it is undeniable that agents do sometimes violate the axioms of standard decision theories, and the choice strategies I described here can help to explain why. Yet, to the extent that these choice strategies are good heuristics, they can help agents conform by the axioms for the most part, so that decision theory would be empirically adequate for the most part, even for agents that do not engage in conscious weighing resulting in a fine-grained attitude.

Those mentalists who do not require the weighing process resulting in a fine-grained attitude to be a conscious process could now respond that the mental mechanisms that ensure both consistency in choice and that everything important to an agent is typically reflected in her choice behaviour could be considered to be types of sub-conscious weighing processes. In our case, whatever makes it the case that, were I to have an important meeting, my affect-based motivation to have some coffee is blocked (but not otherwise) could be considered to be a weighing process in the sense that my ultimate choice depends on how important two competing considerations (enjoyment and jitteriness) are in the particular context of choice.

I do not want to dispute that there is a sense in which we might call this a sub-conscious weighing process. The problem with this line of argument, rather, is that it seems like no explanatory power is lost by considering the result of the weighing to be a choice, rather than a fine-grained mental attitude which then results in a choice. And in those circumstances, parsimony seems to demand we do away with mentalistic preference and stick to a behavioural interpretation of preference – while acknowledging that the choice behaviours behavioural preferences capture may be the result of some unmodelled conscious or unconscious weighing process, one we have already noted standard decision theory has nothing to say about.

I thus think that adopting the mentalist interpretation of preference does come at a

significant cost of generality. Whatever our views on what might warrant the ascription of mental states, there is no plausible basis for ascribing fine-grained mental states of the type mentalistic decision-theoretic preferences would need to be to unreflective agents like the one in our example. Neither appeal to folk psychology, nor introspection, nor theories of choice from the behavioural and cognitive sciences warrant ascription of fine-grained mental states in such cases. Mentalist decision theory thus excludes more unreflective ways of making choices that are not uncommon, and that would have good fit with decision theory under a behavioural interpretation.

This loss of generality might not be as big a cost if we could at least say that the theory captures a class of cases that are of special interest for our explanatory projects. The proponent of the mentalist interpretation could insist, for instance, that the cases excluded by the mentalist interpretation of preference are all cases of choice behaviour that is not intentional or not based on reasons in the right way, or not rational on balance, and that decision theory is meant to be a theory of intentional choice, or choice based on reasons, or rational choice. The next section will argue, however, that having no fine-grained mental attitudes is in fact consistent with intentional, reason-based, and even instrumentally rational choice.

6 Rational Choice Without Fine-Grained Mental Attitude

To start, it would be quite uncharitable to assume that the case described above does not involve intentional or reason-based choice. Our agent appears to intentionally choose to drink the coffee, and she does so for a reason: She prefers the taste of coffee. If that is so, and we grant that she neither consciously nor sub-consciously forms a fine-grained mental attitude, forming a fine-grained mental attitude is not necessary for intentional and reason-based choice. This assessment indeed finds support in the more general literature on practical reason.

There are various different philosophical accounts of what it means to choose intentionally. Many of these require me to have an intention, or goal in action. But none of these accounts require my intention or goal in action to be to end up with a particular fine-grained outcome. In fact, there are entire debates within philosophy that rely on a distinction between the intended and unintended, but foreseen consequences of actions, where the intended consequences of one's action are coarser-grained outcomes.[25] In our coffee case, while the outcome of drinking coffee on an ordinary morning, for the purposes of a decision-theoretic model with wide applicability, will need to include the fact that our agent will leave the house 5 minutes later than if she had tea (after all, this will be

[25]See McIntyre (2014) for an overview of the debate on the Doctrine of Double Effect.

relevant under some circumstances), it would be implausible to say that leaving the house 5 minutes later is part of what she intended. Unless leaving 5 minutes later is something that she seeks independently, this is merely a foreseeable consequence of her choice, not part of what she intends.

