Representation Theorems and Radical Interpretation

Edward Elliott*

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

July 26, 2020

Abstract

David Lewis' theory of radical interpretation for mental content is founded on two key ideas: that beliefs and desires can be understood in terms of their causal-functional roles within folk psychology, and that folk psychology is more or less Bayesian in outline. This paper concerns a puzzle for Lewis' Bayesian functionalism. On the one hand, Lewis argued that the facts about an agent's sensory evidence and choice dispositions will always underdetermine the facts about her beliefs and desires. On the other hand, we have various representation theorems (e.g., in Ramsey 1931; Savage 1954) that are widely taken to show that if an agent's choice dispositions satisfy certain structural conditions, then those dispositions *alone* can suffice to determine her beliefs and desires. Here, I will argue that Lewis' conclusion is correct—any tension with representation theorems is merely apparent, and relates primarily to the difference between how 'choice dispositions' are understood within Lewis' theory and the problematic way they're usually understood in the context of the representation theorems. Indeed, there's no plausible sense in which theorems like Ramsey's or Savage's show that beliefs and desires can be determined by choice dispositions, even in principle—ultimately, they're of very limited relevance to functionalism and to the project of radical interpretation.

Karl is an ordinary human being, with ordinary human beliefs and ordinary human desires. Our task is to work out what those beliefs and desires *are*. The catch is that we cannot help ourselves directly to any facts about Karl's inner mental life, or indeed about any mental states whatsoever, including our own. We are, however, allowed to know any and all facts about his physical constitution, environment, ancestry, possible futures, and counterfactual histories, that may be relevant—but only inasmuch as these are expressed without invoking any concepts that might raise question marks for physicalism. Given that and nothing more as our starting point, we're to derive the facts about Karl's beliefs and desires. Call this the project of *radical interpretation*.[1]

*E.J.R.Elliott@leeds.ac.uk. Comments welcome.

[1] In (Davidson 1973), (Lewis 1974), and elsewhere, radical interpretation also relates to linguistic meaning. We won't be interested in that aspect of the broader project here—sorting out Karl's beliefs and desires will be more than tricky enough.

The point of the project, of course, is not to describe how *we* go about interpreting one another in our everyday lives. Rather, it's to show how the facts about beliefs and desires fit in amongst the facts of a world that's fundamentally physical in nature. But maybe that's too much to ask for at this stage of inquiry. Any complete account of Karl's beliefs and desires will likely need to make some reference to the content of his senses, what he's able to imagine, his phenomenal states, and more besides—and that's a lot of question marks for anyone to deal with. So perhaps we can settle instead for something more realistic, a *not-quite-so-radical interpretation*: we should aim to take significant steps towards a complete physical explanation of the facts Karl's beliefs and desires, while *minimising* the number of question marks for later theorising to handle. If we have to cheat a little and presuppose access to the content of Karl's sensory evidence, for instance, then so be it—we can't solve everything at once, and we can chalk that up as a problem to be dealt with at some point down the road.

There's no shortage of views on how we might go about a project like this, but I happen to think that the strategy Lewis first put forward in 'Radical Interpretation' (1974) and developed in later works (e.g. 1979; 1980a; 1983a; 1983b; 1986; 1994) is *basically* on the right track. I'll say more about it in due course, but briefly for now: Lewis held that beliefs and desires are defined primarily by reference to the typical roles they play in relation to sensory evidence and choice dispositions according to folk psychology; and furthermore he held that folk psychology is more or less *Bayesian* in character—beliefs and desires come in degrees, choices are typically determined by something like expected utility maximisation, and learning works by something like conditionalisation.

From the beginning, though, Lewis held that without substantive 'eligibility' constraints on the contents of beliefs and desires, any definition of those states in terms of their causal-functional relationships with evidence and choice dispositions is doomed to failure. The clearest argument he gives for this is in 'New Work for a Theory of Universals' (1983a, pp. 373ff; see also 1986, pp. 36ff, 105ff). Assuming Karl is indeed an expected utility maximiser who conditionalises on his incoming sensory evidence, Lewis argues that whatever the facts about his life history of evidence and choice dispositions happen to be, there will always be many distinct belief-desire interpretations that *fit* those facts equally well. Hence, the evidence and choice facts radically underdetermine the belief and desire facts, and we need to say something more to pin down the *correct* interpretation. Lewis' solution was in effect to cut some systems of belief and desire from the running: only the most *eligible* systems are possible.

