Synthese
https://doi.org/10.1007/s11229-019-02308-4

Betting against the Zen Monk: on preferences and partial belief

Edward Elliott[1]

Received: 1 April 2019 / Accepted: 26 June 2019

Abstract
According to the preference-centric approach to understanding partial belief, the connection between partial beliefs and preferences is key to understanding what partial beliefs are and how they're measured. As Ramsey put it, the 'degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it' (in: Braithwaite (ed) The Foundations of Mathematics and Other Logical Essays, Routledge, Oxon, pp 156–198, 1931). But this idea is not as popular as it once was. Nowadays, the preference-centric approach is frequently dismissed out-of-hand as behaviouristic, unpalatably anti-realist, and/or prone to devastating counterexamples. Cases like Eriksson and Hájek's (Stud Log 86(2):183–213, 2007) preferenceless *Zen monk* and Christensen's (Philos Sci 68(3):356–376, 2001) *other roles* argument have suggested to many that any account of partial belief that ties them too closely to preferences is irretrievably flawed. In this paper I provide a defence of preference-centric accounts of partial belief.

Keywords Preferences · Partial belief · Betting interpretation · Zen monk · Representation theorems · Functionalism

1 Introduction

The topic of this paper is the metaphysics of partial belief, and in particular the relationship between partial beliefs and preferences. In brief, I want to defend a certain kind of view about what partial beliefs are and how they're measured which takes their connection to preferences to be of special and unique importance. I'll say more

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement No. 703959.

✉ Edward Elliott
e.j.r.elliott@leeds.ac.uk

[1] School of Philosophy, Religion and History of Science, University of Leeds, Leeds LS2 9JT, UK

Published online: 04 July 2019

on what I mean by this as we go along, but the rough idea will be familiar to any readers acquainted with the history of probability theory and Bayesianism. Indeed, it was present already in Ramsey, who famously argued that 'the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it' (1931, p. 169).

Once upon a time, not so long ago, this kind of *preference-centric* approach to understanding partial beliefs was the norm. And in some circles it still is. But in philosophy, nowadays, it gets a bad rap—it's frequently dismissed out-of-hand as outdated and behaviouristic, or as committed to anti-realism or instrumentalism (as if this were automatically a bad thing). Specific instances of the approach have been criticised by, *inter alia*, Christensen (2001), Hájek (2008), Meacham and Weisberg (2011), Eriksson and Rabinowicz (2013), and Stefánsson (2016, 2018). The general approach has been most thoroughly criticised by Eriksson and Hájek (2007), whose preferenceless *Zen monk* and several closely related counterexamples have suggested to many philosophers that any account of partial beliefs and their measurement that ties them too closely to preferences is irredeemably flawed.

I think all this is too quick. There's a lot to be said in favour of the preference-centric approach to partial belief, and too much of the critical discussion so far has focussed on simplistic caricatures of the views that its advocates actually endorse. Consequently, in the remainder of this paper I'll provide an exposition of the preference-centric approach, as well as a defence. I won't try to address all of the criticisms and objections that exist, for they are many and various. But I will explicitly discuss two of them: first, Eriksson and Hájek's Zen monk example; and second, an important issue raised by Christensen (2001) in what I'll call his *other roles* argument. Most other objections that typically get floated against perference-centrism follow similar themes.

After setting down some background assumptions in the next section, in Sect. 3 I'll describe the core tenets of preference-centrism. In Sect. 4 I will consider the Zen monk case in some detail, and I'll show that a suitably generalised version of the problem that the Zen monk poses survives a common first-pass response to Eriksson and Hájek's original case. Then, in Sect. 5 I will argue that the generalised Zen monk problem is of only limited concern to some varieties of preference-centrism. Finally, in Sects. 6 and 7, I will sketch a realist (with a capital 'R') variety of preference-centrism and show how it deals with the Zen monk and the other roles argument.

2 Background assumptions

Let me first set the scene with a few general background assumptions. I won't make any attempt to defend these assumptions here, and I'm not entirely certain that each one of them is true. But I do believe that they are together more probably true than not (by a wide margin), and they'll at least prove useful for structuring the remainder of the discussion.

First, I will assume that agents both real and ideal have partial beliefs. When I imagine what partial beliefs are, I don't picture them as some idealised or fictional attitude that should only be attributed to Bayesian angels contemplating probabilities in some faraway possible world. *You* and *I* have partial beliefs, and I'm inclined to

think that this fact is grounded in something perfectly *objective* about us—presumably, something going on inside our heads, or perhaps something about how the stuff inside our heads is causally linked to our behaviour, sensory inputs, wider community and/or ancestry.

Next, I will assume that every partial belief comes with a *strength*, and that these strengths are *numerically measurable*. That is: there is in principle some assignment of numerical values (broadly construed) to the various strengths of belief, such that the theoretically important relations amongst those strengths are reflected in designated mathematical relations and operations amongst their assigned values. Such an assignment is usually known as a *measurement*.[1] So, for example, some strengths of belief are *greater* than others, while some are *much* greater than others; some are *maximal* and others *minimal*. These (and other) kinds of relations between strengths of belief ought to be appropriately numerically represented in any adequate measurement thereof.

I want to be as neutral as is reasonable about the specifics of how strengths of belief are to be measured. With that said, for expositional purposes it'll be useful to stick to something concrete. Since it's the most familiar way of thinking about these things, I'll therefore assume that strengths of belief are measurable on a ratio scale with real values between 0 and 1. This carries some rather strong implications about the relational structure of the system of strengths of belief. Let me say two things about this. First, I take it as generally uncontroversial that strengths of belief are measurable on more than just an ordinal scale, and that they carry at least interval information (e.g., it must be somehow meaningful to talk of believing one thing with *much* more strength than another), and probably also ratio information (e.g., believing one thing with *twice* the strength as another). And second, nothing about what I have to say will hang on the choice of numbers. Instead of the reals, we might make do for instance with a small few integers, or we could use an expanded space of values including hyperreals or surreals. If we wanted to get really fancy we could use intervals, fuzzy sets, n-dimensional vectors, or what have you.

Finally, I'll assume that the strength of a belief does not belong to the content of that belief. Partial beliefs are not, or need not be, beliefs *about* probabilities. I'm confident that I'll never be a karaoke world champion. The content of this attitude is not *there is a high probability that I will never be a karaoke world champion*, where the sense of 'probability' in question might be cashed out in terms of chances, frequencies, propensities, evidential relations, fair betting prices, etc.—it's *about* nothing more than my limited vocal capacities. It is helpful if we think of the strength of a belief as something that attaches to the attitude itself, and wholly separable from the attitude's content.

[1] My use of 'measurement' and it's cognates throughout this paper will be in reference to the abstract assignment of numbers, not to the empirical process whereby we determine the value of a quantity attaching to some thing. Compare, for example, (i) *measuring temperature* by assigning to each temperature a numerical value in degrees Celsius that appropriately reflects its properties relative to the other temperatures, with (ii) *measuring the temperature* of a liquid using a thermometer. The ambiguity is unfortunate, but long entrenched (cf. Suppes and Zinnes 1963; Krantz et al. 1971). Bunge (1973) once suggested the term 'quantitation' for what I'm calling measurement, which would make the present assumption about the *quantitatability* of partial belief. ('Quantification' was already taken.) Let me emphasise that I will *not* be addressing the empirical question of whether and how strengths of belief might be determined through observations of choice behaviour at any point in this paper.

I do have some partial beliefs that are about probabilities, but most are about plain old non-probabilistic states of affairs. See Christensen (2004, pp. 18ff), Weatherson (2016) and Yalcin (2012) for more discussion on this point.

3 What is preference-centrism?

Summarising the assumptions of Sect. 2, let's say that strength of belief is a *genuine psychological quantity*. If this is correct, then two important questions arise:

JUSTIFICATION. How can we *justify* measuring strengths of belief as we do? That is, on what basis are we allowed to say that the quantity has *this* structure, appropriately measured by *that* assignment of numerical values?

CHARACTERISATION. Under what conditions does an agent have such-and-such partial beliefs, with such-and-such strengths?[2]

Let's say that someone adopts a preference-centric approach to understanding partial belief just in case, in answering these questions, they posit a *uniquely central role* for the relationship between partial beliefs and preferences. I'll explain what I mean by that in more detail momentarily; for now, I take it that paradigm instances of preference-centrism can be found in de Finetti's (1937) *betting interpretation*, including its more recent incarnations that involve lower and/or upper previsions (see esp. Walley 1991). We also find versions of preference-centrism in the betting interpretation's somewhat more general cousins, those accounts of partial belief that relate them in some way to preferences *via* decision-theoretic representation theorems (e.g., Ramsey 1931; Savage 1954; Anscombe and Aumann 1963; Cozic and Hill 2015; see also Fishburn 1967). And at least some varieties of interpretivism and functionalism can be reasonably counted as preference-centric (e.g., Pettit 1991; Maher 1993; Davidson 1980, 1990, 2004; Lewis 1974, 1983a).[3]

So: I've characterised a preference-centric approach as one in which the belief-preference relationship has a *uniquely central role*. That's a little ambiguous, but it's supposed to be. I want to cast a fairly wide net—wide enough at least to catch the varying positions of the authors just cited. There is a common core to all of them,

[2] This question is ambiguous between (amongst other things) a metaphysical reading ('What are the metaphysically necessary and sufficient conditions for believing p with strength x?') and a conceptual reading ('How can we analyse the notion of *believing p with strength x*?'). Depending on your background views, you may or may not think that these questions can come apart in various ways. I'll leave it ambiguous for now, and come back to it in Sect. 5.