Some other prominent accounts of what it means to choose intentionally claim that, in order to choose intentionally, we need to see what we choose "under the guise of the good."[26] That is, we must have a positive mental representation of the outcome we choose. This is also one standard position on what it means to choose for a reason, and indeed some philosophers have held that intentional choice just is choice for a reason or reasons.[27] But again, none of the standard accounts require agents to have formed a positive attitude to fine-grained outcomes. All that is required is that agents see *something* good in the object of their choice, that there is some respect in which they value it. That is, all that is required for the guise of the good thesis is that the agent has some positive attitude to the outcome they choose, and this could be a very coarse-grained attitude. This is taken to be consistent with thinking the outcome bad in many other respects, and even with the agent's attitudes on balance counting against it. But it is also obviously consistent with never forming a fine-grained attitude at all. Similarly, no other accounts of reason-based choice I know of require agents to have formed a fine-grained attitude.[28]

Now the proponent of the mentalist interpretation of preference might say that even if it is possible to act for a reason without forming a fine-grained mental attitude, choosing fully rationally requires us to have formed a fine-grained attitude. The thought could be the following: As previously noted, in many choice situations, agents have many different relevant coarser-grained attitudes that pull in different directions. If I were a more conflicted agent, for instance, I might find myself pulled in different directions on the question of whether to have coffee or tea, even on an ordinary morning. Perhaps I prefer the taste of coffee, but I also have reason to believe that the higher caffeine content is bad for my health. A rational agent, it seems, should find some kind of balance between these competing considerations. She should do what best serves her various desires on balance, and this is more demanding than simply choosing consistently. Perhaps doing what is best on balance requires agents to form a fine-grained mental attitude that properly weighs the different competing considerations at play, even if merely choosing consistently does

[26] See Tenenbaum (2013) for an overview.

[27] See, e.g., Davidson (1963).

[28] Schroeder (2007) briefly considers the possibility of a 'holistic' version of a Humean theory of reasons, according to which an agent has a reason for an action only if that action will "maximize the satisfaction of all of his desires *on balance*." (p.3) This version, on the face of it, looks like it might require the formation of a fine-grained attitude (although the discussion below will show that it doesn't). However, Schroeder quickly dismisses this view as obviously implausible. We all frequently have some reason to do things we ultimately ought not, and if we do them, we still acted for a reason. Similarly, if I failed to form a fine-grained attitude, but acted on some coarser-grained desire I have, I acted for a reason.

not.[29]

I do not wish to dispute that rationality requires agents to do what serves their desires best on balance, or do what is supported by the balance of reasons, whatever that may mean. However, my point is that to do so, it is not required to form a fine-grained mental attitude, either consciously or unconsciously. Perhaps, for instance, I actually do care about my health and there are some adverse health effects of drinking a lot of caffeine, which I know. It could still be true that my desires, on balance, support drinking coffee, as my desire for the taste of coffee outweighs these health considerations. By drinking the coffee, then, I do what best serves my desires on balance. And by drinking coffee, I do what best serves my desires on balance *even if* I never consciously form an attitude to the fine-grained outcome of drinking coffee on this particular morning. And, building on what I argued in the last section, it also need not be mysterious how I managed to do what best serves my desires on balance without consciously forming a fine-grained attitude. Perhaps, for instance, as an agent trained in making decisions quickly, only considerations that are likely to make a difference to my choice in a particular instance ever cross my mind. And, as argued in the last section, when it comes to sub-conscious weighing processes, there is no added explanatory value to supposing that these result in a fine-grained mental state, rather than directly in a choice.

