Betting Against the Zen Monk: On Preferences and Partial Belief

Edward Elliott*

School of Philosophy, Religion and History of Science
University of Leeds

Abstract

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' (Ramsey 1931). I take this as a good starting point for a metaphysics of partial belief—that is, I think that the connection between partial beliefs and preferences is key to understanding what partial beliefs are, and how they're measured. But this view is not as popular as it once was: nowadays, the preference-centric approach is frequently dismissed out-of-hand as behaviouristic, or unpalatably anti-realist. Furthermore, cases like Eriksson & Hájek's (2007) preferenceless *Zen monk* have suggested to many philosophers that any account of partial belief that ties them too closely to preferences is hopelessly flawed. Using the Zen monk as my stalking horse, in this paper I provide a defence of preference-centric accounts of partial belief.

§1. Introduction

The topic of this paper is the metaphysics of partial belief, and in particular the relationship between partial beliefs and preferences. In short, 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 importance. I'll say what I mean by this in more detail 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 (1931), 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' (p. 169).

Once upon a time, the *preference-centric* approach to understanding partial beliefs was the norm, and in some circles it still is. But in philosophy, nowadays, it often gets a bad rap: it's frequently dismissed out-of-hand as outdated and behaviouristic, or as committed to anti-realism or instrumentalism (as if that were automatically a bad thing). Specific instances of the idea have been criticised by, *inter alia*, Christensen (2001), Hájek (2008), Meacham & Weisberg (2011), Eriksson & Rabinowicz (2013), Stefánsson (2016, 2018); and the general approach has been most thoroughly criticised by Eriksson & Hájek (2007),

*e.j.r.elliott@leeds.ac.uk. Draft of October 2, 2018. Comments welcome; helpful comments doubly so.

whose preferenceless *Zen monk* and closely related thought experiments have suggested to many philosophers that any account of partial beliefs and their measurement that ties them too closely to preferences is hopelessly flawed.

I think this is too quick. There's a lot to be said in favour of preference-centric approaches 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 a defence of the approach, as well as an exposition. I won't try to explicitly address all of the specific criticisms that exist, for they are many and various. But the majority of those criticisms share similar themes with the Zen monk; so, using that case as my stalking horse, I'll touch on the most common of them.

After providing some background in the next section, in §3 I will consider the Zen monk in some detail. I'll show that a suitably generalised version of the problem that the Zen monk poses survives a common first-pass response to Eriksson & Hájek's specific case. However, in §4 I argue that the (generalised) problem is of only limited concern to some varieties of preference-centrism. Finally, in §5 I put forward a specific, realist (with a capital 'R') variety of preference-centrism that readily deals with Zen monk problem and related concerns.

§2. Preference-Centrism and the Measurement of Belief

Before we get to preference-centrism and the Zen monk, let me first set the scene with three general assumptions, none of which I'll make any attempt to defend here. I'm not certain that any of these assumptions are true, and when taken in conjunction they're certainly not obvious. 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'll assume that agents both real and ideal have partial beliefs. In particular, I don't imagine partial beliefs as some idealised attitude that can only be meaningfully attributed to Bayesian angels contemplating probabilities in a faraway possible world. *You* and *I* have partial beliefs, and I'm inclined to think that this fact is grounded in something objective about us—presumably, something going on inside our heads, or perhaps something about how the stuff in our heads is causally connected to our behaviour, environment, and/or ancestry.

Second, I'll assume that every partial belief comes with a *strength*, and moreover 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 and operations amongst those strengths are reflected and preserved in mathematical relations and operations amongst their assigned values. Such an assignment is usually known as a *measurement scale*.[1] Since it's the most familiar way of

[1] My use of the word 'measurement' and it's cognates is in reference to abstract measurement scales, not to the empirical process also known as *measurement*. Compare, for example, (i) measuring temperatures by assigning to each temperature a numerical value (e.g., in degrees Celsius) that reflects its properties relative to other temperatures, with (ii) measuring the temperature of a liquid by means of a thermometer. The ambiguity is unfortunate, but it's also long entrenched (cf. Suppes & 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. Let me emphasise that this paper will not address the empirical question of whether and how strengths of belief might be determined through observations of choice behaviour.

doing these things, I will suppose throughout that strengths of belief are measurable on a ratio scale with real values between 0 and 1. But nothing really hangs on that choice. Instead of real numbers, we might make do with a few integers, or we could use an expanded space of values including hyperreals or surreals and so on—and if we wanted to get really fancy we could use intervals, fuzzy sets, n-dimensional vectors, or whatever other exotic numerical structures mathematicians might cook up.

Third, I'll assume that the strength of a belief does not belong to the *content* of that belief. Partial beliefs are not, or they need not be, *about* probabilities. I'm confident that Melbourne has better coffee than Sydney, and this isn't a belief about chances, frequencies, propensities, evidential relations, fair betting prices, or any other vaguely probabilistic phenomenon—it's just a (very well-justified) belief about the best place to find coffee in Australia. It's helpful if we think of the strength of a belief as something that attaches to the attitude itself, and wholly separable from its content. See (Christensen 2004, pp. 18ff) and (Weatherson 2016) for more discussion on this point.

Summarising my three assumptions, let's say that strength of belief is a *genuine psychological quantity*. If this is correct, then a number of important and closely interlocking questions naturally arise:

Measurement. How should we *measure* strength of belief? i.e., should we use real numbers? Are strengths of belief ratio scalable? Are the beliefs of different agents measurable on the same scale, and hence comparable?