[3] The reader may have noticed that I haven't discussed what I mean by 'preference' and what the basic objects of preference are. Broadly speaking, there are two main ways preferences are conceived. First, there's preference as a kind of comparative propositional attitude, the kind we'd have when we'd prefer the facts to be one way rather than another. We find this conception especially in Jeffrey (1965). Second, there's preference *qua* disposition to choose some act (or bet, or commodity bundle, etc.) over another. This conception is most closely associated with Samuelson (1938) and Savage (1954). I think of preference as a propositional attitude with close ties to choice behaviour, such that in most (but not all) cases we can read the former off of the latter, and the latter off of the former. But I want my discussion to be neutral between the two readings, and I won't be precious about keeping them separated. What I have to say shouldn't depend too much on the disambiguation, so feel free to pick whichever 'preference' is your preference.

though, and I've always found the relative simplicity of the betting interpretation especially useful for teasing it out.

Suppose Sally would prefer winning \$1 to the status quo, and that she's uncertain as to whether p. Given that, let β designate a bet with prize \$1 if p is true, and \$0 if p is false—i.e.,

$$\beta = \langle \$1 \text{ if } p,\ \$0 \text{ otherwise} \rangle$$

We should expect that Sally will prefer being given \$1 unconditionally to taking the bet β, and that she'll prefer β to nothing. That much is obvious. The central idea of the betting interpretation, however, is that the *extent* to which Sally prefers \$1 to β (and β to \$0) is directly proportionate to the strength of her belief in p. Combine that simple idea with the assumption that the strength of the preference that Sally has for β can be straightforwardly read off the prices she'd be willing to buy and sell it at, and that's the betting interpretation in a nutshell.

Let $\mathcal{U}$ here be a measure of Sally's preferences on an interval scale (in other words, a utility function). Ramsey showed us how, under appropriate conditions, we might construct a scale like this for the measurement of preferences without first measuring beliefs (Ramsey 1931; see Elliott (2017a, b), for a recent exposition of and improvements to Ramsey's method.) Next, let $\mathcal{P}$ designate an appropriate measure of her beliefs on the [0, 1] interval. Then, suppressing a few fiddly details, according to expected utility theory,

$$\mathcal{U}(\beta) \ = \ \mathcal{P}(p) \cdot \mathcal{U}(\$1) \ + \ \mathcal{P}(\neg p) \cdot \mathcal{U}(\$0)$$

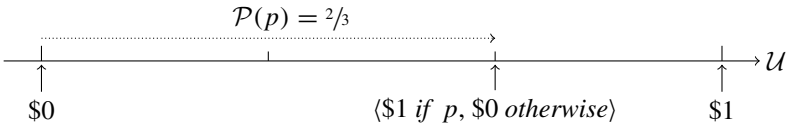
As Ramsey noted, if we assume that $\mathcal{P}$ satisfies at least the condition

$$\mathcal{P}(p) = 1 - \mathcal{P}(\neg p),$$

then we can then re-write the above in terms of ratios of differences in utility to better represent the role that Sally's partial belief regarding p is playing:

$$\mathcal{P}(p) \ = \ \frac{\mathcal{U}(\beta) - \mathcal{U}(\$0)}{\mathcal{U}(\$1) - \mathcal{U}(\$0)}$$

Call this the *betting formula*; it is just another way to say that *where* Sally's utility for β sits between her utilities for \$1 and for \$0 varies directly in proportion to the strength with which she believes p. For example, the utility she attaches to β will sit two thirds of the way from \$0 to \$1 if she believes p to degree $^2/_3$:



Now, you *might* take the betting formula as a complete answer to the CHARACTERISATION question. And some people certainly seem to have done so. However, the

point of this discussion *isn't* that the betting formula is a good definition of the strength of belief—it's not. There are compelling counterexamples, and those fiddly details I dismissed earlier do matter. (See especially Bradley and Leitgeb 2006; Eriksson and Hájek 2007; Eriksson and Rabinowicz 2013). More generally, we probably don't conform perfectly to the simplified version of expected utility theory just described even for simple bets like β—at least not in all cases and at all times. (This is one of the central complaints of Meacham and Weisberg (2011) against versions of preference-centrism that rely on representation theorems.) The betting interpretation comes with a few built-in simplifying assumptions that don't seem too plausible if treated as universal generalisations about how decisions are made.

But what I want to highlight here is that almost every formal decision theory posits a strong correlation between strength of belief and strength of preference for choices under uncertainty that operates along *roughly* the lines I've just described. The various alternatives to expected utility theory that we see today—whether developed for normative or descriptive purposes—all retain the same basic structure of that theory, and they tend to have the betting formula or something in the nearby vicinity as a special case. Indeed, most alternatives to expected utility theory just *are* expected utility theory with some extra bells and whistles: add a fudge factor here, maybe throw in a risk function there, and allow for some non-additive probabilities, and you've got yourself the prototypical non-expected utility theory. See, for example, Luce and Fishburn (1991), Tversky and Kahneman (1992), Sarin and Wakker (1992), Seidenfeld et al. (2010) and Buchak (2013).

According to advocates of the preference-centric approach, that tells us something *centrally important* about our beliefs, and how they're measured. Note, especially, that the betting formula doesn't just suggest a way of answering the CHARACTERISATION question—it also suggests natural answers to questions of JUSTIFICATION. Roughly: strengths of belief can be measured on the [0,1] interval of the reals because the ratios of utility differences that correspond to those beliefs will always sit somewhere on that interval (and where one's partial beliefs are vague and imprecise, so too will their preferences be); strengths of belief are ratio-scalable because ratios of utility differences are meaningful; and they admit interpersonal comparisons because similar strengths of belief play similar roles in determining strength of preference for choices under uncertainty for different agents. We know that we don't yet know *exactly* how strengths of belief and strengths of preference relate in ordinary, non-ideal agents like ourselves. We've probably still got more to learn about how exactly they relate for ideally rational agents. But there can be no reasonable doubt that the betting formula gets something fundamentally *right* about that relationship, and further work along these lines promises to help us make sense of why and how partial beliefs are measurable in the ways we take them to be.

This isn't just a point about how we assign numbers, but an essential aspect to understanding what our partial beliefs *are*. The CHARACTERISATION and JUSTIFICATION questions are closely bound up with one another. Partial beliefs are individuated along two dimensions: *content* and *strength*. We've assumed that these dimensions are independent. Whatever it is that answers the CHARACTERISATION question must at least differentiate between strengths of belief appropriately—and, having specified the conditions under which an agent believes p to degree x for each strength x, it's only

plausible that we'd find *something* in how those conditions relate to one another that will explain why those strengths can be measured in the way they are. Just imagine if that were false: we find that an agent believes p to strength 0.5 just in case condition c is satisfied, and believes p with strength 0.6 just in case condition c' is satisfied, but there's nothing at all in how c and c' relate to one another that might explain why in c' she believes p *more* than she does in c. Not impossible, perhaps, but certainly unlikely.

So it's fair to assume that we cannot answer the CHARACTERISATION question without also providing the basis for answering the JUSTIFICATION question. In the other direction: an answer to the JUSTIFICATION question need not be all there is to answering the CHARACTERISATION question—there might be more to characterising what it is to have partial beliefs than merely explaining why they're measured the way they are. We'll talk about that more below. But, for the same reason we've just seen, whatever answers the JUSTIFICATION question will likely form a *major* and *essential* part of the answering the CHARACTERISATION question.

Given this, the kind of preference-centric approach that I would want to defend is committed to saying that the belief-preference relationship, as captured at least *roughly* by the betting formula,

1. Constitutes *one of* the most central theoretical roles for partial beliefs,
2. Is *necessary* for answering the CHARACTERISATION question, and
3. Is *uniquely important* for answering the JUSTIFICATION question.

I think almost everyone would agree with the first claim. My guess is that many Bayesians would also agree with the second, though I'd also guess that plenty would want to disagree as well. (I did not say 'uniquely necessary', though I didn't say 'non-uniquely' either.) The third claim is likely to the most controversial. Their conjunction constitutes what I take to be the common core of preference-centrism. Further reasons to think this will be raised in Sect. 5, where I discuss the views of some historically influential preference-centric authors.

4 Will the real Zen monk please stand up?

We now have enough background on preference-centrism to talk about Eriksson and Hájek's Zen monk:

> Imagine a Zen Buddhist monk who has [partial beliefs] but no preferences. Gazing peacefully at the scene before him, he believes that Mt. Everest stands at the other side of the valley, that K2 does not, and so on. But don't ask him to bet on these propositions, for he is indifferent among all things. *If the monk is conceptually possible, then any account that conceptually ties degrees of belief to preferences is refuted.* (2007, p. 194, emphasis added)

I think the Zen monk is clearly cause to reject some versions of the preference-centric approach—those according to which partial beliefs are to be reduced to nothing more than the preferences an agent has at a given time. And good riddance. That kind of preference-centrism is often at the forefront of critical discussions, but I'm skeptical

about how many people have ever actually held such a view. (I'll discuss this in Sect. 5.) With that said, the conceptual possibility of the Zen monk is entirely consistent with (in fact: *implied by*) many possible and actual accounts of partial belief that tie them *very* closely to preferences indeed. It's the purpose of the later sections below to establish this.

But before we get to all that, in the remainder of this section I want to get clearer on exactly what the issue *is*, and how fans of preference-centrism *shouldn't* respond. In particular, here's a common first-pass response to the case:

> It's not obvious that the monk really *is* conceptually possible. Is being indifferent among *all* things really conceptually coherent? Surely the monk would prefer, for example, an end to suffering over a swift kick to the shins?

I like to think that a state of utter preferencelessness is conceptually possible. It's at least clear enough that the Zen monk isn't *prima facie* negatively inconceivable, to use David Chalmers' two-fold categorisation scheme (see Chalmers 2002). There are no obvious contradictions implicit in the description of the case. I do not know whether the monk is ideally positively conceivable, but there's no point butting heads over it. We don't need the Zen monk to bring out the worry that's driving Eriksson and Hájek's case.