Again, this is reflected in the non-formal literature on instrumental rationality.[30] One important reason why it has not been taken to be necessary to form a fine-grained attitude in order to count as instrumentally rational is that in many cases, especially when we need to act fast, doing what is supported by the balance of reasons is inconsistent with time- and cognitive labour-intensive deliberation – and it is not clear how one would form a fine-grained mental attitude without incurring at least some such costs.[31] As previously noted, we do not seem to come readily equipped with them, but rather construct them "on the fly". And forming a fine-grained attitude is not sufficient for instrumental rationality for similar reasons: If forming a fine-grained attitude that reflects what best serves one's

[29]Hausman (2012), for instance, treats the completeness axiom, the requirement that agents have preferences over all fine-grained outcomes they might face, as "a boundary condition on rational choice."' He continues, "an inability to compare alternatives is not itself a failure of rationality, but when people are unable to compare alternatives, they are unable to make a choice on the basis of reasons." (p.19) Strictly speaking, failures of completeness on a mentalistic interpretation of preference need not mean that agents are unable to compare outcomes, merely that they haven't done so. Still, Hausman seems to imply that a failure to have formed fine-grained mentalistic preferences means that one can neither act for reasons, nor act rationally.

[30]For instance, while Schroeder (2007) is not a 'holist' about reasons, he is a holist about the normative, in that he thinks agents ought to do what is supported by the balance of reasons. Moreover, his account of what makes reasons weightier appeals to the weight an agent would put on a reason in ideal deliberation. However, he does not require agents to actually fully deliberate and form a fine-grained attitude that tracks the balance of reasons. What agents are rationally required to do is simply to actually choose what is supported by the balance of reasons.

[31]Also see Angner (forthcoming) on the excessive cognitive demands Hausman's interpretation of preference would impose on agents.

desires on balance is laborious, then agents may well make mistakes. In fact, there is evidence that agents often make better decisions (measured in terms of ex post satisfaction) when they do not go through a process of weighing reasons.[32]

And so the kinds of choices decision theory can't accommodate under a mentalist interpretation of preference are not just unintentional, non-reason-based, and instrumentally irrational ones. It also excludes those intentional and instrumentally rational choices that, while abiding by the standard consistency requirements, are too unreflective to warrant the ascription of fine-grained mental attitudes. It is only in those cases where agents consciously form fine-grained attitudes that we seem to have a solid basis for ascribing mentalistic preferences. And so what the mentalist interpretation of decision theory seems to restrict the theory to are choices that are brought about by particularly thorough deliberation, resulting in an attitude to the fine-grained outcomes the theory deals with.[33] But this is not a particularly interesting class of decisions if we are interested in decision theory for the purpose of describing, predicting, explaining and rationalizing choice behaviour. And so this amounts to a very problematic loss of generality for the theory.

Now we are in a position to state what ultimately makes the behavioural interpretation of preference superior, even for those interested in providing folk psychological explanations, or insight into the mental causes of choice. As argued in section 3, due to their fine-grained nature, mentalistic preferences do not offer folk psychological explanation in their own right. Decision-theoretic models can potentially play an indirect role in facilitating inference to coarser-grained attitudes that do explain folk psychologically, but they can also do so under a behavioural interpretation of preference. The last two sections argued that in fact there are cases of unreflective, but intentional and instrumentally rational choice that can be captured with a behavioural decision-theoretic model, but where the ascription of fine-grained mental states, and thus mentalistic preferences is implausible. Mentalism thus holds no advantage over the behavioural interpretation when it comes to folk psychological explanation, and has a narrower range of applicability. It should be dropped in favour of the behavioural interpretation.

[32] See Wilson and Schooler (1991) and Wilson et al. (1993).

[33] This does seem to be acknowledged by some of the main defenders of the mentalist interpretation of preference. For instance, Sen (1973) writes, "There is, of course, the problem that a person's choices may not be made after much thinking or after systematic comparisons of alternatives. I am inclined to believe that the chair on which you are currently sitting in this room was not chosen entirely thoughtlessly, but I am not totally persuaded that you in fact did choose the particular chair you have chosen through a careful calculation of the pros and cons of sitting in each possible chair that was vacant when you came in. Even some important decisions in life seem to be taken on the basis of incomplete thinking about the possible courses of action." (p. 247) Sen takes this to count *against* the behavioural interpretation, as it means preferences are assigned even when deliberation is incomplete, which he takes to be implausible. Instead, I think it shows the opposite: Choices made after incomplete deliberation may be intentional, based on reasons and instrumentally rational, and the range of applicability of decision theory would be needlessly stifled if we could not apply the theory to such choices. Such examples provide an argument in favour of, and not against the behavioural interpretation.