Justification. How can we *justify* measuring strengths of belief as we do? i.e., on what basis are we allowed to say that strengths of belief have *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?

Say that someone adopts a preference-centric approach to understanding partial belief if, in answering these kinds of questions, they posit a *uniquely central role* for the relationship between partial beliefs and preferences. (I'll explain that more momentarily.) I take it that paradigm instances of preference-centrism can be found in de Finetti's (1937) *betting interpretation*, including its most recent incarnations that involve lower and/or upper previsions (see esp. Walley 1991). We also find preference-centrism in the betting interpretation's more general cousins, those accounts of partial belief that relate them to preferences *via* representation theorems (e.g., Ramsey 1931; Savage 1954; Anscombe & Aumann 1963, Cozic & Hill 2015, amongst many others); 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, 1983).[2]

[2] The views of the authors listed here have been called many names—'pragmatism', 'behaviourism', 'the preference interpretation', 'the thesis of the primacy of practical reason'. I apologise for adding yet another name to the mix with 'preference-centrism', but I do so because those terms don't have a fixed meaning across different users. For example, some use 'the betting interpretation' to refer to a very literalistic reading of de Finetti's proposal in his (1937), while others take it to also include the views of Ramsey, Savage, or anything else that vaguely resembles defining partial belief in terms of preference. Another example: 'pragmatism' as defined in (Joyce 1999, p. 90) designates a normative thesis that's logically independent of preference-centrism (roughly: epistemic rationality is derivative upon pragmatic rationality); whereas in (Ramsey 1927, p. 5), 'pragmatism' picks out a non-normative, metaphysical thesis (roughly: beliefs are sets of behaviours).

So I've just said that a preference-centric approach is 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 paradigm instances just cited. There is a common core to all of them, however, and I find the relative simplicity of the betting interpretation especially useful for teasing it out.

Suppose that Sally would prefer winning \$1 to winning nothing, 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.,

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

We should expect that Sally will prefer being given \$1 straight to taking the bet β, and β to nothing. That much is obvious. The central insight of the betting interpretation is that the *extent* to which Sally prefers \$1 to β (and β to \$0) is proportionate to the strength of her belief in p. Combine that with the idea that the utility Sally attaches to β can be read off the prices she'd be willing to buy and sell it at, and you've got the betting interpretation in a nutshell.

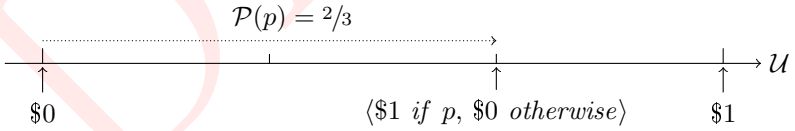
Let $\mathcal{U}$ be a measure of Sally's preferences on at least an interval scale (i.e., a utility function), and $\mathcal{P}$ a measure of her beliefs. Then, suppressing some annoying details, according to expected utility theory $\mathcal{P}$ is a probability function and

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

We can then re-write this in ratio form to better represent the role that Sally's belief in p plays:

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

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



Now, you could take the ratio formula as a more or less complete answer to the Characterisation question, and some people certainly seem to have done so. But the point of this discussion *isn't* that the ratio formula is a good definition of the strength of belief—it's not. There are compelling counterexamples, and those annoying details do matter (see esp. Eriksson & Rabinowicz 2013). Moreover, we mere human beings probably don't conform quite so well to expected utility theory, at least not in all cases and at all times. The betting interpretation comes with a built-in rationality assumption that doesn't seem too plausible if treated as a universal generalisation about decisions are made.

What I want to highlight, rather, 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* these lines. The various

alternatives to expected utility theory that we see today (whether developed for normative or descriptive/empirical purposes) all retain the same basic structure of that theory. Most of them, it's fair to say, just *are* expected utility theory with some added bells and whistles—a fudge factor here and a bit of generalisation there—and they tend to have the ratio formula or something close by as a special case. (See, for example, Luce and Fishburn 1991, Tversky and Kahneman 1992; Sarin and Wakker 1992; Seidenfeld et al. 2010; Buchak 2013.)

According to advocates of the preference-centric approach, that tells us something crucially important about our beliefs and how they're measured. Note especially that the ratio formula doesn't just suggest a way of answering the Characterisation question—it also suggests answers to the 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 always sit somewhere on that interval; they're ratio-scalable because ratios of utility differences are meaningful; and they admit interpersonal comparisons because similar strengths of belief will have similar roles in fixing the utilities of choices under uncertainty for different agents.

Given this, the kind of preference-centric approach that I would want to defend says that the belief-preference relationship, as captured at least roughly by the ratio formula,

(i) constitutes *one of* the most central theoretical roles for partial beliefs,
(ii) is *necessary* for answering the Characterisation question, and
(iii) is *uniquely necessary* for answering the Justification question.

Almost everyone would agree with the first claim. Many would also agree with the second, which says that if something's not connected to preferences in that kind of way, then it's just not a partial belief. My guess is that just as many would be happy to deny it. But it's claim (iii) which is the most controversial, and which makes the conjunction of the three preference-centric.