To see this, let's describe the issue abstractly.[4] Suppose we start with some background decision theory, we'll label it $\mathcal{D}$. If $\mathcal{D}$ looks anything at all like expected utility theory—and a reminder: they basically all do—then $\mathcal{D}$ will say that an agent's overall preferences at a given time are determined by

1. her *partial beliefs* at that time (represented with $\mathcal{P}$),
2. her *basic desires* at that time (represented henceforth with $\mathcal{U}$), and
3. any additional factor(s) (represented with $\mathcal{F}$), e.g., the agent's attitudes to risk

The determinants $\mathcal{P}, \mathcal{U}$, and $\mathcal{F}$ may furthermore be restricted to some special class, the 'admissible' determinants—e.g., $\mathcal{D}$ might build in the presupposition that $\mathcal{P}$ satisfies certain constraints, such as that $\mathcal{P}(p) = 1 - \mathcal{P}(\neg p)$.

(So, for instance, on Savage's (1954) way of modelling decision-making, agents' preferences over their actions are determined by their basic preferences over the ultimate and fully-specified consequences of their choices ($\mathcal{U}$), plus a probabilistic partial belief function ($\mathcal{P}$) defined over the different states of the world consistent with those actions. In almost all alternatives to expected utility theory, we usually find the same kind of structure: a (not necessarily probabilistic) partial belief function $\mathcal{P}$, a function representing preferences over consequences $\mathcal{U}$, and in some cases a third factor $\mathcal{F}$, such as the agent's attitudes towards risk (e.g., Luce and Fishburn 1991; Starmer 2000; Buchak 2013). In Ramsey's (1931) and related systems [e.g., Davidson and Suppes 1956; Davidson et al. 1957, Elliott (2017a, b)], preferences for bets are fixed by agents' utilities for the bets' potential outcomes plus their confidence in the conditions under which those outcomes would be obtained if they were to take up the bet. And in (Jeffrey 1965), agents' preference rankings over arbitrary propositions can be described as a

[4] The level of formal abstraction that follows is almost certainly unnecessary for understanding the Zen monk case, which is very straightforward. But the notation and formalisms introduced here will be helpful later in the paper, so please bear with me.

function of their basic desirabilities for possible worlds plus a probability distribution over the space of those worlds.)

Given this, we could represent the content of decision theory $\mathcal{D}$ as a function from admissible $\langle \mathcal{P}, \mathcal{U}, \mathcal{F} \rangle$ triples to an overall pattern of preferences, which we'll represent hereafter with $\succsim$. Now, here's one more thing that holds true for most varieties of decision theory: if an agent's basic desires are *trivial* (i.e., $\mathcal{U}$ encodes no preferences between any basic objects of desire, whatever they may be taken to be), then her overall preferences will be likewise trivial (i.e., $\succsim$ encodes no preferences whatsoever). An easy and obvious example is expected utility theory. If Sally has no interests in money, such that $\mathcal{U}(\$1) = \mathcal{U}(\$0)$, then the strength of her preference for the bet β (from Sect. 3) will remain constant regardless of the value of $\mathcal{P}(p)$. More generally, if *every* pair of outcomes a and b are the same as far as your basic desires are concerned, then there's no reason to choose any bet over any other bet, and there's never any risks involved in making one choice rather than any other.

So let $\mathcal{U}^{tr}$ designate some *trivial* set of basic desires. (I'll assume without argument that $\mathcal{U}^{tr}$ is admissible, because we'll soon see that it makes no difference either way.) Then, for $\mathcal{P} \neq \mathcal{P}^{\star}$, under most decision theories $\mathcal{D}$,

$$\mathcal{D}(\mathcal{P}, \mathcal{U}^{tr}, \mathcal{F}) \;=\; \mathcal{D}(\mathcal{P}^{\star}, \mathcal{U}^{tr}, \mathcal{F}^{\star})$$

The upshot is that if it's conceptually possible to have two agents—two Zen monks, let's call them *Zee* and *Zed*—with the same trivial basic desires but different partial beliefs, then they will have the same preferences. In that case there can be no way to extract any differences between their beliefs out of the facts about the preferences they both have at that time. In other words, if Zee and Zed are conceptually possible, then the following is false:

ENTAILMENT. The facts about an agent's partial beliefs at a time are in all circumstances conceptually entailed by the facts about what preferences she actually has at that time.

And *that's* the extent of what the Zen monk case can be taken to establish.

So with that in mind, let's suppose that it really is conceptually impossible for two agents to be in the states $\langle \mathcal{P}, \mathcal{U}^{tr}, \mathcal{F} \rangle$ and $\langle \mathcal{P}^{\star}, \mathcal{U}^{tr}, \mathcal{F}^{\star} \rangle$. Indeed, let's suppose more generally that the very idea of having no preferences whatsoever is incoherent. Does this affect anything of critical importance? Not at all: there are other ways to get the same result. This is evident from the many representation theorems for the various decision theories that we've managed to find over the past century. (This will take some work to spell out, so readers not interested in the details should at this point feel free to skip through to the final paragraph of this section.)

A representation theorem for a decision theory $\mathcal{D}$ states that a certain sequence of constraints $C_1, C_2, \ldots, C_n$ on a set of preferences $\succsim$ is sufficient (and perhaps also necessary) to ensure:

$\mathcal{D}$-EXISTENCE. There exists an admissible $\langle \mathcal{P}, \mathcal{U}, \mathcal{F} \rangle$ such that $\mathcal{D}(\mathcal{P}, \mathcal{U}, \mathcal{F}) \;=\; \succsim$.

The constraints $C_1, C_2, \ldots, C_n$ ensure $\mathcal{D}$-EXISTENCE, in other words, only if *any set of preferences that satisfies* $C_1, C_2, \ldots, C_n$ *is consistent with* $\mathcal{D}$. A representation

theorem will also usually be given alongside a uniqueness theorem, where this might come in a variety of forms. The particular form of uniqueness result that will be of interest to us tells us that whenever $\succsim$ satisfies $C_1, C_2, \ldots, C_n$, then

$\mathcal{P}$-UNIQUENESS. If $\mathcal{D}(\mathcal{P}, \mathcal{U}, \mathcal{F}) = \mathcal{D}(\mathcal{P}^\star, \mathcal{U}^\star, \mathcal{F}^\star) = \succsim$, then $\mathcal{P} = \mathcal{P}^\star$.

That is, the constraints $C_1, C_2, \ldots, C_n$ will ensure $\mathcal{P}$-UNIQUENESS just in case, *if some set of preferences satisfies* $C_1, C_2, \ldots, C_n$, *then those preferences could only have been generated consistently with* $\mathcal{D}$ *by exactly one set of partial beliefs*.

If $\mathcal{P}$-UNIQUENESS holds, and it's assumed that beliefs and preferences are in fact related to one another in the way that $\mathcal{D}$ describes, then—*and only then*—we will it be possible to determine what partial beliefs an agent has at a time merely be considering what preferences she has at that time. To be clear: there's no guarantee that there *will* be constraints on $\succsim$ consistent with $\mathcal{D}$ that ensure $\mathcal{P}$-UNIQUENESS. Such constraints exist for *some* decision theories, but not for all, and finding the relevant constraints can be hard work. More importantly, even where there *are* constraints sufficient to ensure $\mathcal{P}$-UNIQUENESS, there will in general be a significant gap between those constraints sufficient for $\mathcal{D}$-EXISTENCE and those constraints sufficient for $\mathcal{P}$-UNIQUENESS. This is usually easy to prove, and it means that we should in general expect there to be at least *some* sets of preferences that are consistent (according to $\mathcal{D}$) with multiple distinct sets of beliefs.

So, back to the Zen monk. Included within either the constraints that are used to ensure $\mathcal{D}$-EXISTENCE *or* those that are used to ensure $\mathcal{P}$-UNIQUENESS, we typically find the following condition:

NON- TRIVIALITY. It's not the case that $\succsim$ is trivial.[5]

And here's the important part: as a rule, the constraints that ensure $\mathcal{D}$-EXISTENCE *plus* NON- TRIVIALITY don't by themselves ensure $\mathcal{P}$-UNIQUENESS. Or to put that more directly: *even with the Zen monk ruled out by fiat*, we still find that on any of the usual decision theories there will be sets of preferences consistent with multiple distinct sets of beliefs. We simply do not need the Zen monk to raise Eriksson and Hájek's objection to ENTAILMENT.

