Folk Psychology and the Interpretation of Decision Theory

July 17, 2019

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 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 (forthcoming).

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 objects, such as, for instance, the desire to drink some coffee. However, preferences, utilities and probabilities as they feature in decision theory have fine-grained objects, such as the outcome of drinking coffee under a specific set of circumstances. The mentalistic interpretation of preference, utility and probability thus involves the ascription of fine-grained mental states. But I argue that such fine-grained mental states cannot play the role mentalists would like mentalistic preference, utility and 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;

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

Best explanations of choice behaviour, be they folk psychological or by reference to theories from the behavioural sciences, reference coarse-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 coarse-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 coarse-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 present a *prima facie* case in favour of the behavioural interpretation of preference which appeals to its greater generality. This serves as a backdrop to the mentalist argument from folk psychology that is the main target of this paper, described in Section 3. Section 4 lays the groundwork for the main argument of this paper by arguing that the attitudes featuring in ordinary folk psychological explanations have coarse-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. Sections 5 and 6 show that this is a much stronger commitment than the platitudes of folk psychology. Section 5 argues that there are many cases of decision-theoretically consistent choice behaviour where ascription of fine-grained mental states is both intuitively implausible and not presupposed by our best explanations. Section 6 argues that this is also specifically the case for intentional, reason-based and instrumentally rational choice. Section 7 addresses the argument from folk psychology directly by arguing that folk psychological explanation is only successful with reference to coarse-grained mental states. To the extent that decision theoretic models can facilitate inference to such coarse-grained mental states, they can do so better under the behavioural interpretation of preference, due to its more general applicability. Section

8 extends my argument to utility and probability, and Section 9 concludes.

2 A *prima facie* Case for the Behavioural Interpretation of Preference

The behavioural interpretation of preference has rarely been cast in a favourable light by philosophers.[6] I would nevertheless like to start here with what I take to be a strong *prima facie* case in favour of a behavioural interpretation of preference, to use as a backdrop for the mentalist argument from folk psychology. This argument is notably not behaviourist, in that it doesn't rule out appeal to mental states where it is necessary for the explanatory purpose at hand. It merely points to advantages of black-boxing the causes of choice for the purposes of decision theory.

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 (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.

Proponents of mentalistic and behavioural interpretations of preference disagree about what it means to ascribe preferences to agents within a decision-theoretic framework. While mentalists think that preference ascriptions are ascriptions of conative mental states (such as comparative evaluations), proponents of behavioural interpretations take them to be mere re-descriptions of agents' choice behaviours. According to behavioural interpretations of preference, roughly, what it means to prefer option a to option b is just that one actually does, or hypothetically would choose option a rather than option b when faced with a choice between them. This understanding of preference is very pop-

[6] This has recently seemed to change, however. See Clarke (2016) and Vredenburgh (2019) for more favourable treatments.

ular 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 therefore 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.

Before I turn to the *prima facie* case in favour of the behavioural picture, 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 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,[9] 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 with the argument from folk psychology is significantly stronger, as it involves the ascription not only of cognitive states, but also of conative states. And moreover, the *prima facie* case in favour of the behavioural interpretation of preference, which I shall make presently, is not undermined by the concession of a limited mentalism about the description of options. And that is because, as just mentioned, this *prima facie* case does not rely on any strong commitment to a more general behaviourism.[10]

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

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

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

[9] See [redacted].

[10] See Clarke (2016) for a similar argument in the context of economics. It should also be added, in the light of my argument below, that this concession to mentalism does not involve the ascription of fine-grained mental states, as, e.g. the mentalist interpretation of probability assignments would.

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, 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, 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 theory to find fruitful application is that they consistently respond to what they believe or register about their choice environments. 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.

3 The Argument from Folk Psychology

There is an obvious response here for the mentalist. There are various predictive and explanatory goals we might pursue with decision theory. According to the mentalist,

[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.

one crucial one requires the ascription of mental states via preference. Decision theory under a behavioural interpretation of preference might 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 preference fits into a pattern of preferences.[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 allows 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 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.