Let me put some flesh on those bones. (More still will be said again in §5.) I suspect that any plausible answer to the Characterisation question will be broadly functionalist in character, and will make central reference to the various causal and explanatory roles that partial beliefs play in the theories in which they figure (*à la* Lewis 1970). As such, I agree with Christensen (2001), Eriksson & Hájek (2007), and others, that we ought not to focus *only* on the belief-preference relationship when we characterise what partial beliefs *are* in terms of what they *do*. That role is only one amongst many. Christensen notes several examples of other roles that partial beliefs also play:

> … 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. (Christensen 2001, p. 361)

Another clear and important example is the relation between confidence and the placebo effect. And, perhaps most obviously, our partial beliefs are formed as a rational response to evidence—some reference to *that* role at least should surely

turn up somewhere in any plausible functionalist characterisation of what it is to have partial beliefs. It may even show up as a necessary condition, though I've yet to be convinced that it should.

I would argue, however, that the connection with preferences is *primary*: other roles might matter, at least in principle, to saying what partial beliefs are, but the belief-preference role should be given the most weight. A simple toy model might help to give a concrete illustration of what I mean here. Imagine that there are exactly three causal roles—A, B, and C—associated with being in a particular 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 each role a weight to reflect its absolute importance—let's say $A = 3$, $B = C = 1$—and 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 actually manages to satisfy. We define what it is to be in state $\mathcal{P}$ as being in that state S that's the *unique best deserver* of $\{A, B, C\}$; i.e., that state, whatever it may be, that fits the total role *well enough* (> 3), and *better* than any other state does. Now, if any state is the unique best deserver of $\{A, B, C\}$, then it must satisfy role A. But satisfying that role isn't *sufficient*, and either of the minor non-necessary roles B and C might still matter in principle when it comes to deciding whether to count S as the unique best deserver of $\{A, B, C\}$. I imagine that the role of partial belief in connection with preferences is much like role A in this example: almost all of the work, as it were, in fixing its meaning *can* be achieved by reference to that role alone, but other factors *might* make a difference as well, and thus can't be excluded from any complete analysis.

Why should that role be given more weight? Because it's *only* through that connection that we'll get an entirely satisfactory answer to the JUSTIFICATION questions. There's a special role for preferences in our understanding of what partial beliefs are, and that's in explaining where the numbers come from and why we're justified in interpreting those numbers the way we do. Nothing else comes close to 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 myself, and how we can make plausible theoretical sense of interpersonal comparisons. I have developed and defended these claims at length in other works (see especially Elliott 2017b; Elliott MS), so I won't repeat my reasons again here. What's more important for present purposes is that you get a sense of what preference-centrism involves, and what it might take to argue in favour of it.

Finally, the reader may have noticed at this point that I haven't pinned down what I mean by 'preference', and what the basic objects of preference are. Broadly speaking, there are two 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), and it tends to be what philosophers have in mind when they talk about preferences. And 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 usually think of preference as a propositional attitude with very close causal and conceptual ties to our choice behaviour, such that in the most (but not all) cases we can read the former off of the latter, and the latter off of

the former.[3] 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 won't hang on the disambiguation, so feel free to pick whichever 'preference' is your preference.

§3. Will the Real Zen Monk Please Stand Up?

We now have enough background to talk about the 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* [*partial beliefs*] *to preferences is refuted.* (Eriksson & Hájek 2007, p. 194, emphasis added)

The Zen monk is clearly a problem for some versions of the preference-centric approach—those according to which partial beliefs just *are* preferences, or more precisely, according to which the facts about our partial beliefs are in general *a priori* entailed by the facts about our preferences. That kind of preference-centrism is often at the forefront of critical discussions, but I'm skeptical that many people have ever actually held such views. There are plenty of ways that preferences can *and have been* taken to play a uniquely central role in our account of what partial beliefs are without playing *those* roles. More importantly, the conceptual possibility of the Zen monk is consistent with (in fact: implied by) many accounts of partial belief that conceptually tie them to preferences.

But before we get to all that, in the rest 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 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 that the Zen monk isn't *prima facie* negatively inconceivable, to use David Chalmers' two-fold categorisation scheme (see Chalmers 2002). There's no obvious contradiction implicit in the description of the case. I don't know whether the monk is ideally positively conceivable, but there's no value to butting heads over it. We simply don't need the Zen monk to get at the worry that's driving Eriksson & Hájek's case.

To see this, let's put that worry very abstractly.[4] Suppose we start with some background decision theory, $\mathcal{D}$. If $\mathcal{D}$ looks anything at all like expected utility theory—and a reminder: the vast majority of decision theories do—then $\mathcal{D}$ will say that an agent's overall preferences are determined by

[3] See Maher 1993, pp. 12ff, for a thorough discussion on the relationship between the two notions of preference, and some of the difficulties associated with equating the two.

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

(i) her *partial beliefs* (symbolised $\mathcal{P}$),
(ii) her *basic desires* (symbolised henceforth with $\mathcal{U}$), and
(iii) any additional factor(s) (symbolised $\mathcal{F}$),

where the $\mathcal{P}$, $\mathcal{U}$, and $\mathcal{F}$ may be restricted to some special class, the 'admissible' determinants—e.g., $\mathcal{D}$ might build in the presupposition that $\mathcal{P}$ is probabilistically coherent, or that $\mathcal{U}$ is finitely bounded.

(So, for example, on Savage's (1954) way of modelling decision-making, agents' preferences over their actions are determined by their basic preferences over the ultimate 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 & Fishburn 1991; Starmer 2000; Buchak 2013). In Ramsey's and related models (Ramsey 1931; Davidson et al. 1957; Elliott 2017a,b), preferences for bets are fixed by agents' utilities for the bets' possible 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), the agents' preference rankings over arbitrary propositions can be described as a function of their utilities for possible worlds plus a probability distribution over the space of those worlds.)