7 Extension to Utility and Probability

This paper has advocated for a behavioural interpretation of preferences as they feature in decision theory. What shall we say about the related concepts of utility and probability? This will depend to some extent on what view we take on the relationship between preferences on the one hand, and utilities and probabilities on the other. One common view is to regard either just utility (in the von Neumann-Morgenstern framework), or both utility and probability (in the Savage framework) to be mere convenient constructs we use to represent binary preference (the possibility of which is established by the respective representation theorems). In the case of utility, this is often called the 'constructivist' interpretation of utility.[34] If we adopt this view, then a behavioural interpretation of utility (and probability) follows from interpreting preference behaviourally: Utility (and probability) are a mere convenient representation of binary preference, which in turn is just a convenient way to represent an agent's actual and hypothetical choices.

However, it is sometimes also argued that utility and probability functions are meant to capture mental attitudes that are not reducible to binary preference. E.g. it is sometimes argued that utility functions provide a cardinal measure of *strength* of desire and preference, going beyond a binary kind of attitude. In the case of utility, this is sometimes called the 'realist' interpretation of utility, and it potentially creates room for interpreting utility and probability mentalistically while granting a behavioural interpretation of preference. However, even if we adopt this interpretation of utility and probability, within decision-theoretic models, utility and probability functions still have the same kinds of objects as preferences, namely fine-grained outcomes. And then the first part of my previous argument applies *mutatis mutandis*.

On the realist picture, preferences should track (expected) real utility. Clearly, in our coffee example, the agent can't be modelled as maximizing a single utility function ranging over the coarse-grained outcomes describing merely what beverage she drinks. The utility function an agent is modelled as maximizing must range over fine-grained outcomes, for the same reasons that preference must have fine-grained outcomes as its object. And likewise, the probabilities used to calculate the expectation of utility must range over those same fine-grained outcomes (or the states that bring them about). And so, just as was the case for preference, the mentalistic interpretation of utility and probability as they feature in decision-theoretic models involves the ascription of fine-grained mental states which there is often no plausible basis to ascribe to agents. This is not to say that the notions of real utility and probability functions as graded measures of desire and belief are generally implausible, in particular when we think about probabilites and utilities with coarser-grained objects. The point is merely that the utilities and probabilities that feature in the

[34] See Dreier (1996) and Velleman (1993/2000) for defences of constructivism about utility

decision-theoretic models we use to explain choice behaviour have fine-grained objects, and it is the ascription of such fine-grained mental states that is both often implausible and not very explanatory in its own right.

Granted, where it is plausible to ascribe such fine-grained mental states, there is a potential explanatory gain from showing how preferences are derived from utilities and probabilities not themselves reducible to preference. But the fine-grained nature of the objects of utility and probability still means these explanations are unlike, and less informative than ordinary folk psychological explanations, in that they don't tell us what it is about outcomes that makes them choice-worthy. It is thus not clear whether the potential gain in explanatoriness is worth the loss in generality. Moreover, I take there to be good reasons against the realist interpretation of utility, at least.[35]

8 Conclusions

It is common to think of decision theory as capturing, with just a few more "frills", the platitudes about belief, desire and choice that ordinary folk psychological explanation draws on. The observations I have made in this paper about the objects of preference, utility and probability in decision-theoretic models should curb this enthusiasm. Preferences in decision-theoretic models have fine-grained outcomes as their objects, as it is only at this level of description that we find the consistency in preference decision-theoretic models presuppose. The idea that agents have mental states that correspond to such fine-grained preferences is much more controversial than the platitudes about belief, desire and choice that folk psychology appeals to. In fact, I have argued that agents often choose without being plausibly ascribed fine-grained mental states, and that their choices can nevertheless be intentional, reason-based and instrumentally rational.