The argument doesn't come with a nice pithy case to exercise your imagination, and the instances where $\mathcal{P}$-UNIQUENESS fails but NON- TRIVIALITY holds are usually going to be complicated and boring. But if you're not yet convinced, then one thing in particular is worth pointing out: the constraints over and above those required for $\mathcal{D}$-EXISTENCE under which $\mathcal{P}$-UNIQUENESS also holds true, *minus* NON- TRIVIALITY, are universally regarded as the *least* plausible constraints of any representation theorem (both descriptively and normatively). These include things like Savage's widely criticised requirement that the set of 'acts' includes all functions from states to outcomes, or his constraint P6 which (in effect) says that an agents' relative likelihood

[5] In Savage's (1954) theorem, NON- TRIVIALITY is his condition P5, which is necessary for his strong uniqueness theorem. In Ramsey's (1931) theorem, NON- TRIVIALITY is a consequence of his first constraint, 'There is an ethically neutral proposition p believed to degree ½'. And in Jeffrey's (1965) theorem, which famously does not mention constraints sufficient to entail $\mathcal{P}$-UNIQUENESS, we still find NON- TRIVIALITY in the so-called *G Condition*: 'In the preference ranking there is a good proposition, *G*, of which the denial is bad'. This is just a small selection of examples—NON- TRIVIALITY shows up in some form or another everywhere.

rankings must be 'atomless' and defined over an uncountable set of states. (See Joyce 1999, Sect. 3.3, and Fisburn 1970, pp. 193ff, for discussion.) More generally, one typically only manages to prove a result like $\mathcal{P}$-UNIQUENESS by making very strong richness assumptions about the domain of the preference relation, and then combining that with strong structural constraints on preferences that are difficult to justify either empirically or normatively. As James Joyce put it,

> … when one looks closely at the way these theories obtain unique representations what one finds is mostly smoke and mirrors… unique representations are secured only by making highly implausible assumptions about the complexity of the set of prospects over which the agent's preferences are defined. (2000, p. S7)

To summarize: even if you think that Zen monks are conceptually impossible, if your favourite decision theory $\mathcal{D}$ looks more or less like expected utility theory, then it's likely that there will still be possible sets of preferences which, according to $\mathcal{D}$, are consistent with more than one set of partial beliefs. Call this the *non-denominational monk problem*. Debating the possibility of the Zen monk seems parochial when there are non-denominational monks that need to be dealt with.

5 Who's afraid of the big bad monk?

I'm inclined to think that ENTAILMENT is false. There is, at least in some cases, more to understanding what it is for an agent to have partial beliefs than can be captured merely by talking about what preferences that agent actually has at a given time. And if that's all that Eriksson and Hájek meant when they said that "any account that conceptually ties credences to preferences is refuted," then I'm happy to agree with their conclusion.

As I mentioned earlier, though, I'm doubtful that many advocates of preference-centrism—including the historically paradigmatic advocates—ever believed otherwise. So, before I go on to present my own favourite way of answering the non-denominational monk problem in Sect. 6, in this section I want to consider the extent to which ENTAILMENT (or something close to it) can be thought of as a 'core commitment' of the preference-centric approach.

Let's get some general points out of the way first.[6] First: there are many ways to read the CHARACTERISATION question, and we certainly don't have to think that every attempt to answer it must be aimed at faithfully recapturing our *conceptual* commitments—to supply an *analysis* of the concept of partial belief. That's one kind of project that a fan of preference-centrism might be engaged in. Another possible project is *explication*: to isolate an especially clear notion in the vicinity of the ordinary concept that will be useful for this or that theoretical purpose. At least some preference-centric authors have had something like this in mind, some more explicitly than others (e.g., Walley 1991 and arguably Savage 1954). Relatedly, in some cases a preference-centric account is intended as nothing more than an *operationalisation*: not an explication or even a definition, but just enough to pin down a clear enough meaning for present purposes (where those purposes don't call for a particularly sophisticated or plausible

[6] See also Christensen (2001, p. 360) for similar points.

theory of the mind). This seems to be the case for many uses of the betting interpretation in economics. Finally, another answer to the CHARACTERISATION question can come in the form of an *a posteriori identification*: to characterise the metaphysically necessary and sufficient conditions under which one has such-and-such partial beliefs, where those conditions might come apart from (and indeed conflict with) any commitments built into the concept of partial belief.[7] Like Eriksson and Hájek, I'm particularly interested in *analyses*, so let's pretend from here on that everyone else is as well.

Next, we need to be careful when reading what look like endorsements of the ENTAILMENT thesis, or nearby theses. Consider Peter Walley, who writes:

> According to the psychological model outlined above [and advocated here], beliefs and values are *behavioural dispositions*: abstract, theoretical states of intentional systems, which can interact in suitable circumstances to produce actions… You have a higher degree of belief in one event A than another B when You are disposed to choose a bet which yields a desirable reward if A occurs, rather than a bet which yields the same reward if B occurs. (1991, p. 18)

This reads like an *equality*: partial beliefs *are* preferences (specifically: they *are* choice dispositions). Over the following pages, however, we find:

> Logical behaviourists, notably Ryle, identified mental states such as beliefs and values with certain kinds of behavioural dispositions … Logical behaviourism is consistent with the psychological model [advocated here] and with the behavioural interpretation of probability adopted in this book, although it goes somewhat further than we need to. We require only that beliefs and values entail certain behavioural dispositions, *there may be more to them than that* […] We are requiring only that beliefs and probabilities should (potentially) influence behaviour. That does seem to be an essential part of their meaning. (pp. 19–20; emphasis added)

Similarly, if we ignore the 'roughly' in the following passage, then it's easy to read Ramsey as *equating* having partial beliefs with having a certain pattern of preferences when he says:

> … This amounts roughly to defining the degree of belief in p by the odds at which the subject would bet on p, the bet being conducted in terms of differences of value as defined. (1931, pp. 179–80)

But to interpret Ramsey in this way is to neglect the core idea which motivated his theory—that the degree of a belief is a *causal* property of it. A belief is something separate from and causally prior to preferences, that comes with a strength that explains the extent to which that belief feeds into our preferences when we're rational. Ramsey

[7] My favourite example of this: it's plausibly a priori that water is the watery stuff, where 'watery stuff' picks out a disjunction of the kinds of things we associate with the concept *water* (e.g., being the clear, colourless, potable liquid that fills the lakes and oceans around here and falls from the sky as rain). But it's not metaphysically necessary that water is watery, since water is H_2O, and there are metaphysically possible worlds where H_2O is not even remotely watery (e.g., black and tarry). See Braddon-Mitchell (2003) for some discussion of this example. The point is familiar from the literature on two-dimensional semantics, but need not presuppose it—only that conceptual impossibility need not imply metaphysical impossibility.

uses this kind of causal language throughout his paper (see esp. pp. 169–75); and in an earlier passage, he writes:

> I suggest that we introduce as a law of psychology that his behaviour is governed by what is called the mathematical expectation; that is to say that, if p is a proposition about which he is doubtful, any goods or bads for whose realization that p is in his view a necessary and sufficient condition enter into his calculations multiplied by the same fraction, which is called the 'degree of his belief in p'. (p. 174)

Note the emphasis here: given that the subject *has* a belief towards p, what makes it the case that he believes it *with a certain strength*? On a charitable reading, Ramsey only ever purported to show that the *strengths* with which an agent believed certain propositions could be determined through their preferences given the assumption that those preferences were formed in accord with a 'general psychological theory' (p. 173), one that relates partial beliefs and basic desires to choices in specifically causal terms and applies only under 'suitable circumstances' (pp. 170, 172, 173). The beliefs themselves are causal antecedents to preferences, not themselves reducible to preferences—there might (for all he says) be more to what it is to have a belief than can be stated purely in terms of preferences. But what it is to have a belief *with a given strength* can be determined (in suitable circumstances) by reference to what that belief *does* in the context of rational decision making under uncertainty. *That's* the central thesis of his discussion, *not* that partial beliefs are nothing over and above preferences.

Note, also, that Ramsey certainly never thought that satisfying the axioms of his representation theorem was a *necessary* condition for having partial beliefs. He argued that anyone who satisfied those axioms would have probabilistically coherent beliefs, yet he also clearly allowed that it was possible to have incoherent beliefs. (His Dutch Book argument on p. 182 makes this abundantly clear.) Unfortunately for us, Ramsey never said anything very specific about what an agent's beliefs would be like or how they might be measured if her preferences violated his axioms. So, in print, Ramsey *at most* committed to the claim that the facts about an agent's actual preferences determine the $\mathcal{P}$-facts *under the assumption that those preferences satisfy the conditions of his representation theorem*. I'm not convinced he would have been quite happy with saying even this much, but even if he were then it could at most establish:

> LIMITED ENTAILMENT. The facts about an agent's partial beliefs at a time are *in some cases* conceptually entailed by the facts about what preferences she actually has at that time.

And here's the rub: a set of preferences $\succsim$ will satisfy all eight of his theorem's conditions only if $\mathcal{P}$-UNIQUENESS holds. 'Truth and Probability' provides us with an unfinished sketch of an account of the measurement of partial belief that was only ever intended to apply in exactly those cases where the Zen monk and non-denominational monk problems do not and cannot arise!

The same general point here applies Savage, *mutatis mutandis*. Savage was perfectly clear that his preference axioms only characterised 'a highly idealized theory of the behaviour of a "rational" person with respect to decisions' (1954, p. 7), while also recognizing that we non-idealized and potentially irrational mortals can violate those

axioms and still have partial beliefs (see esp. pp. 19–20, pp. 27ff). So Savage, too, gave us a theory of measurement that was only supposed to apply in exactly those cases where the non-denominational monk problem cannot arise.

This means that Ramsey's and Savage's accounts were incomplete, of course, and it would obviously be nice if we had something more general to say. That's hardly a criticism for the first early steps in formulating a view. Moreover, it's not like there are *no* other considerations to which Ramsey or Savage could have appealed to in order to help distinguish between agents whose preferences do not provide enough information to decide their partial beliefs.