Given this, we could represent the content of the theory $\mathcal{D}$ as a function from admissible $(\mathcal{P},\mathcal{U},\mathcal{F})$ triples to an overall pattern of preferences, which we'll symbolise 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), then the agent's general preferences will be likewise trivial (i.e., $\succsim$ encodes no preferences whatsoever). An easy and obvious example is expected utility theory, as it was described in §2. If Sally has no interests in money, so $\mathcal{U}(\$1) = \mathcal{U}(\$0)$, then $\mathcal{U}(\beta)$ remains unchanged 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 interests are concerned, then there's no reason to choose anything over anything else, and there's no 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 $\mathcal{P} \neq \mathcal{P}'$, under most decision theories $\mathcal{D}$,

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

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 beliefs, then they will have the same (trivial) preferences, so there's no way to read the differences between their beliefs off of their preferences. In other words, if Zee and Zed are conceptually possible, then the following thesis is false:

Entailment. The facts about an agent's partial beliefs are in all cases *a priori* entailed by the facts about what preferences she has.

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 $(\mathcal{P}, \mathcal{U}^{tr}, \mathcal{F})$ and $(\mathcal{P}', \mathcal{U}^{tr}, \mathcal{F}')$ really *aren't* conceptually possible. Indeed, let's suppose more generally that the very idea of having no preferences whatsoever is incoherent. Does this change anything? Not at all: there are other ways to get the same result. This is evident from the many representation theorems for 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 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 set of constraints $C_1, C_2, ..., C_n$ on a set of preferences $\succsim$ is sufficient (and in some rare cases, necessary) to ensure:

Existence. There exists an admissible $(\mathcal{P}, \mathcal{U}, \mathcal{F})$ such that $\mathcal{D}(\mathcal{P}, \mathcal{U}, \mathcal{F}) = \succsim$.

The constraints $C_1, C_2, ..., C_n$ ensure Existence, in other words, only if any set of preferences that satisfies $C_1, C_2, ..., C_n$ is *consistent* with $\mathcal{D}$. A representation theorem will also usually be given alongside a uniqueness theorem, which might come in a variety of forms. The particular kind of uniqueness result that will be of interest to us here is:

Uniqueness. For any admissible $(\mathcal{P}, \mathcal{U}, \mathcal{F})$, $(\mathcal{P}', \mathcal{U}', \mathcal{F}')$, if $\succsim = \mathcal{D}(\mathcal{P}, \mathcal{U}, \mathcal{F}) = \mathcal{D}(\mathcal{P}', \mathcal{U}', \mathcal{F}')$, then $\mathcal{P} = \mathcal{P}'$.

That is, constraints $C_1, C_2, ..., C_n$ imply Uniqueness just in case, if a set of preferences satisfies $C_1, C_2, ..., C_n$, then those preferences could only have been generated by one set of beliefs consistently with $\mathcal{D}$. If Uniqueness holds, and it's assumed that beliefs and preferences are related as per the theory $\mathcal{D}$, then—and only then—we will be able to determine what beliefs an agent has merely be considering what preferences she has.

To be clear: there's no guarantee that there *will* be constraints on $\succsim$ consistent with $\mathcal{D}$ that ensure 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 Uniqueness, there will in general be a gap between those constraints sufficient for Existence, and those constraints sufficient for Uniqueness. This is usually easy to prove, and it means that we should expect there to be at least *some* sets of preferences that are consistent (according to $\mathcal{D}$) with distinct sets of beliefs.

So back to the Zen monk. Included within either the constraints that imply Existence or those that imply Uniqueness, we always find the following condition:[5]

Non-Triviality. It's not the case that $\succsim$ is trivial.

And here's the important part: as a rule, the constraints that ensure Existence, *plus* Non-Triviality, don't ensure Uniqueness. That is, *even with the Zen*

[5] In Savage's (1954) theorem, Non-Triviality is his condition P5, which is necessary to prove his strong uniqueness theorem. In Ramsey's (1931) theorem, Non-Triviality is an immediate 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 include constraints sufficient to entail 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 regularly, since it's needed (amongst other things) to deal with the Zen monk.

monk ruled out by fiat, we will still find that on any of the usual decision theories there are going to be sets of preferences consistent with distinct sets of beliefs. And note that, since Non-Triviality holds only if the agent has interesting and varied basic desires, the earlier assumption that $\mathcal{U}^{tr}$ is admissible is superfluous.

Just like the original Zen monk case, the upshot of this argument is that Entailment is false. The argument doesn't come with a nice pithy case to exercise your imagination, and its typical instances 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 under which Uniqueness holds true, *minus* Non-Triviality, are universally regarded as the *least* reasonable 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 rankings must be 'atomless' and defined over an uncountable set of states. (See Joyce 1999, §3.3, and Fishburn 1970, pp. 193ff, for discussion.)

More generally, one typically only ensures a result like Uniqueness by making very strong richness assumptions about the domain of the preference relation, and then combining that rich structure with 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)

In short: even if you think Zen monks are conceptually impossible, if your favourite decision theory looks more or less like expected utility theory, then it's still highly plausible that there will be possible sets of preferences which, according to that theory, are consistent with more than one set of partial beliefs. Call this the *non-denominational monk problem*. Debating about the possibility of the Zen monk seems parochial when there are still plenty of non-denominational monks to deal with.

§4. Who's Afraid of the Big Bad Monk?