Leading on from this, proponents of the mentalist interpretation of preference may also dispute that there is a significant loss in generality when moving to a mentalist interpretation. If decision theory under a mentalistic interpretation provided us with something like the folk psychological explanations we are all familiar with, then maybe the theory is not saying anything too controversial about what causes choice. At least in cases where agents act intentionally, and we can cite reasons for their choices, we can usually give some kind of folk psychological explanation. If decision theory under a mentalistic interpretation of preference applies to the same kinds of cases and offers just a more formal version of the folk psychological explanations we are all used to providing, then the theory still has a good deal of generality. In further support of this claim to generality, recent proponents of the mentalist interpretation have stressed that preference can be understood as a kind of summary attitude that takes into account all of the agent's motivations, acknowledging that these can be of diverse kinds. This is why Hausman

[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.

(2012) insists on preference being a 'total' comparative evaluation, and Bradley (2017) and Joyce (1999) take it to be an 'all-things-considered' evaluative judgement. This makes a theory that is based on a mentalistic notion of preference seemingly consistent with the variety of motivation we observe.

In short, the mentalist appeals to folk psychology to argue that one crucial explanatory goal of decision theory does require preference to ascribe mental states, and others are not undermined by doing so: The mentalist interpretation allows for a deeper kind of explanation, while not resulting in a significant loss of generality. In the following, I do not wish to dispute that choice is caused by the kinds of mental states folk psychological explanations appeal to,[15] that it can be explained by citing them, and that providing such explanations 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 comes with no explanatory advantages, and a more significant loss of generality than one might think.

4 On the Objects of Preference

My argument will be based on an observation about the objects of preference in decision theory and the objects of preference in ordinary preference-talk and ordinary folk psychological explanation. 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 [redacted] drank coffee on the first day because she prefers the taste of coffee to the taste of 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 my options as simply "drinking coffee" and "drinking 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 options. But suppose my flatmate also has reason to believe that I am not indifferent between the options on any of the days. 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

[15] Though note that folk psychology is controversial as a theory or model of how people in fact make choices, even if it correctly captures how people standardly explain each other's choices and attribute mental states to each other. See, e.g., Churchland (1981).

my beverage of choice. To capture my choice behaviour with a consistent preference relation, it seems he has to describe my options in a more fine-grained way. In particular, at a minimum, he should include in the description of my options not only the taste of the beverage involved, facts about 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 in the description facts about 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 fine-grained options, and thus the various combinations of circumstances that might affect my choice.

The fine-grained nature of the options that feature in decision-theoretic models distinguishes the preferences decision theory invokes from preference as it is typically appealed to in ordinary language. Agreeing with my flatmate's original folk psychological assessment of me, I might say that I prefer the taste of coffee to the taste of tea. While this preference featured in my flatmate's folk psychological explanation of my choice behaviour above, it will not feature as such in a decision-theoretic model of my choice situation. That model will feature only preferences over specific outcomes involving me tasting coffee or tea, combined with various other relevant features of the options open to me. At best, we can read off a general preference for the taste of coffee from the fact that I tend to prefer options that involve me drinking coffee to ones that involve me drinking tea, unless other more important considerations lead me to prefer tea.

More generally, decision theorists usually assume that agents' options have to be described in enough detail that everything that might affect an agent's choice 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 options 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.

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. The standard objects of preferences in decision theory are such prospects, or lotteries over them. Each such prospect 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 will be instantiated by many different finer grained prospects. In the following, I will thus speak of coarse-grained objects of desire or preference and fine-grained objects, outcomes or options. I submit that the ordinary notions of preference and desire that feature in standard folk psychological explanations are what Pettit calls preferences and desires for properties, or coarse-grained objects. I will in the following also refer to these as coarse-grained attitudes. The preferences that feature in decision theory, on the other hand, are preferences over fine-grained objects, or options. If these preferences 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 options, then these are at least partly explained by our coarse-grained attitudes.[16] For instance, if I prefer the fine-grained option involving me drinking coffee from my rainbow streetcar mug this morning to the fine-grained option 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 options, if we form them, are at least partly the result of weighing various coarse-grained attitudes to the different coarse-grained objects that comprise the fine-grained options.

There are two important lessons to be drawn from these observations for the argument from folk psychology, which the next two sections will, in turn, elaborate on. Firstly, if the preferences featuring in decision-theoretic models ascribe conative mental attitudes, then these are fine-grained mental attitudes. In the following three sections, I argue that this results in a significant loss of generality, as ascriptions of fine-grained mental states are often intuitively implausible, and not presupposed by our best explanations of choice behaviour. The claim that agents regularly do form such attitudes is certainly a much more controversial claim than the platitudes that are involved in ordinary folk psychological explanation. And secondly, even on a mentalistic interpretation of preference, standard decision-theoretic models do not ascribe the coarse-grained mental attitudes that feature in ordinary folk psychological explanations. Pettit argued that this renders standard decision theory incomplete. I will argue 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.