Preferences as they feature in decision-theoretic models and preferences as they feature in ordinary folk psychological explanation differ at least in their objects: The former have fine-grained objects, the latter coarser-grained ones. I have argued that this means that decision-theoretic models can only ever help furnish satisfactory folk psychological explanations indirectly. No matter how we interpret preference, the decision-theoretic model alone can at best furnish an extremely thin kind of causal explanation of an agent's choices. Successful folk psychological explanation only comes from inferring coarser-grained attitudes from a pattern of preferences, which decision-theoretic models can help facilitate.

An enthusiast for folk psychological explanation might be satisfied that this indirect role for decision-theoretic models is still crucial, and that facilitation of folk psychological

[35] As discussed, e.g., in Broome (1991) or Okasha (2016).

explanation is an important function of decision theory. I am happy to grant that, as my main goal here has been to show that the standard case for a mentalistic interpretation of the preferences in decision-theoretic models has been undercut. And this is because decision-theoretic models can play a facilitating role in inferring the coarser-grained mental states featuring in successful folk psychological interpretation whether we think of preference mentalistically or behaviourally. As the models have wider applicability under a behavioural interpretation, in fact enthusiasts for folk psychological explanation have a reason to stick to a behavioural interpretation of preference.

References

Sidney N. Afriat. The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77, 1967.

Erik Angner. What preferences really are. *Philosophy of Science*, forthcoming.

James R. Bettman. *An information processingtheory of consumer choice*. Addison-Wesley, 1979.

James R. Bettman, Mary Frances Luce, and John W. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25:187–217, 1998.

Ken Binmore. *Rational Decisions*. Princeton University Press, 2008.

Richard Bradley. *Decision Theory with a Human Face*. Cambridge University Press, 2017.

John Broome. *Weighing Goods*. Blackwell, 1991.

Jerome R. Busemeyer, Joseph G. Johnson, and Ryan K. Jessup. Preferences constructed from dynamic microprocessing mechanisms. In Sarah Lichtenstein and Paul Slovic, editors, *The Construction of Preference*, pages 220–234. Cambridge University Press, 2006.

Paul Churchland. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78(2):67–90, 1981.

Christopher Clarke. Preferences and positivist methodology in economics. *Philosophy of Science*, 83(2):192–212, 2016.

Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

Franz Dietrich and Christian List. Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics and Philosophy*, 32(2):249–281, 2016.

James Dreier. Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision*, 40(3):249–276, 1996.

Allan Gibbard. Preferences and preferability. In Christoph Fehige and Ulla Wessels, editors, *Preferences*, pages 239–259. Walter de Gruyter, 1998.

W. M. Goldstein and H. J. Einhorn. Expression theory and the preference reversal phenomena. *Psychological Review*, 94:236–254, 1987.

Till Gruene. The problems of testing preference axioms with revealed preference theory. *Analyse Kritik*, 26:382–397, 2004.

Faruk Gul and Wolfgang Pesendorfer. The case for mindless eonomics. In Andrew Caplin and Andrew Schotter, editors, *The Foundations of Positive and Normative Economics*. Oxford University Press, 2008.

Daniel Hausman. Problems with realism in economics. *Economics and Philosophy*, 14(2): 185–213, 1998.

Daniel Hausman. *Preference, Value, Choice, and Welfare*. Cambridge University Press, 2012.

H. S. Houthakker. Revealed preference and the utility function. *Economia*, 17:159–174, 1950.

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1965/1983.

James Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

John H. Kagel, Raymond C. Battalio, and Leonard Green. *Economic Choice Theory: An Experimental Analysis of Animal Behavior*. Cambridge University Press, 1995.