We should all agree that Entailment is false. There is 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 has. And if that's all that Eriksson & Hájek meant when they said that "any account that conceptually ties credences to preferences is refuted," then I agree. But as I mentioned earlier, 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 §5, in this section I want to consider the extent to which Entailment can be thought of as a 'core commitment' of preference-centrism.

Let's get some general points out of the way first. First: there are many ways to read the Characterisation question, and we certainly don't have to think that every attempt to answer is aimed at faithfully recapturing our *conceptual*

commitments. That's one kind of project a preference-centric author might be engaged in. Another 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 (especially modern applications of the betting interpretation) is intended as nothing more than an *operationalisation*: enough to pin down a clear enough meaning for present purposes, where those purposes don't call for a particularly sophisticated or plausible theory of the mind. 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 our concept of *partial belief*.[6]

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. (pp. 179–80)

[6] 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 'Water is the watery stuff' isn't metaphysically necessary, 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 doesn't imply metaphysical impossibility.

But to interpret Ramsey in this way is to forget the very idea which motivated his theory—that the degree of a belief is a *causal* property of it. A partial belief is something separate from and prior to preferences, that comes with an attached quantity that explains the extent to which the belief feeds into our preferences when we're rational. Ramsey uses this kind of causal language throughout (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 read off of their choices given the assumption that those choices were made 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—but what it is to have a belief with a given strength is characterised by reference to what that belief does in the context of rational decision making.

Relatedly, we frequently don't have enough information to know with confidence what some historical advocate of preference-centrism would say about the non-denominational monk problem. I'll take Ramsey as my example again—the same point will apply to Savage, Anscombe & Aumann, de Finetti, and others, *mutatis mutandis*. At most, Ramsey in 'Truth and Probability' might be read as committed to the claim that the $\succsim$-facts entail the $\mathcal{P}$-facts *under the assumption that* $\succsim$ *satisfies the conditions of his representation theorem*. And here's the rub: a set of preferences $\succsim$ will satisfy all eight of his conditions only if Uniqueness holds.[7] Given this, we can't attribute to Ramsey anything stronger than:

Limited Entailment. The facts about an agent's partial beliefs are in some cases *a priori* entailed by the facts about what preferences she has.

Or to put the point more directly: Ramsey (and others like him) can be read as at most giving us a story about what partial beliefs are that applies in exactly those cases where the Zen and non-denominational monk problems don't arise!

Turning to more recent advocates of preference-centrism, consider 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

[7] See (Bradley 2001) and (Elliott 2017a) for detailed expositions of Ramsey's theorem.

> 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 clearly accepts Limited Entailment, but it's not clear whether he would have accepted full-fledged Entailment. Set aside the brief mention of other 'criteria that go into evaluating interpretations' for now, we'll return to that in a short moment—what already makes it hard hard to know what Maher would say about non-denominational monk cases is that the representation theorem he employs to flesh out his interpretivist theory comes with a very strong uniqueness theorem, on par with Ramsey's and Savage's theorems (and explicitly includes a Non-Triviality constraint: see pp. 187–8, Axiom 2).

Here's two things Maher might say when confronted with the Zen monk:

(i) Since the monk's preferences maximise expected utility relative to both $(\mathcal{P}, \mathcal{U}^{tr})$ and $(\mathcal{P}', \mathcal{U}^{tr})$, there's no fact of the matter as to which of $\mathcal{P}$ or $\mathcal{P}'$ represents the monk's partial beliefs.

(ii) Since the monk's preferences maximise expected utility relative to both $(\mathcal{P}, \mathcal{U}^{tr})$ and $(\mathcal{P}', \mathcal{U}^{tr})$, other interpretive criteria might be called on to determine whether $\mathcal{P}$ or $\mathcal{P}'$ better represents the monk's partial beliefs.

I don't know which response Maher would prefer. But the more interesting and plausible of the two, I think, is (ii). 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 alignment with a long history of interpretivist thought.

Take, for example, the kind of interpretivism endorsed by David Lewis in 'Radical Interpretation' (1974). (See also Williams 2016, 2018, for a more recent version of interpretivism similar to and inspired by Lewis' own.) 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*. Rationalisation is very 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 choices pragmatically rational—i.e., to the extent that her preferences maximise expected utility with respect to those beliefs and utilities. Charity, on the other hand, is perhaps better seen as a motley collection of interpretive constraints, which in summary say 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 was not as clear as he could have been about how Rationalisation and Charity were supposed to interact, especially in cases where they might pull in different directions. In 'Radical Interpretation', he merely says that given

> … a source of information on Karl's behaviour and as a source of information on his life history of evidence, fill in [an assignment of partial beliefs and utilities] by means of the Rationalisation Principle and the Principle of Charity. (1974, p. 341)

Nevertheless, a later discussion is helpful. In 'New Work for a Theory of Universals', Lewis writes:

> If we rely on principles of [Rationalisation] to do the whole job, we can expect radical indeterminacy of interpretation. We need further constraints, of the sort 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 the principles of* [*Rationalisation*]. (1983, p. 375, emphasis added)

In other words, Lewisean interpretation is a two-step process. First and foremost, we solve for fit with preferences using the principle of Rationalisation. Charity then plays a secondary, tie-breaking role, filtering between those interpretations reckoned equally good by the principle of Rationalisation.