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

5 Choice Without Fine-Grained Mental Attitude

Imagine an agent of the following kind. Our agent has various coarse-grained attitudes to the features of options she might face. For instance, she prefers the taste of coffee to the taste of 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 other point in time ever consciously forms a fine-grained mental attitude to the fully described option 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 is motivated simply by her coarse-grained attitude of preferring the taste of coffee to the taste of tea. 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 description of options, 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. 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 description of the options. 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.

Introspection is of course often an imperfect guide to what mental states are in fact important in bringing about choice.[17] We often ascribe mental states that are not conscious or introspectively accessible in cases where those mental states are part of our best explanations of agents' choice behaviours or feelings. Indeed, functionalists about mental states think that mental states just are those states that play a certain role in explaining an agent's choice behaviours. Perhaps, then, appealing to fine-grained mental states is an essential feature of our best theories of what explains choice, even if these mental states need not be conscious. This would give us a solid basis for the ascription of such mental states even to unreflective agents.

[17] See, for instance, Nisbett and Wilson (1977) for evidence of imperfect introspection about the mental causes of choice.

A first thing to note here is that folk psychological explanation, at least, does not require appeal to a fine-grained mental state. In fact, it seems perfectly fine to simply explain my choice with appeal to my coarse-grained desire for coffee. And it would be begging the question against the behavioural interpretation of preference to cite decision theory itself to make the case that we need to appeal to fine-grained mental states to explain choice.

How about theories of choice from the behavioural and cognitive sciences? While there are some theories describing choice mechanisms that seem to presuppose that fine-grained mental states are formed,[18] 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*,[19] 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 directly, without much reflection.[20] 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.[21]

Now the mentalist might respond that, while individual choices can be explained without presupposing fine-grained mental attitudes, perhaps the consistency of choice be-

[18] 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.

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

[20] 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. In cognitive science, it is now acknowledged that there might be more sophisticated processing going on 'under the hood' in generating affective responses than previously thought. See Railton (2014) for a philosophical discussion. These processes could potentially involve fine-grained representations of options. However, even if this were so, appealing to these representations won't help the mentalist hoping to offer rationalizations akin to folk psychological explanation, as unlike the resultant coarse-grained affective response, they are assumed to occur at the sub-personal level.

[21] 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.

haviour (which we have granted in our example) cannot be explained without appealing to fine-grained mental states. In line with Hausman's claim that preferences are total 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.

However, there are mechanisms other than any mental weighing conducted by the agent that could lead to consistency in choice behaviour. For instance, 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 avoid inconsistent choice behaviour, simply by learning to avoid the kind of behaviour that leads them to make such losses. This also provides a more plausible explanation of cases where decision theory has a good fit, but the ascription of 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 evidence that some non-human animals exhibit choice patterns consistent with decision theory.[22]

I thus think that adopting the mentalist interpretation of preference does come at a cost of generality: We are excluding 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

[22]One might worry here that any of the affect-based choice mechanisms or choice strategies involving selective processing of only some aspects of options described in the psychological literature can lead to decision-theoretic inconsistency if used exclusively. However, this need not speak against applying decision theory to agents who do not form fine-grained attitudes. And this is because it is also 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.

meant to be a theory of intentional choice, or choice based on reasons, or rational choice. Proponents of the argument from folk psychology, in particular, might be happy to accept such a restriction, as these might be thought to be the only plausible candidates for folk psychological explanation anyway.

6 Rational Choice Without Fine-Grained Mental Attitude

Having no fine-grained mental attitudes is in fact consistent with intentional, reason-based, and even instrumentally rational choice. To start, it seems 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 does not form 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 option. 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 coarse-grained objects.[23] In our coffee case, while the description of the option of drinking coffee on an ordinary morning, for the purposes of the decision-theoretic model, 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 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."[24] That is, we must have a positive mental representation of the option 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.[25] But again, none of the standard accounts require agents to have formed a positive attitude to their fine-grained options. All that is required is that agents see *something* good in the

[23]See McIntyre (2014) for an overview of the debate on the Doctrine of Double Effect.
[24]See Tenenbaum (2013) for an overview.
[25]See, e.g., Davidson (1963).

object of their choice, that there is some respect in which they value them. That is, all that is required for the guise of the good thesis is that the agent has some positive coarse-grained attitude to the option they choose. This is taken to be consistent with thinking the option 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.[26]

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 coarse-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 not.