More recent advocates of preference-centrism *do* tend to make use of a strictly wider range of resources, at least some of which could be put to use in dealing with the non-denominational monk problem. Consider first Patrick Maher's oft-cited endorsement of interpretivism:

> I suggest that we understand attributions of [partial belief] and utility as essentially a device for interpreting a person's preferences. On this view, an attribution of probabilities and utilities is correct just in case it is part of an overall interpretation of the person's preferences that makes sufficiently good sense of them and better sense than any competing interpretation does. This is not the place to attempt to specify all the criteria that go into evaluating interpretations… For present purposes it will suffice to assert that… having preferences that all maximise expected utility relative to $\mathcal{P}$ and $\mathcal{U}$ is a sufficient (but not necessary) condition for $\mathcal{P}$ and $\mathcal{U}$ to be one's [partial belief] and utility functions. (Maher 1993, p. 9)

Maher explicitly accepts Limited Entailment here. But the representation theorem he employs to flesh out the details of his interpretive theory explicitly includes a Non- Triviality constraint (see pp. 187–8, Axiom 2) and comes with a strong $\mathcal{P}$-Uniqueness result. Furthermore, the other interpretive criteria aren't fully discussed. So it's hard to know what he'd say about the best way to interpret members of the non-denominational monkhood. Here's two (non-exhaustive) things he might say when confronted with the Zen monk:

1. Since the monk's preferences maximise expected utility with respect to both $\langle \mathcal{P}, \mathcal{U}^{tr} \rangle$ and $\langle \mathcal{P}^\star, \mathcal{U}^{tr} \rangle$, there's no fact of the matter as to which of $\mathcal{P}$ or $\mathcal{P}^\star$ represents the monk's partial beliefs.
2. Since the monk's preferences maximise expected utility with respect to both $\langle \mathcal{P}, \mathcal{U}^{tr} \rangle$ and $\langle \mathcal{P}^\star, \mathcal{U}^{tr} \rangle$, other interpretive criteria might be called on to determine whether $\mathcal{P}$ or $\mathcal{P}^\star$ better represents the monk's partial beliefs.

I don't know which response Maher would prefer, if either. But the more interesting and plausible of the two, I think, is the second. If we have other criteria to play with—criteria that might help us decide between interpretations that look equally good when considered only in terms of how well they fit with preferences—then why wouldn't we make use of them?

And this response would be in close alignment with a long history of interpretivist thought. Take, for example, the kind of interpretivism endorsed by David Lewis in 'Radical Interpretation' (1974), from which Maher draws inspiration. According to

the Lewisean position, the correct assignment of partial beliefs and utilities to an agent is decided primarily on the basis of two interpretive principles: *Rationalisation* and *Charity*. As he describes it in 'Radical Interpretation', Rationalisation is close to what Maher says in the passage above: an assignment of partial beliefs and utilities to an agent is better to the extent that makes her behaviour seem pragmatically rational—i.e., to the extent that her behavioural preferences maximise expected utility with respect to those beliefs and utilities.[8] Charity, on the other hand, in summary says that (a) an assignment of partial beliefs is better to the extent that it maximises the agent's epistemic rationality given her life history of evidence and any reasonable constraints on her prior probabilities; and (b) attributes basic desires that make sense given the kind of being she is and the kind of life she has lived.

Lewis wasn't as clear in 'Radical Interpretation' as he could have been about how Rationalisation and Charity were supposed to interact, whether one was to be given priority or who they were supposed to be weighed against each other in cases where they might pull in different directions. The most he says is that once we're given appropriate information on the agent's behaviour and her life history of evidence, we're to "fill in [an assignment of partial beliefs and utilities] by means of the Rationalisation Principle and the Principle of Charity" (1974, p. 341). Nevertheless, two of Lewis' later discussions are helpful in teasing out the full picture and the special role that preferences play within it. Since several of the issues here will be relevant again in the next section, this will be worth spending a little time on.

First, in 'New Work for a Theory of Universals', Lewis gives his account of interpretation a slightly more detailed gloss: an assignment of $\mathcal{P}$ and $\mathcal{U}$ shouldn't just rationalise the agent's *actual* preferences—her preferences at the time of interpretation—but also the agent's dispositions to change her preferences in various ways upon receiving different kinds of evidence. Let '$\mathcal{P}_p$' designate $\mathcal{P}$ conditionalised on the proposition p; then,

> Say that $\mathcal{P}$ and $\mathcal{U}$ *rationalise* behaviour B after evidence p iff the system of [preferences] given by the $\mathcal{P}_p$ expectations of $\mathcal{U}$ ranks B at least as high as any alternative behaviour. Say that $\mathcal{P}$ and $\mathcal{U}$ *fit* iff, for any evidence-specifying p, [learning] p yields a state that would cause behaviour rationalised by $\langle \mathcal{P}_p, \mathcal{U} \rangle$. That is our only constraining principle of fit. (1983a, p. 374, notation altered for consistency)

Thus, if the agent would change her preferences from $\succsim$ to $\succsim^\star$ if she were to learn p, then $\langle \mathcal{P}, \mathcal{U} \rangle$ is a better interpretation not only to the extent that her actual preferences, $\succsim$, maximise expected utility with respect to $\langle \mathcal{P}, \mathcal{U} \rangle$, but also to the extent that the preferences she would have, $\succsim^\star$, maximise expected utility with respect to $\langle \mathcal{P}_p, \mathcal{U} \rangle$. Call this principle *Counterfactual Rationalisation*. A little further on, Lewis writes:

> If we rely on [Counterfactual Rationalisation] to do the whole job, we can expect radical indeterminacy of interpretation. We need further constraints, of the sort

[8] A little more precisely, Lewis only ever said that our systematised folk theory of rational decision should "look a lot like Bayesian decision theory" (1979, pp. 533–534; cf. 1974, pp. 337–338). A more complete statement of Rationalisation would involve fewer idealisations (1983a, p. 375), and would allow us to factor in situations in which the folk recognise that decision-makers might predictably choose irrationally (e.g., when they're intoxicated, tired, or stressed).

> called principles of (sophisticated) charity, or of 'humanity'. Such principles call for interpretations according to which the subject has attitudes that we would deem reasonable for one who has lived the life that he has lived. These principles select among conflicting interpretations that equally well conform to [Counterfactual Rationalisation]. (1983a, p. 375)

In other words, the process of interpreting an agent has two stages. First and foremost, we solve for fit with the agent's (actual and counterfactual) preferences using Counterfactual Rationalisation. The remaining principles of Charity then play a secondary, tie-breaking role, filtering between those interpretations reckoned equally good by Counterfactual Rationalisation. Call this the *individualistic* version of Lewisean interpretivism.

Second, in the postscripts to 'Radical Interpretation' (1983b), Lewis notes that his initial discussions made his views seem unduly individualistic. On his later view, those principles aren't supposed to be used to interpret agents on a one-by-one basis. Rather, assignments of beliefs and desires are to be associated with psychophysical states which can be independently identified across different members of the same kind; the correct interpretation of any individual agent is then the best interpretation of their psychophysical states according to the causal roles that those states play in the typical or average member of the agent's kind. For some individuals—so-called 'madmen'—the correct interpretation need not be the one that best rationalises their preferences at all: "[t]he best interpretation for Karl is the one assigned to him by the scheme of interpretation that does best overall, even if it does not do so well in his exceptional case" (1983b, p. 120; cf. also 1980). At the level of individuals, the belief-preference connection might be quite loose indeed—as long as it not quite so loose *on average*. Thus we also have the *anti-individualistic* version of Lewis' view.

So here's the general pattern we're seeing in the authors discussed above: preferences have a special role in our account of what partial beliefs are and how they're measured, or what makes it correct to attribute such-and-such partial beliefs with such-and-such strengths to a person, but there may be more to partial beliefs than just their connection to (actual) preferences. Sometimes, *perhaps*, the facts about an individual's preferences at a time might be enough to pin down the facts about her partial beliefs at that time, though if so then only in special cases where those preferences are non-trivial and satisfy certain other (very strong) assumptions. This need not be true in general. *This* is what advocates of preference-centrism typically believe, and have believed for a long while—not that partial beliefs just *are* preferences, not that the partial beliefs facts are directly entailed by the facts about actual preferences, and not that partial beliefs can be defined entirely in terms of preferences. There are no doubt many who take the hard line and really believe that partial beliefs are conceptually nothing over and above preferences. But in general, ENTAILMENT and LIMITED ENTAILMENT aren't core commitments of preference-centrism.

6 Preference-centric functionalism and counterfactual uniqueness

There are many ways to develop a theory of partial belief consistent with what I've called the 'core commitments' of preference-centrism (Sect. 3). Some might be considered anti-realist or instrumentalist. I'm not always sure I know what those terms mean when they're discussed in this context. In any case, though, I suspect that any plausible answer to the CHARACTERISATION question ought to be broadly functionalist in character—and so will make central reference to the various causal, explanatory, and/or normative roles that partial beliefs play in the theories in which they figure (*à la* Lewis 1970, 1972).

In these final two sections of the paper, I want to sketch how I think a version of preference-centrism should be developed along these lines. In this section I'll discuss show how a simplified version of the view can deal with the non-denominational monk problem, and in the next section I'll flesh the view out a bit more and discuss a final objection to preference-centrism. My view draws on some of the insights already present in Lewis (as any good functionalist theory should), but it differs on the details.

The key to solving the non-denominational monk problem, I think, is to consider not only those preferences that having a given set of partial beliefs will generate directly, but also the preferences that those beliefs might generate under different conditions. We know from the existence of non-denominational monks that there will *at least sometimes* be sets of preferences $\succsim$ that could have been generated by distinct sets of partial beliefs, $\mathcal{P}$ and $\mathcal{P}^\star$. This doesn't mean that $\mathcal{P}$ and $\mathcal{P}^\star$ must therefore have exactly the same overall causal profile—just that, if they *do* have distinctive causal properties, then those won't be manifest in the preferences that they generate *as a matter of fact*. If we want to have a more complete picture of what $\mathcal{P}$ and $\mathcal{P}^\star$ *can* do, sometimes we need to consider what they *would* do in non-actual circumstances.