Since it will be relevant again in the next section, it's also worth noting that in 'New Work', Lewis gives Rationalisation a slightly updated gloss: an assignment of $\mathcal{P}$ and $\mathcal{U}$ shouldn't just rationalise the agent's *actual* preferences, but also the agent's present dispositions to change her preferences in various ways upon receiving different kinds of evidence (see p. 374). Thus, if the agent is disposed to change her preferences from $\succsim$ to $\succsim'$ upon receipt of some evidence E, then $(\mathcal{P},\mathcal{U})$ is a better interpretation not only the extent that $\succsim$ maximises expected utility with respect to $\mathcal{P}$ and $\mathcal{U}$, but also to the extent that $\succsim'$ maximises expected utility with respect to $(\mathcal{P}^E,\mathcal{U})$, where $\mathcal{P}^E$ picks out $\mathcal{P}$ conditionalised on E. Call this *Rationalisation**; since $\mathcal{P}$ conditionalised on a tautology is just $\mathcal{P}$, Rationalisation* incorporates strictly more information than Rationalisation as it was originally described in 'Radical Interpretation'.

Note, though, that the focus of Rationalisation* still on fit with preferences—that is, the principle is still geared towards interpretations that maximise the pragmatic rationality of the agent's preferences, the difference being that now it's the rationality of the agent's *actual* and *dispositional* preferences that matter. Fit with previous life history of evidence don't come into the picture at this stage of the interpretation; the only role for that consideration is to help us sift through what remains once we've done as much as we can fitting an interpretation to the agent's preferences. The view that results is clearly preference-centric, but isn't committed to Entailment nor even Limited Entailment.

So here's the general pattern we're seeing: 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 agent's preferences are enough to pin down the facts about her partial beliefs, in special cases where those preferences are non-trivial and satisfy certain other (very strong) constraints. But this need not be true in general, and outside of those special cases things might get messy. *This* is what advocates of preference-centrism typically believe, and have believed for a long while—*not* that partial beliefs just are preferences, nor even that the facts about partial beliefs are in general entailed by the facts about preferences. Entailment isn't a core commitment of preference-centrism.

§5. Preference-centric Functionalism

It's one thing to reject ENTAILMENT; that's easy if you're happy to assume away the hard cases, and merely put forward a story about what partial beliefs are and how they're measured that only applies under the very special circumstances where the preferences have the kind of rich structure they need to have to make the account work. It's quite another thing to develop an entirely general account of partial belief, one that actually has something definite and plausible to say about the proper interpretation of the non-denominational monkhood.

I believe that such an account can be developed along functionalist lines, and in this final section I want to sketch how I think it should go.[8] The key to solving the non-denominational monk problem is to consider not only those preferences that a set of partial beliefs actually gives rise to, but also the preferences they would give rise to under different kinds of conditions. It's only through the latter that we'll get a complete picture of the causal role that partial beliefs play in relation to preferences.

We've just seen a version of this insight in Lewis' Rationalisation*. From what was said in §3, we know that there will sometimes be sets of preferences $\succsim$ that could have been generated by distinct sets of partial beliefs, $\mathcal{P}$ and $\mathcal{P}'$. Consequently, if there *is* a causal difference between $\mathcal{P}$ and $\mathcal{P}'$, then that difference won't always be evident in the preferences that the beliefs generate *as a matter of fact*. But the distinctive causal potentials of $\mathcal{P}$ and $\mathcal{P}'$ *might* be evidenced instead if we consider how they dispose an agent towards different kinds of preferences given some evidence E. I suspect that taking these additional considerations into account will help to significantly narrow down the range of possible interpretations in *many* non-denominational monk cases, though not all. Lewis himself thought that Rationalisation* was never enough to pin down a unique interpretation, hence the need for Charity. (See Williams 2016 for a recent exposition of the issues here.)

However, a variation on the original Zen monk case suggests that even the Lewisean combination of Rationalisation* and Charity will be insufficient:

> Imagine two Zen Buddhist monks, Zee and Zed, who have partial beliefs but no preferences, because they both have trivial basic desires. Gazing peacefully at the scene before him, Zee believes that Mt. Everest stands at the other side of the valley, that K2 does not, and so on. Though they've lived essentially identical lives—each has the same basic life experiences and the same history of evidence—Zed disagrees with some of Zee's beliefs: he believes that K2 stands at the other side of the valley, and that Mt. Everest does not.

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 E, $\mathcal{P}^E$ 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 updated their beliefs in light of some evidence E. And since they're the same kind of being, with the same basic experiences

[8] To be clear: I intend for the following to be read as a first pass at a functionalist analysis or definition of *partial belief*, as capturing the basic structure of the concepts involved. I don't want to put it forward as an explication, operationalisation, or *a posteriori* identification.

and life history of evidence, there's nothing for Charity to operate on to let us distinguish between them. Conclusion: if Zee and Zed are conceptually possible, then there's more to our concept of partial belief than can be captured by those considerations that feed into Rationalisation* and Charity.

Note, by the way, that we'll come to the same conclusion regardless of how Rationalisation* and Charity are supposed to interact—that is, whether Charity merely plays a tie-breaking role, or whether the two considerations are weighted against each other somehow to produce the final interpretation. Rationalisation* is trivially satisfied, and the two monks are identical as far as Charity is concerned. We need an even richer 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 at this point for getting across the kind of functionalist account that 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. Moreover, we're told something about what this means, in terms of 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 a representation (or representations) of the way Robo takes the world to be, and a 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 in his interior. After a while, we make some crucial discoveries.