But now you might be wondering whether this 'eligibility' solution was really needed. After all, don't we have a number of representation theorems—such as those of Ramsey (1931) and Savage (1954), amongst many others—that are widely taken to show that if an agent's choice dispositions satisfy certain structural conditions $c_1, \ldots, c_n$, then there's a *unique* system of graded beliefs and desires under which they maximise expected utility? And if that's what these theorems really tell us, then if we again assume that Karl's an expected utility maximiser, it follows that *if* his choice dispositions satisfy the conditions $c_1, \ldots, c_n$, *then* his choice dispositions alone could in principle suffice to determine his beliefs and desires—we don't even need the evidence facts!

So we have what appears on the surface to be a conflict. In the one corner stands Lewis, who argues that there will always be radically distinct systems of belief and desire that equally fit the facts about Karl's evidence and choice

dispositions, whatever they may be, assuming that Karl is a rational Bayesian agent. And in the other corner are the representation theorems, which are widely taken to imply that in the right conditions the facts about Karl's choice dispositions alone might be enough to determine his beliefs and desires, under the very same assumption. Lewis' argument is compelling, or so I'll argue. But there can be no reasonable doubt that these theorems express mathematical truths. So what are we to make of all this?

In this paper, I'll argue for two main conclusions. First: there's no genuine conflict between Lewis' underdetermination argument and representation theorems like Ramsey's and Savage's. I'm not the first to suggest this, but the manner of reconciliation I propose is new. Where others have suggested that Lewis probably denied some of the conditions $c_1, \ldots, c_n$ used by Ramsey and Savage to derive a unique interpretation from choice dispositions (e.g., Schwarz 2014; Williams 2016), I think the issue goes deeper than that. The key to reconciliation, I'll argue, lies in the difference between how 'choice dispositions' are understood in the context of Lewis' theory, and the (problematic) way they're usually treated in the context of the representation theorems.

And this leads to my second conclusion: on any plausible view about how beliefs and desires might be related to choice dispositions, representation theorems like Ramsey's and Savage's simply do not tell us how the former are determined by the latter (i.e., assuming the agent maximises expected utility). A closer look at how we understand 'choice dispositions', and how these relate to beliefs and desires, shows that the representation theorems don't have the right kind of *structure* to tell us anything much about how beliefs and desires relate to choices. Ultimately, these theorems are of no significant relevance to the problem of radical indeterminacy.

The remainder of the paper comes in three parts. The first is a description of Lewis' Bayesian functionalism (§1). The second describes Lewis' underdetermination argument (§2). The third concerns representation theorems: how to reconcile them with Lewis' argument, and why they're of little relevance to the project of (not-quite-so-)radical interpretation (§3).

1. Bayesian Functionalism

I'll start with some 'big picture' matters on analytic functionalism and Bayesianism (§1.1), after which I'll describe the what Lewis' Bayesian functionalist theory looks like (§1.2–§1.3).

1.1 The big picture

Any functionalist account of belief and desire will take as its starting point some psychological theory, call it $\mathcal{T}$, which tells us:

a) What kinds of contents are *eligible* to be believed and/or desired—i.e., a way of indexing the possible belief and desire states
b) What beliefs and desires *do*—i.e., how beliefs and desires relate to one another, other mental states, and the world (e.g., via the senses and behaviour)

For instance, a very simplistic theory might say that (a) for every proposition p there's a *belief that* p, and (b) a belief that p arises from evidence that p and

causes behaviours appropriate to worlds where p is true. Change either part and you've got a different theory.

Let a *scheme of interpretation* be a way of identifying momentary physical states $S, S', \ldots$ (e.g., brain states) with the specific belief and desire states $M, M', \ldots$ that are possible according to $\mathcal{T}$. Say next that a scheme of interpretation *fits* to the extent that, if it identifies $M, M', \ldots$ with $S, S', \ldots$ respectively, then the $S, S', \ldots$ behave in relation to one another, other mental states, and to the external world in the same way that their corresponding interpretations $M, M', \ldots$ behave in relation to one another, other mental states, and to the external world according to $\mathcal{T}$. According to functionalism, the *correct* scheme of interpretation will be one that maximises fit.

If more than one scheme of interpretation maximises fit, then the usual thing is to say that the correct scheme is *indeterminate* between them (e.g., Lewis 1983b, p. 120). Where the resulting degree of indeterminacy is small, then this raises no special issues. After all, there's presumably some fuzziness in the content of one's beliefs and desires at a given time, and it would be strange if our theory entailed there isn't. A problem arises, however, if there are multiple *radically* different schemes tied for equal best. This is the issue Lewis thought we'd face if the eligibility conditions imposed on the contents of belief and desire aren't strong enough. But we'll come back to that. For now, let me say more about the specific varieties of functionalism Lewis preferred.