Hausman (2012), for instance, seems to think so. Regarding the completeness axiom, the requirement that agents have preferences over all fine-grained options they might face, he acknowledges that abiding by it would be a "remarkable intellectual achievement" resulting from "an unmodeled process of exhaustive comparative evaluation" (p.18). Nevertheless, he thinks that completeness is "a boundary condition on rational choice. 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 options, merely that they haven't done so. Still, Hausman seems to imply that a failure to have formed fine-grained preferences means that one can neither act for reasons, nor act rationally. I have already argued that one can act for reasons without having formed fine-grained attitudes. But I also think one can act fully rationally without having done so.

[26]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 coarse-grained desire I have, I acted for a reason.

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, the point is that to do so, it is not required to form a fine-grained mental attitude. 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 form an attitude to the fine-grained option of drinking coffee on this particular morning. It also need not be mysterious how I managed to do what best serves my desires on balance without 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. Moreover, it seems like forming a fine-grained mental attitude is not even sufficient for being instrumentally rational. Even apart from the possibility of counter-preferential choice, agents may also form fine-grained attitudes that do not correctly reflect the balance of reasons.

Again, this is reflected in the non-formal literature on instrumental rationality.[27] 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.[28] We certainly do not come readily equipped with them.[29] 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 desires on balance is laborious, then agents may well make mistakes. In fact, there is evidence that agents often make better decisions (measured in

[27]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.

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

[29]There is fairly solid evidence that when mentalistic preferences over anything but very coarse-grained objects are elicited, these are typically constructed "on the fly". See Lichtenstein and Slovic (2006) for an overview and Bettman (1979) for an early proponent. Note that this also makes the completeness axiom featured in standard versions of expected utility theory highly implausible on a mentalistic interpretation of preference, even if choice always involved the formation of a fine-grained attitude. On a behavioural interpretation, on the other hand, completeness is only controversial regarding options for which it is questionable whether they are possible objects of choice. This is one respect, at least, in which the theory's core assumptions are *more* likely to hold on the behavioural interpretation of preference. See also Mandler (2001) on this.

terms of ex post satisfaction) when they do not go through a process of weighing reasons.[30]

And so the kinds of choices decision theory excludes under a mentalist interpretation of preference are not just unintentional, non-reason-based, and instrumentally irrational ones. And it is not even the case that all the choices that are captured by the theory under a mentalist interpretation are instrumentally rational (as mental preferences may fail to reflect the balance of reasons). Instead, what the mentalist interpretation of decision theory restricts the theory to are choices that are brought about by particularly thorough deliberation, resulting in an attitude to the fine-grained options the theory deals with.[31] 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. This class of cases certainly has no special connection to folk psychological explanation. And so this amounts to a loss of generality for the theory that should be problematic even for those who see the main value in decision theory in providing folk psychological explanation.

Now a last redeeming feature of the mentalist interpretation of preference might be that, in the restricted class of cases in which it does apply, it enables us to give explanations that are deeper than the kinds of explanations that the proponent of the behavioural interpretation can offer. If that is so, there would be a generality/explanatoriness trade-off that the mentalist might want to resolve in favour of explanatoriness. The next section argues this claim, too, is mistaken.

7 Against the Argument from Folk Psychology

There are surely some cases where agents can credibly be ascribed fine-grained conative attitudes of the kind standard decision-theoretic models presuppose on the mentalist interpretation. These would feature agents more reflective than I am before my first morning coffee. In these cases, does the mentalist interpretation at least buy us the deeper kind

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

[31] 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.

of explanation that the argument from folk psychology promised? It does not. Insofar as decision theory helps to capture folk psychological explanations, it can do so equally well under a behavioural conception of preference. Where folk psychological explanation is successful, this explanation relies on our hypothesizing coarse-grained attitudes that explain preferences over fine-grained options. 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, but let us suppose we are dealing with my much more reflective twin Y [redacted]. All other features of the case remain the same. Suppose you ask on the second day, on which Y drinks tea: Why did Y drink tea this morning? As in X's case, Y's flatmate's ordinary folk psychological explanation of her choice says that Y 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. For the reasons given above, a decision-theoretic explanation of Y's choice behaviour would involve making a model of Y's choice situation that describes her options in a fine-grained way: The description of the option of drinking tea will include the fact that she will taste tea, but also the fact that this tea has a low caffeine content, that Y has an important meeting later that day, that this particular tea is not horrible, that Y is not using up the last tea bag, and so on. All of these things can potentially affect her choice behaviour. Next, the model attributes preferences to Y: Most importantly, Y prefers the fine-grained option involving drinking tea that day to the other options available to her. As Y is very reflective, she actually has formed conative mental states corresponding to the preferences featuring in the decision-theoretic model. In citing these preferences, have we provided a folk psychological explanation of her choosing tea?