First, we can see that inside 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*, each which can be easily identified and distinguished. 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. Unfortunately, the causal relationship here is complicated: more often than we'd like there are quite distinct combined states that produce exactly the same behavioural dispositions. There's 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 states of either one of the compartments 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} of his *des* compartment. The result is a complete map of the different sets of preferences $\succsim$ that each combined (S_{bel}, S_{des}) state would generate. This gives us a way to start testing out some interpretive hypotheses. We start with possible assignments of sets of partial beliefs $\mathcal{P}$ to the states S_{bel}.

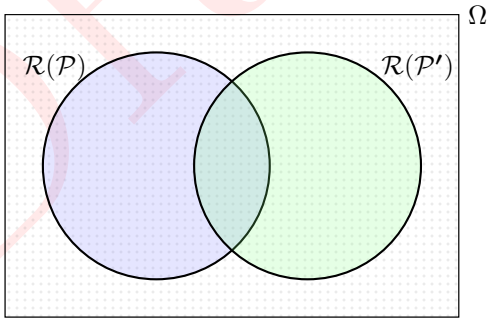
Since Robo was designed to instantiate $\mathcal{D}$, and he's functioning exactly as he's supposed to, his preferences under any combined state are always consistent with that theory. Therefore, let Ω pick out the space of all sets of preferences consistent with $\mathcal{D}$. We are able to associate 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 show that $\mathcal{D}$ implies:

Counterfactual Uniqueness. If $\mathcal{P} \neq \mathcal{P}'$, then $\mathcal{R}(\mathcal{P}) \neq \mathcal{R}(\mathcal{P}')$.

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 two different sets of beliefs, $\mathcal{R}(\mathcal{P})$ and $\mathcal{R}(\mathcal{P}')$, 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 thus belong to the intersection of every $\mathcal{R}(\mathcal{P})$, for every admissible $\mathcal{P}$.

So suppose Robo's actual preferences are indeed somewhere the intersection of $\mathcal{R}(\mathcal{P})$ and $\mathcal{R}(\mathcal{P}')$. If you want to, you can even suppose that he has the trivial set of preferences. This is just the sort of case where Uniqueness fails, so on

the basis of Robo's actual preferences alone we won't have enough information to decide between $\mathcal{P}$ or $\mathcal{P}'$ as the better interpretation. But if we allow ourselves counterfactual information, then we don't *need* Uniqueness: $\mathcal{R}(\mathcal{P})$ and $\mathcal{R}(\mathcal{P}')$ are very different subsets of Ω, so $\mathcal{R}(S_{bel})$ can correspond to at most one of them at a time.

More generally: if we've established Counterfactual Uniqueness, then whenever *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's a $\mathcal{P}$ 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} have a unique interpretation in terms of partial beliefs, by virtue of the fact that their counterfactual profiles fit the profiles of the different belief states as specified by $\mathcal{D}$. (To be clear: given what we were told about Robo's design, it was expected that we'd find states that fit such a profile. 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. Furthermore, we need to show that there's no way to rejig the interpretation such that the S_{bel} and S_{des} states pick out basic desires and partial beliefs respectively. Both of these steps of the interpretation will require their own Counterfactual Uniqueness conditions. But at this stage it's 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}.

* * *

As a representation of how functionalists ought to approach the assignment of beliefs and desires to human beings, the metaphor is deficient in two important respects. We'll talk about those in a short moment. But first let me say a bit about Counterfactual Uniqueness, which is the linchpin of the idea. Why should we expect it to be true?

Obviously, a formal proof of Counterfactual Uniqueness would depend on the exact nature of $\mathcal{D}$, which I've intentionally left vague. I don't know the exact 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 $\mathcal{P}'$. (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 $\mathcal{P}'$ will assign distinct values to some proposition p, so now combine $\mathcal{P}$ and $\mathcal{P}'$ with some utility function, $\mathcal{U}$, such that $\mathcal{U}(a) = 1$, $\mathcal{U}(b) = 0$, and $\mathcal{U}(c) = \mathcal{P}(p)$. 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 $\mathcal{P}'$. We'll see the same sort of thing 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.

Of course, if Counterfactual Uniqueness does end up being false, then it will be possible to elaborate more on our partial beliefs' causal roles. Something like the considerations that go into Rationalisation* would help to further

narrow down the range of possible interpretations. In particular, for each S_{bel}, we define S^E_{bel} as the state Robo's *bel* compartment would be in, if he were to receive evidence E while in state S_{bel}. Each S^E_{bel} can then be paired with a region of Ω, and we'd be able to run more or less the same interpretive strategy given the following, strictly weaker premise:

Weak Counterfactual Uniqueness. If $\mathcal{P} \neq \mathcal{P}'$, then there is an E such that $\mathcal{R}(\mathcal{P}^E) \neq \mathcal{R}(\mathcal{P}'^E)$.

At this point, we have a very rich representation of partial belief's causal role in relation to preferences—one that builds in information about (a) how an agent's preferences would be expected to change given that she has such-and-such partial beliefs now and receives some new evidence E, and also (b) how her preferences would be expected to change if her basic desires were altered in various ways. To fix upon a unique interpretation even in hard cases like Zee and Zed above, it's not clear that we'd need any more information than that.

Nevertheless, we haven't taken into account the *other* causal roles that our partial beliefs play. That is the first respect in which the metaphor is deficient. Recall the discussions in §2 and §4: nothing about preference-centrism—as I've defined it or as it's often understood by its advocates—requires us to say that the belief-preference relationship is the *only* functional relationship that's relevant to giving an analysis of what partial beliefs are. We should expect that other kinds of causal roles will also be mentioned in our final analysis, even if they don't weigh very heavily when it comes to deciding matters of overall fit.