Lewis' Counterfactual Rationalisation principle incorporates something like this insight, but I think we can do better. There will be hypothetical cases where a given set of preference is consistent (according to expected utility theory) with distinct sets of partial beliefs $\mathcal{P}$ and $\mathcal{P}^\star$, and *furthermore*, for any evidence p the agent could learn, $\mathcal{D}(\mathcal{P}_p, \mathcal{U}) = \mathcal{D}(\mathcal{P}^\star_p, \mathcal{U})$. After all, consider again the two Zen monks, Zee and Zed. Since they have only trivial basic desires, Zee and Zed should both be preferenceless. Moreover, they'll *remain* preferenceless regardless of whatever information they might learn: for any p, $\mathcal{P}_p$ gives rise to the same preferences as $\mathcal{P}$ when combined with a set of trivial basic desires. So long as their basic desires remain unchanged, there's no difference between their actual preferences and the preferences they would have if they were to update their beliefs in light of p.

This suggests a variant on the non-denominational monk problem for the individualistic Lewisean interpretivist, because it's not clear that appeal to the principles of Charity will always be enough. (An essentially similar problem can be raised against the anti-individualistic Lewisean, though because that raises a number of complications I will set it to one side.) We can suppose that Zee and Zed are both human, they belong to the same community, they've had essentially the same basic life experiences and the same history of evidence. Consequently, there's nothing for the principles of Charity to operate on that would let us distinguish between Zee and Zed. Yet they have (perhaps only slightly) different partial beliefs. That seems conceptually possible to

me, and I'd like an account of what their partial beliefs are that makes adequate sense of that possibility.[9]

Note, by the way, that we'll come to the same conclusion regardless of how Counterfactual Rationalisation and Charity are supposed to interact—that is, whether Charity merely plays a secondary tie-breaking role, or whether the degree of fit with each principle is weighted against the other somehow to produce a final interpretation. Zee and Zed satisfy Counterfactual Rationalisation in exactly the same trivialised way, and they're identical as far as Charity is concerned. So it looks like we need another way to think about the causal role of our partial beliefs. We need to be able to ask: *what would Zee and Zed's preferences be like if we held their partial beliefs fixed, but altered their basic desires?*

A metaphor will be helpful for establishing the gist of what I have in mind. Suppose that Robo is a robot, whose internal design is presently unknown to us. Our task is to interpret Robo, to assign to him a set of partial beliefs and basic desires that best fits with whatever we're able to figure out about him, and (hopefully) to also pin down what those attitudes correspond to internally. We're not given very much information to start with, though we do get some helpful pointers.

We're told that Robo was built and is functioning exactly as intended. He is the perfect exemplar of his kind. Moreover, we're told something about what this means in terms of the high-level features of his design. His designers tell us that they wanted him to follow their favourite decision theory, $\mathcal{D}$, which (we'll suppose) looks more or less like expected utility theory. Consequently, somewhere inside Robo we can expect there will be some representation (or representations) of the way Robo takes the world to be, and some representation (or representations) of the basic ways Robo would like it to be, and we can expect that these will interact somehow in a manner that's appropriately modelled at some level by $\mathcal{D}$ so as to determine his choices and behaviour. The designers also wanted to go for maximum flexibility: for any $\mathcal{P}$ and $\mathcal{U}$ that are admissible according to $\mathcal{D}$, there should in principle be a way for Robo to represent $\mathcal{P}$ and $\mathcal{U}$.

We have a lot of time on our hands and plenty of technical know-how. Robo is made almost exclusively from semi-transparent materials, so without interfering with his behaviour we're able to observe everything going on inside. After a while, we make some interesting discoveries.

First, we can see that inside what looks like Robo's head are two little compartments, one labelled '*bel*' and the other labelled '*des*'. The way these compartments work is mystery for now, but we do note that each compartment can be set in any number of different *states*, where those states can be easily identified and distinguished from one another. Second, we learn that in *normal* circumstances—i.e., the kinds of conditions that Robo was designed to operate in, where we'd typically expect to find him or robots like him—Robo's behaviour and behavioural dispositions are a function of the combined state of his *bel* and *des* compartments.

[9] A referee notes that if, for any sequence of evidence, there's only one rational response, then if Zee and Zed have different partial beliefs at least one of them must be irrational. So be it—I never said that Zee and Zed had to both be perfectly rational. Moreover, it'd be nice to have a theory of what partial beliefs are the adequacy of which doesn't hang on the contested issue of whether more than one response to evidence is rationally permissible.

Unfortunately, the causal relationship here is highly complicated: more often than we'd like there are quite distinct combined states that produce exactly the same behavioural dispositions. There is simply no one-one relationship between Robo's preferences and his combined *bel*/*des* state; still less is there any direct connection between his behavioural dispositions and the state of either compartment by itself. This makes our task tricky. But it doesn't make it impossible.

Eventually, we learn enough about Robo's internal set-up to know what would happen if we held fixed the state of either one of the two compartments while varying the other. So here's what we do: for each state S_{bel} of his *bel* compartment, we map out all of the ways Robo's preferences *would be* given each variation to his *des* compartment. We do the same, *mutatis mutandis*, for each state S_{des}. The result is a complete map of the many different sets of preferences $\succsim$ that each combined $\langle S_{bel}, S_{des}\rangle$ state would generate under normal circumstances.

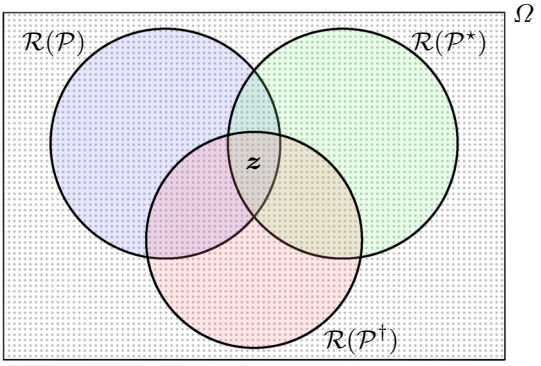
This gives us the basic information we need to start testing out some interpretive hypotheses. We begin with possible assignments of whole sets of partial beliefs $\mathcal{P}$ to the states S_{bel}. Since Robo was designed to instantiate $\mathcal{D}$, and he's functioning exactly as he's supposed to and in normal circumstances, his preferences under any combined *bel*/*des* state will be consistent with $\mathcal{D}$. Therefore, let Ω pick out the space of all sets of preferences consistent with $\mathcal{D}$. We are able to associate each state S_{bel} with a certain region of Ω, designated $\mathcal{R}(S_{bel})$, which contains all and only those sets of preferences that result from holding S_{bel} fixed while varying the state of the *des* compartment. We are *also* able to associate each admissible set of partial beliefs $\mathcal{P}$ with a region of Ω, designated $\mathcal{R}(\mathcal{P})$, that *fits* with those beliefs according to $\mathcal{D}$—or more formally,

$$\mathcal{R}(\mathcal{P}) = \{\succsim \in \Omega : \exists \mathcal{U} \text{ such that } \mathcal{D}(\mathcal{P}, \mathcal{U}) = \succsim\}$$

Finally (and this is the really important bit) we're able to prove that $\mathcal{D}$ implies:

COUNTERFACTUAL $\mathcal{P}$- UNIQUENESS. If $\mathcal{P} \neq \mathcal{P}^{\star}$, then $\mathcal{R}(\mathcal{P}) \neq \mathcal{R}(\mathcal{P}^{\star})$.

Pictorially, we could represent the situation like this, where every point in the box labelled 'Ω' represents a set of preferences consistent with $\mathcal{D}$:



We find a non-denominational monk case whenever the regions of Ω associated with (at least) two different sets of beliefs intersect. That just means that, for any set

of preferences within the intersection, those preferences are consistent with more than one set of partial beliefs (according to $\mathcal{D}$). The Zen monk's trivial preferences will belong to the intersection of every $\mathcal{R}(\mathcal{P})$ for each admissible $\mathcal{P}$ (here labelled z).

So, suppose Robo's *actual* preferences are somewhere in the intersection of $\mathcal{R}(\mathcal{P})$ and $\mathcal{R}(\mathcal{P}^\star)$. If you want to, you can even suppose that he has the trivial set of preferences z. This is just the sort of case where $\mathcal{P}$-UNIQUENESS fails, so if you've got nothing more than Robo's actual preferences to go on you won't have enough information to fix an interpretation of the current state of his *bel* compartment. *But that's no problem.* If we allow ourselves counterfactual information, then we don't *need* $\mathcal{P}$-UNIQUENESS to fix an interpretation. $\mathcal{R}(\mathcal{P})$ and $\mathcal{R}(\mathcal{P}^\star)$ are very different subsets of Ω, and $\mathcal{R}(S_{bel})$ can correspond to at most one of them.

More generally: if we've established COUNTERFACTUAL $\mathcal{P}$-UNIQUENESS, then if *any* interpretation fits the counterfactual properties of S_{bel}, it must fit it *uniquely*. And that's where we get lucky. We find, first of all, that for each S_{bel} there is some $\mathcal{P}$ such that $\mathcal{R}(S_{bel}) = \mathcal{R}(\mathcal{P})$; and second, for each $\mathcal{P}$, there's one or more S_{bel} such that $\mathcal{R}(\mathcal{P}) = \mathcal{R}(S_{bel})$. In other words, we find that each of the different states S_{bel} has a unique interpretation in terms of partial beliefs, by virtue of the fact that their *counterfactual* profile fits the causal profile of the relevant partial belief states as specified by $\mathcal{D}$. (Given what we were told about Robo's design, we knew that states like this had to exist. The lucky part was in finding just what those states actually are.)

Of course, we're not finished with our interpretation of Robo just yet. We still have to make sure that every state S_{des} can be assigned a unique set of basic desires. We can do that using a process exactly analogous to what we've just used for partial beliefs, which would require establishing a COUNTERFACTUAL $\mathcal{U}$-UNIQUENESS condition as well. Furthermore, we need to show that there's no way to rejig the overall interpretation such that the S_{bel} and S_{des} states pick out basic desires and partial beliefs respectively. (We can expect this part to be quite easy.) But at this stage it should be clear enough how the rest of the story goes, and what more is needed to fix a unique interpretation for each state S_{bel} and S_{des}.