Tobias Kalenscher and Marijn van Wingerden. Why we should use animals to study economic decision making – a erspective. *Frontiers in Neuroscience*, 5(82):1–11, 2011.

David Lewis. Radical interpretation. *Synthese*, 23:331–344, 1974.

Sarah Lichtenstein and Paul Slovic. The construction of preference: An overview. In Sarah Lichtenstein and Paul Slovic, editors, *The Construction of Preference*, pages 1–40. Cambridge University Press, 2006.

Ian Little. A reformulation of the theory of consumer's behaviour. *Oxford Economic Papers*, 1(1):90–99, 1949.

Michael Mandler. A difficult choice in preference theory: Rationality implies completeness or transitivity but not both. In Elijah Millgram, editor, *Varieties of Practical Reasoning*, pages 373–402. MIT Press, 2001.

Alison McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, `https://plato.stanford.edu/archives/win2014/entries/double-effect/`, winter 2014 edition, 2014.

R. E. Nisbett and Timorthy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977.

Samir Okasha. On the interpretation of decision theory. *Economics and Philosophy*, 32: 409–433, 2016.

Ellen Peters. The functions of affect in the construction of preference. In Sarah Lichtenstein and Paul Slovic, editors, *The Construction of Preference*, pages 454–463. Cambridge University Press, 2006.

Philip Pettit. Decision theory and folk psychology. In Michael Bacharach and Susan Hurley, editors, *Foundations of Decision Theory: Issues and Advances*, pages 147–175. Blackwell, 1991.

Peter Railton. The affective dog and its rational tale: Intuition and attunement. *Ethics*, 124(4):813–859, 2014.

Don Ross. Estranged parents and a schizophrenic child: choice in economics, psychology and neuroeconomics. *Journal of Economic Methodology*, 18(3):217–231, 2011.

L. R. Santos and M. Keith Chen. The evolution of rational and irrational economic behavior: evidence and insight from a non-human primate species. In E. Fehr P. W. Glimcher, C. F. Camerer and R. A. Poldrack, editors, *Neuroeconomics: Decision Making and the Brain*, pages 81–93. Academic Press, 2009.

Leonard Savage. *The Foundations of Statistics*. Wiley, second revised edition edition, 1972.

Mark Schroeder. *Slaves of the Passions*. Oxford University Press, 2007.

Amartya Sen. Behaviour and the concept of preference. *Economica*, 40(159):241–259, 1973.

Eldar Shafir, Itamar Simonson, and Amos Tversky. Reason-based choice. *Cognition*, 49: 11–36, 1993.

Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. Macgregor. The affect heuristic. In T. Gilovich, D. Griffin, and D. Kahnemann, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 397–420. Cambridge University Press, 2002.

Katie Steele. Choice models. In Nancy Cartwright and Eleonora Montuschi, editors, *Philosophy of Social Science: A New Introduction*, pages 185–207. Oxford University Press, 2014.

H. Orri Stefansson and Richard Bradley. What is risk aversion? *British Journal for the Philosophy of Science*, 70(1):77–102, 2019.

Sergio Tenenbaum. Guise of the good. In Hugh LaFollette, editor, *The International Encyclopedia of Ethics*. Wiley-Blackwell, 2013.

David Velleman. The story of rational action. In *The Possibility of Practical Reason*. Oxford University Press, 1993/2000.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Kate Vredenburgh. A unificationist defense of revealed preferences. *Economics and Philosophy*, 2019.

Timothy D. Wilson and J. W. Schooler. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60: 181–192, 1991.

Timothy D. Wilson, Douglas J. Lisle, Jonathan W. Schooler, Sara D. Hodges, Kristen J. Klaaren, and Suzanne J. LaFleur. Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin*, 19:331–339, 1993.

R. B. Zajonc. Feeling and thinking: Closing the debate over the independence of affect. *American Psychologist*, 35:151–175, 1980.