The second major deficiency in the metaphor is related. 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 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 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 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 exemplar of a human being *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 $\mathcal{D}$.

Finally, 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. But there are standard functionalist technologies for dealing with these kinds of cases (cf. Lewis 1980). 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. In this case, the natural thing to consider is the causal profile of his *bel* and *des* compartments, *were* he to be functioning properly. Likewise for 'mad belief': what matters is not so much what a person's beliefs do, but what they *ought* to do—what they do when everything is working well.

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 unique best deserver of the total causal role associated with $\mathcal{P}$, where the belief-preference connection—both actual and counterfactual—is given special importance in fixing what makes S a good enough deserver of that role. The result isn't behaviourist, instrumentalist, or anti-realist; it doesn't ignore other functional roles besides the belief-preference connection; and it doesn't apply only under idealised cases or given strong constraints on preferences.

§6. Conclusion

Of course, it's obvious that preference-centrism as a whole yet faces further worries and objections, and I don't expect my particular brand of functionalism to convince everyone. There's still significant work to be done in actually 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.

So, in closing I want to emphasise just how *distant* this is from the Entailment thesis. Even in the highly idealised metaphor, where we ignored all of the other causal roles that partial beliefs play, still the relation between partial beliefs and preferences isn't an entailment relation. Given that Robo is functioning perfectly, and operating in normal circumstances, then *maybe* the facts about the preferences he actually has might sometimes entail the facts about his partial beliefs—that is, *if* the relevant decision theory $\mathcal{D}$ admits of constraints under which Uniqueness holds.

At most, then, we might get something like Limited Entailment, if $\mathcal{D}$ has the right kind of character, and we restrict ourselves to the interpretation of perfectly built robots operating in normal circumstances. The human situation isn't going to be like Robo's, and in our case even Limited Entailment looks a pipe dream. The most we can expect are probabilistic correlations between certain patterns of preferences and certain partial belief states.

Nevertheless, the functionalism I've been describing is up to its ears in references to preferences: what it is to have such-and-such partial beliefs (and such-and-such basic desires) is to be characterised *primarily* through the preferences that those beliefs (and desires) do generate and would generate under different conditions. In that sense, even an entire monastery of Zen and non-denominational monks would present no deep concerns for account of partial belief that conceptually ties them to preferences, including accounts that tie them together very closely indeed.[9]

[9] Thanks to 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). This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 703959.

References

Anscombe, F. J. and R. J. Aumann (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, R. (2001). Ramsey and the Measurement of Belief. In D. Corfield and J. Williamson (Eds.), *Foundations of Bayesianism*, pp. 263–290. Kluwer Academic Publishers.

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

Bunge, M. (1973). On Confusing 'Measure' with 'Measurement' in the Methodology of Behavioral Science. In *The Methodological Unity of Science*, pp. 105–122. Dordrecht: D. Reidel Publishing.

Chalmers, D. (2002). Does Conceivability Entail Possibility? In T. Gendler and 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 University Press.

Cozic, M. and B. Hill (2015). Representation theorems and the semantics of decision-theoretic concepts. *Journal of Economic Methodology 22*(3), 292–311.

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. In *Problems of Rationality*, pp. 19–38. Oxford: Oxford University Press.

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

de Finetti, B. (1937). Foresight: its logical laws in subjective sources. In S. Kotz and N. L. Johnson (Eds.), *Breakthroughs in Statistics*, Volume 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. (MS). Comparativism and the Measurement of Partial Belief.

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

Eriksson, L. and W. Rabinowicz (2013). The interference problem for the betting interpretation of degrees of belief. *Synthese 190*, 809–830.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: John Wiley & Sons.

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., R. D. Luce, P. Suppes, and A. Tversky (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. Academic Press.

Lewis, D. (1970). How to Define Theoretical Terms. *The Journal of Philosophy 67*(13), 427–446.

Lewis, D. (1974). Radical interpretation. *Synthese 27*(3), 331–344.

Lewis, D. (1980). Mad Pain and Martian Pain. In N. Block (Ed.), *Readings in Philosophy of Psychology*, pp. 216–222. Harvard University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Luce, R. D. and P. C. Fishburn (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. and J. Weisberg (2011). Representation Theorems and the Foundations of Decision Theory. *Australasian Journal of Philosophy 89*(4), 641–663.

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach and S. Hurley (Eds.), *Foundations of Decision Theory: Issues and Advances*, pp. 147–175. Oxford: Basil Blackwater.

Ramsey, F. P. (1927). Facts and Propositions. *Proceedings of the Aristotelian Society 7*(1), 153–170.

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. and P. Wakker (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., M. J. Schervish, and J. B. Kadane (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. (2018). On the Ratio Challenge for Comparativism. *Australasian Journal of Philosophy 96*(2), 380–390.

Stefánsson, H. O. (2016). What Is Real in Probabilism? *Australasian Journal of Philosophy 97*(3), 573–587.

Suppes, P. and J. Zinnes (1963). Basic Measurement Theory. In D. R. Luce (Ed.), *Handbook of Mathematical Psychology*. John Wiley & Sons.

Tversky, A. and D. Kahneman (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.

Williams, J. R. G. (2016). Representational Scepticism: The Bubble Puzzle. *Philosophical Perspectives 30*, 419–442.

Williams, J. R. G. (2018). Normative Reference Magnets. *Philosophical Review 127*(1), 41–71.