The metaphor is only meant to give the gist of how things might go. As a representation of how functionalists ought to approach the assignment of beliefs and desires to actual human beings, Robo is deficient in a number of important respects. We'll talk about those in the next section. But first, let me say a bit about COUNTERFACTUAL $\mathcal{P}$-UNIQUENESS, which is the linchpin of the whole idea.

Why should we expect COUNTERFACTUAL $\mathcal{P}$-UNIQUENESS to be true? A formal proof would depend on the exact nature of $\mathcal{D}$, which I've intentionally left vague. I don't know the final shape of the decision theory that I'd want to use to delineate the causal-functional profile of our partial beliefs in relation to preference, and I certainly don't want to stipulate anything here. But here's a reason why you might want to believe it. Suppose $\mathcal{D}$ looks a lot like expected utility theory, and consider two partial belief functions, $\mathcal{P}$ and $\mathcal{P}^\star$. (For simplicity we'll assume they're defined on the same domain, but the point can be made just as easily without this assumption.) $\mathcal{P}$ and $\mathcal{P}^\star$ will assign distinct values to some proposition p, so now combine $\mathcal{P}$ and $\mathcal{P}^\star$ with some utility function, $\mathcal{U}$, such that

$$\mathcal{U}(a) = 1, \quad \mathcal{U}(b) = 0, \quad \mathcal{U}(c) = \mathcal{P}(p)$$

Now given $\mathcal{P}$, we'd expect indifference between c and the bet $\langle a \textit{ if } p, b \textit{ otherwise}\rangle$—not so given $\mathcal{P}^\star$. Moreover, there won't be any *other* utility function $\mathcal{U}^\star$ which is consistent with the same overall pattern of preferences and such that $\mathcal{D}(\mathcal{P}, \mathcal{U}) = \mathcal{D}(\mathcal{P}^\star, \mathcal{U}^\star)$. Consequently, $\mathcal{R}(\mathcal{P}) \neq \mathcal{R}(\mathcal{P}^\star)$.

The same sort of reasoning will apply more generally. Unless the 'admissible' $\mathcal{U}$ are severely restricted for some reason, or the decision theory $\mathcal{D}$ is quite unlike expected utility theory, if you give me two distinct partial belief functions then I'll give you a context in which they generate different patterns of overall preferences. So it's plausible that we can have COUNTERFACTUAL $\mathcal{P}$-UNIQUENESS. We can alter the case to provide reasons to believe COUNTERFACTUAL $\mathcal{U}$-UNIQUENESS.[10]

Taking into account the counterfactual properties of partial beliefs in relation to preferences gives us a richer way of representing that connection. And what's especially interesting about this is that it shows that there *is*, at least potentially, enough information in the facts about actual *and* counterfactual preferences *alone* to distinguish between Zee and Zed's partial beliefs, *before* we consider any 'other interpretive criteria'. If we wanted to, we *could* define up a simplistic variety of preference-centric functionalism that makes no mention of other functional roles, just the connection that partial beliefs share with actual and counterfactual preferences. I don't think that's the right way to go, but the mere possibility of that view puts the lie to Eriksson and Hájek's claim that "if the Zen monk is conceptually possible, then any account that conceptually ties degrees of belief to preferences is refuted". If we're restricted to considering only actual preferences, then that may well be true. But the preference-centric functionalist need not be restricted to actual preferences. The kind of preference-centric functionalism I'm describing here would have us characterise partial beliefs *entirely* in terms of their connections to preferences, yet it is straightforwardly consistent with (and in fact *implies*) the possibility of Zen monks.

7 Preference-functionalism and other theoretical roles

There are three major respects in which the Robo metaphor is deficient, and these highlight ways in which a more complete preference-centric functionalism would need to be fleshed out. I'll discuss these in increasing order of importance.

First of all, we were able to find states that precisely fit the causal role associated with partial beliefs in Robo's head because it was effectively stipulated in the hypothetical set up that they were there. That is, we imagined that Robo was designed to instantiate $\mathcal{D}$ precisely, that he was functioning exactly as he was supposed to, and that he was operating in normal circumstances. In that sense Robo represents the best possible

[10] If COUNTERFACTUAL $\mathcal{P}$-UNIQUENESS does end up being false, then it will be possible to re-describe our representation of partial belief's causal roles. For example, for each S_{bel}, we define S^p_{bel} as the state Robo's *bel* compartment would be in if he were to receive evidence p while in state S_{bel}. Each S^p_{bel} can then be paired with a region of Ω, and then we'd be able to run more or less the same interpretive strategy on Robo this time using a strictly weaker counterfactual uniqueness condition: If $\mathcal{P} \neq \mathcal{P}^\star$, then there is an p such that $\mathcal{R}(\mathcal{P}_p) \neq \mathcal{R}(\mathcal{P}^\star_p)$.

case for a functionalist, where we know that the actual facts on the ground are going to match up perfectly with the theories we're using to define our theoretical terms. In a more realistic case—say, the interpretation of the psychophysical states of a human—we can't presuppose that an interpretee will *precisely* instantiate our favourite decision theory $\mathcal{D}$, whatever that theory ends up being. Indeed, we should expect that even the perfect human exemplar *won't* exactly instantiate $\mathcal{D}$. The kinds of decision theories that exist nowadays at most represent the major *difference makers* that factor into our choices—those factors that do most of the causal heavy lifting in typical cases. They almost certainly do not represent the full and complete causal structure of our decision-making processes. Human psychology is a messy thing; the human brain is even messier. Since no ordinary, well-functioning human being is likely to perfectly instantiate any decision theory we come up with *precisely*, the most we should expect is to find states that fit *close enough* with expectations according to theory $\mathcal{D}$.

Secondly, we also need to account for cases of 'madness'—the assignment of beliefs and desires even to those whose mental faculties aren't functioning like they're supposed to. Imagine, for example, that some of Robo's screws came loose, and his behaviour went haywire: his actual and counterfactual preferences no longer conform so well with $\mathcal{D}$, and we can't run the kind of interpretive strategy we used earlier. Some of the anti-individualistic moves that Lewis makes can help here, as well as a touch of normalisation. The natural thing would to consider the causal profile of Robo's *bel* and *des* compartments, *were* he to be functioning properly in normal circumstances. Likewise for 'mad belief': what matters is not so much what an individual person's beliefs do, but what they *ought* to do—their causal properties when everything is working well, in the typical member of the relevant kind. Of course, all this means that it will be even harder to "read off" an agent's partial beliefs from her preferences—at best we could expect strong correlations between having such-and-such partial beliefs and having such-and-such preferences, but we won't get anything like an entailment relation between them. I take it that this is a good result, and strong correlations are all we should be expecting to find.

Finally, we haven't taken into account any of the (many) other roles that our partial beliefs play. Christensen, for example, notes that:

> … it seems that even within the realm of explaining behaviour, degrees of belief function in ways additional to explaining preferences (and thereby choice-behaviour). For example, we may explain someone coming off well socially on the basis of her high confidence that she will be liked. Or we may explain an athlete's poor performance by citing his low confidence that he would succeed… it is an important psychological fact that a person's beliefs—the way she represents the world—affect the way she behaves in countless ways that have nothing directly to do with the decision theorist's paradigm of cost-benefit calculation. (2001, p. 361)

Another example is the relation between confidence and the placebo effect. And, perhaps most obviously, many of our partial beliefs are formed and adjusted as a rational response to perceptual evidence—some reference to *that* role at least should surely turn up somewhere in any good functionalist definition of what it is to have partial beliefs. Because of this, as Christensen put it, "the move of settling on just one

of these connections—even an important one—as definitional comes to look highly suspicious" (p. 362). I'm strongly inclined to agree. When we try to characterise what partial beliefs *are* in terms of what they *do*, we ought not to focus solely on the belief-preference connection. That role is one amongst many, and at least some of those other roles are plausibly required to fully answer the CHARACTERISATION question.

We've already seen that at least some historical proponents of preference-centrism allowed that other factors might be relevant to charactering what partial beliefs are. But to close off the essay I want to discuss another, more interesting response that can be made on behalf of preference-centrism: merely noting the existence of other theoretical roles and the need to take those into account when formulating any adequate functional definition is entirely compatible with saying that the connection with preferences is *special*.

A simple toy model will help to illustrate what I mean. Imagine that there are exactly three causal roles—A, B, and C—associated with being in a particular total state of partial belief, $\mathcal{P}$. We want to characterise what it is to be in state $\mathcal{P}$ in terms of overall *fit* with those three roles, where considerations of fitness are weighted by the relative importance or centrality of each role. So we assign to each role a numerical weight to reflect its centrality—let's say: $A = 3$, $B = 2$, $C = 1$. We mark the extent to which some state S fits the total causal role associated with $\mathcal{P}$ by summing the weights of the specific roles that S manages to satisfy. Finally, we define what it is to be in state $\mathcal{P}$ as being in that state (or one of the states) that is a *best deserver of* $\{A, B, C\}$; i.e., some state, whatever it may be, that fits the total causal role *well enough* (≥ 4), and *at least as well* as anything else does. Now, if any state is a best deserver of $\{A, B, C\}$, then it must satisfy role A. But satisfying that role isn't *sufficient*, and the non-necessary roles B and C are still important in their own way. We cannot fully characterise the state of *being in* $\mathcal{P}$ without mentioning roles B and C somewhere, and in that sense it's crucial to take them into account. However, those roles are not essential to *being in* $\mathcal{P}$ in the same way that role A is.