One way in which we might think we have is the following: In coming up with the model of the choice situation, we had to think about all the things that could possibly be relevant for Y's choice. To do so, we considered the coarse-grained preferences and desires that we think she might have and that we would also standardly appeal to in ordinary folk psychological explanations. Moreover, when we then look at the differences between the option she chose and preferred on this occasion, the other options available now, and the options she 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 Y 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 options that feature in the decision-theoretic model. This desire explains Y's 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 Y's flatmate gave in the first place: Y 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. 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 described Y's options in a way that captures everything relevant to her choice and is consistent with her relevant beliefs, we ascribe to her a behavioural preference to choose the fine-grained option involving drinking tea over each of the others. Looking at the salient differences between the options available to her, the opportunity to keep her caffeine consumption low on the day of an important meeting stands out, and we infer a desire to keep her nerves down, which explains her choice.

One might respond here that the inference from a pattern of preferences understood as fine-grained mental attitudes to a coarse-grained desire is safer than the inference from a pattern of choice behaviours to a coarse-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 coarse-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 coarse-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 coarse-grained attitudes they facilitate. Perhaps the best 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 coarse-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 option involving owning the new card over the fine-grained option involving keeping her old card. Does citing merely this preference, without appeal to underlying coarse-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.

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, if agents form fine-grained attitudes at all, these stand at the end of deliberation, and are at least partly explained by coarse-grained attitudes. By only citing the fine-grained attitude, we have at best cited the immediate mental cause of choice. By citing coarse-grained attitudes, ordinary folk psychological explanations, on the other hand, give greater insight into the agent's deliberation, by citing the mental causes of fine-grained preferences.

Secondly, what makes the ordinary folk psychological explanations offered above superior is that they cite much more general attitudes: Y and I generally prefer the taste of coffee to the taste of tea, in a wide variety of circumstances. Likewise, we generally prefer to be less nervous for important meetings. These attitudes have coarse-grained states of affairs, or, in Pettit's terms, properties of prospects, as their object. These coarse-grained states of affairs can be part of many different fine-grained options: I face different kinds of options involving me tasting coffee all the time. Consequently, the coarse-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 coarse-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. However, it is not clear why this 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 offer an advantage over the behavioural one when it comes to offering folk psychological explanations, even in the restricted class of cases where it is plausible to assume fine-grained mental attitudes. Folk psychological explanation gets its explanatory force from citing general, or coarse-grained preferences and desires. These coarse-grained attitudes do not explicitly feature in decision-theoretic models, whether we think of preference mentalistically or not. We can potentially infer them 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.

Moreover, drawing on our analysis in the last section, I think the argument can be strengthened: In fact, the behavioural interpretation can facilitate folk psychological explanation in situations where the mentalistic interpretation would render decision-theoretic models inapplicable. This is so in choice situations where agents act, and act for reasons, without ever forming a fine-grained mental attitude. A behavioural decision theoretic model of the choice behaviour of such agents may allow us to infer the coarse-grained attitudes that formed the reasons and causes for their choices, even in the absence of a fine-grained mental attitude. The mentalist about preference, on the other hand, would have to say that such agents lack the kinds of preferences necessary for the decision-theoretic model to even apply. And thus, the behavioural interpretation in fact puts us in a better position to at least facilitate valid folk psychological explanations.

8 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.[32] 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 very fine-grained states of affairs. And then the first part of my previous argument applies *mutatis mutandis*: The mentalistic interpretation of utility and probability as they feature in decision-theoretic models involves the ascription of implausibly fine-grained mental states. 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.[33]

9 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 successful decision-theoretic models have fine-grained descriptions of options as their objects, as it is only at this level of description that we find the consistency in prefer-

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

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

ence 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 coarse-grained. As I have explored in the last section of this paper, 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 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 coarse-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.

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*, forthcoming.

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.