I imagine that the role of partial belief in connection with preferences is something like role A. Other theoretical roles matter to characterising what partial beliefs are, some more than others. It might even be impossible to formulate an adequate characterisation of partial belief without mentioning all of the associated theoretical roles. But a large part of the 'work' in fixing the characterisation of what our partial beliefs *are* can be achieved by reference to the belief-preference connection alone—sort that part out and you're most of the way there.

Why should that role be given so much weight? Partly this comes down to intuitions about the relative centrality of different theoretical roles. In principle, I could imagine myself accepting that a state could be a partial belief while not explanatorily tied to the placebo effect in any interesting way; on the other hand, if it's not connected up to preferences in that kind of way then it just ain't partial belief. The same *arguably* goes even for the connection between partial belief and evidence. When we interpret each other, we seem to give primacy to the belief-preference connection over the evidence-belief connection. Imagine the child who has no evidence that there's a werewolf in the closet, and overwhelming evidence that there isn't, but acts in a way that would be best explained by high confidence in the existence of closet-dwelling monsters—the

interpretation goes with what explains the child's behaviour, however poorly that fits with the evidence.

However, here's another, more substantive reason: it's *only* through that connection that we'll get an entirely satisfactory answer to the questions of JUSTIFICATION. Whatever our partial beliefs are, and whatever other roles they might play in our cognitive economy, they must have a certain rich relational structure which explains why it makes sense to measure them in the way that we do. As I argued in Sect. 3, whatever answers the JUSTIFICATION question will likely form a *major* and *essential* part of the answering the CHARACTERISATION question. The belief-preference connection promises to explain why we think partial beliefs have this rich structure. Moreover, nothing else comes as close to adequately accounting for why we're able to measure partial beliefs on the [0,1] interval or some appropriate generalisation thereof, how strengths of belief can be ratio-scalable even for non-ideal agents like us, and how we can make plausible theoretical sense of interpersonal comparisons.

There are other accounts of how we might answer the JUSTIFICATION question that I haven't discussed here, and if you're a fan of those other ways then I don't expect to have said enough to convince you.[11] That's a debate better left for elsewhere. I have developed and defended my preferred view on the measurement of partial belief and the shortcomings of other accounts in other works (see especially Elliott 2017b, 2019) and I won't repeat my reasons again here. This paper is a defence of preference-centrism against prominent objections, not an attack the views of others. What's important is to recognise that (a) there is plenty of scope within preference-centrism for accepting that partial beliefs can play many theoretical roles, and (b) that those other roles can also be important for answering the CHARACTERISATION question, yet (c) taking the belief-preference connection to be of special or unique importance is not at all unmotivated or inherently implausible.

8 Conclusion

To summarise, then, the kind of functionalism I'm advocating says that a human agent has partial beliefs $\mathcal{P}$ just in case that agent is in a state S (perhaps a brain state) that, in a typical and well-functioning human being, is the best deserver of the total causal role associated with $\mathcal{P}$, where the belief-preference connection—both actual and counterfactual—is to be given special importance in fixing what makes S a good enough deserver of that role. The resulting view isn't behaviourist, instrumentalist, or anti-realist. It doesn't ignore other functional roles, and it has a straightforward answer to the non-denominational monk problem. More generally, it is quite distantly removed from any implication of ENTAILMENT and LIMITED ENTAILMENT, even in

[11] For example, many have argued we can explain the facts about how we measure partial beliefs by reference to rational constraints on comparative confidence relations, where no assumption is made about how comparative confidences connect to preferences. Alternatively, if you think partial beliefs are beliefs about probabilities, then you'll presumably think that whatever explains the measurement of those probabilities will derivatively explain the measurement of partial belief. Still others might think that there's something to the way our beliefs are formed in response to evidence which underlies the correct theory of their measurement, though I've yet to see this idea developed anywhere in print.

special circumstances—though it does (rightly) entail strong correlations between having certain patterns of preferences and having certain partial belief states.

There is still work to be done in fully spelling out the theoretical role associated with our partial beliefs and basic desires. There are also long-standing objections and worries with functionalism in general which must be dealt with. But there's a broad class of objections to preference-centrism that *don't* work, and it's high time we put them to rest.

Acknowledgements Thanks are due to two referees, Thomas Brouwer, Nicholas DiBella, Will Gamester, Al Hájek, Jessica Isserow, Jessica Keiser, Gerald Lang, Robbie Williams, and audiences at the 2018 AAP/NZAAP AGM (Wellington) and the 'What Are Degrees of Belief?' workshop (Leeds).

References

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics*, *34*(2), 199–205.

Braddon-Mitchell, D. (2003). Qualia and analytical conditionals. *The Journal of Philosophy*, *100*(3), 111–135.

Bradley, D., & Leitgeb, H. (2006). When betting odds and credences come apart: more worries for dutch book arguments. *Analysis*, *66*(2), 119–127.

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

Bunge, M. (1973). On confusing 'measure' with 'measurement' in the methodology of behavioral science. *The methodological unity of science* (pp. 105–122). Dordrecht: D. Reidel Publishing.

Chalmers, D. (2002). Does conceivability entail possibility? In T. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 145–200). Oxford: Oxford University Press.

Christensen, D. (2001). Preference-based arguments for probabilism. *Philosophy of Science*, *68*(3), 356–376.

Christensen, D. (2004). *Putting logic in its place: formal constraints on rational belief*. Oxford: Oxford University Press.

Cozic, M., & Hill, B. (2015). Representation theorems and the semantics of decision-theoretic concepts. *Journal of Economic Methodology*, *22*(3), 292–311. https://doi.org/10.1080/1350178X.2015.1071503.

Davidson, D. (1980). Toward a unified theory of meaning and action. *Grazer Philosophische Studien*, *11*, 1–12.

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, *87*(6), 279–328.

Davidson, D. (2004). Expressing evaluations. *Problems of rationality* (pp. 19–38). Oxford: Oxford University Press.

Davidson, D., & Suppes, P. (1956). A finitistic axiomatization of subjective probability and utility. *Econometrica*, *24*(3), 264–275.

Davidson, D., Suppes, P., & Siegel, S. (1957). *Decision making; an experimental approach*. Palo Alto: Stanford University Press.

de Finetti, B. (1937). Foresight: Its logical laws in subjective sources. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (Vol. I, pp. 134–174). New York: Springer.

Elliott, E. (2017a). Ramsey without ethical neutrality: A new representation theorem. *Mind*, *126*(501), 1–51.

Elliott, E. (2017b). A representation theorem for frequently irrational agents. *Journal of Philosophical Logic*, *46*(5), 467–506.

Elliott, E. (2019). Comparativism and the measurement of partial belief. http://www.edwardjrelliott.com/research.html.

Eriksson, L., & Hájek, A. (2007). What are degrees of belief? *Studia Logica*, *86*(2), 183–213.

Eriksson, L., & Rabinowicz, W. (2013). The interference problem for the betting interpretation of degrees of belief. *Synthese*, *190*, 809–830.
Fishburn, P. C. (1967). Preference-based definitions of subjective probability. *The Annals of Mathematical Statistics*, *38*(6), 1605–1617.
Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.
Hájek, A. (2008). Arguments for-or against-probabilism? *British Journal for the Philosophy of Science*, *59*(4), 793–819.
Jeffrey, R. C. (1965). *The logic of decision*. Chicago: University of Chicago Press.
Joyce, J. M. (1999). *The foundations of causal decision theory*. New York: Cambridge University Press.
Joyce, J. M. (2000). Why we still need a logic of decision. *Philosophy of Science*, *67*(Supplement), S1–S13.
Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. vol. I: Additive and polynomial representations*. Cambridge: Academic Press.
Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, *67*(13), 427–446.
Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258.
Lewis, D. (1974). Radical interpretation. *Synthese*, *27*(3), 331–344.
Lewis, D. (1979). Attitudes De Dicto and De Se. *The Philosophical Review*, *88*(4), 513–543.
Lewis, D. (1980). Mad pain and martian pain. In N. Block (Ed.), *Readings in philosophy of psychology* (pp. 216–222). Cambridge: Harvard University Press.
Lewis, D. (1983a). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377.
Lewis, D. (1983b). Postscripts to "Radical Interpretation". *Philosophical papers* (Vol. 1, pp. 119–121). New York: Oxford University Press.
Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty*, *4*, 29–59.
Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.
Meacham, C. J. G., & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, *89*(4), 641–663. https://doi.org/10.1080/00048402.2010.510529.
Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory: Issues and advances* (pp. 147–175). Oxford: Basil Blackwater.
Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). Oxon: Routledge.
Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour: An addendum. *Economica*, *5*, 353–354.
Sarin, R., & Wakker, P. (1992). A simple axiomatization of nonadditive expected utility. *Econometrica*, *60*(6), 1255–1272.
Savage, L. J. (1954). *The foundations of statistics*. New York: Dover.
Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2010). Coherent choice functions under uncertainty. *Synthese*, *172*, 157–176.
Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*, 332–382.
Stefánsson, H. O. (2016). What Is 'Real' in probabilism? *Australasian Journal of Philosophy*, *97*(3), 573–587. https://doi.org/10.1080/00048402.2016.1224906.
Stefánsson, H. O. (2018). On the ratio challenge for comparativism. *Australasian Journal of Philosophy*, *96*(2), 380–390.
Suppes, P., & Zinnes, J. (1963). Basic measurement theory. In D. R. Luce (Ed.), *Handbook of mathematical psychology*. New York: Wiley.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.
Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman & Hall.
Weatherson, B. (2016). Games, beliefs and credences. *Philosophy and Phenomenological Research*, *92*(2), 209–236.
Yalcin, S. (2012). Bayesian expressivism. *Proceedings of the Aristotelian Society*, *112*, 123–160.