Given the above schematic characterisation of functionalism, we can define a number of different types by reference to characteristics of the underlying theory $\mathcal{T}$. The first type is one of Lewis' core commitments throughout his career:

Analytic Functionalism. $\mathcal{T}$ is a systematisation of folk psychology

The 'systematisation' is important: to whatever extent we have a shared implicit understanding of intentional psychology, this is likely be at least a little messy, perhaps incomplete, and maybe even inconsistent. So what we're *really* after will be a systematic reconstruction of folk psychological thinking, with the holes filled in, the inconsistencies smoothed over, and the messiness tidied up—i.e., the idea is to define beliefs and desires in terms of their functional roles within the best systematisation of folk psychology properly so-called. (Compare Jackson 1998, ch. 5, on functionally defining moral concepts using hypothetical systematisations of the folk's implicit moral theories.)

Next, we have a style of functionalism that Lewis advocated from at least (1980a) onwards (see also 1983b; 1986, pp. 39–40):

Anti-Individualist Functionalism. $\mathcal{T}$ concerns typical agents

The functional roles we extract from a systematised folk psychology are to be understood as *typical* roles. The theory $\mathcal{T}$ places no specific constraints on how *Karl's* beliefs and desires will actually behave, but rather on how we'd expect his beliefs and desires to behave *if* he were perfectly typical and acting under normal conditions. A scheme of interpretation on this picture assigns mental state interpretations to physical state types on the basis of how those states *typically* behave, and Karl is said to be in the mental state M just in case he happens to be in a physical state S that's assigned M by the correct scheme of interpretation—i.e., regardless of whether *for Karl* S happens to behave anything at all like M is supposed to according to $\mathcal{T}$.

(In Lewis' work, 'typicality' is usually characterised in statistical terms: roughly, S might be interpreted as a *belief that p*, for example, because across the many actual and possible individuals in which it recurs, it tends to be caused by the kinds of things that cause a belief that p, and it tends to cause behaviour that fits with believing that p. See, e.g., (Lewis 1980a; 1983b; 1986, pp. 39–40). It's better, I think, to understand 'typicality' in terms of archetypes. The core principles of folk psychology tell us what our beliefs and desires are *supposed* to be like when everything's functioning properly in normal conditions. The agent who instantiates the folk psychological theory of decision-making will therefore be more like the Platonic ideal of a decision-maker than the average decision-maker. But I'm going to leave the relevant sense of 'typical' ambiguous for the discussion that follows, since very little of what I have to say will depend on it.)

The final type of functionalism we'll define is:

Bayesian Functionalism. $\mathcal{T}$ is recognisably Bayesian in character

That is, $\mathcal{T}$'s a theory wherein, in one form or another, beliefs and desires come in degrees, they're more or less coherent according to something like the Bayesian standard, choices are determined by something at least similar to expected utility maximisation, and learning works by something similar to conditionalisation.

A reason you might have for being a Bayesian functionalist would be if you thought that our best scientific theories of learning and decision-making are more or less Bayesian in character. I think that's plausible, but for this paper it's neither here nor there. In Lewis' case, Bayesian functionalism is a consequence of his commitment to analytic functionalism, under the assumption that any good systematisation of folk psychology 'should look a lot like Bayesian decision theory' (1979, pp. 533–4). As he put it in 'Radical Interpretation',

> [Bayesian] decision theory (at least, if we omit the frills) is not esoteric science, however unfamiliar it may seem to an outsider. Rather, it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference, and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematised. (1974, pp. 337–8; cf. p. 335)

Not exactly like the orthodox Bayesian theory, mind you, but not too far from it either. Nevertheless, I expect that this will strike many readers as surprising. "Folk psychology is *nothing at all* like Bayesianism!", I hear them cry. But I'm with Lewis on this, and since the matter is likely to be controversial, it's worth pausing to consider some points in its defence of the idea before moving on.

First, you might worry that folk psychology cannot be much like Bayesianism, since the folk themselves will often struggle to comprehend Bayesian theories. But the tenets of our systematised folk psychological theory need not be constituted from statements that members of the folk would themselves spontaneously *assert*, or even *readily recognise* as things that they themselves tacitly believe. Despite what Lewis said at points earlier in his career, we don't have to see folk psychology as a simple collection of 'platitudes' gathered from the folk themselves. Our understanding of intentional psychology can instead be likened to our tacit understanding of grammar, and as such can involve sophisticated principles clearly expressible only using technical machinery with which the folk may be unfamiliar (cf. Lewis 1994, p. 416; Jackson and Pettit 1990, pp. 33–6).

Second, you might worry that Bayesian theories tend to be unrealistically demanding—more so than we usually expect of ourselves or others. Part of the problem here comes down to the many idealisations present in standard Bayesian models, and Lewis clearly thought the more extreme of these would need to be removed from any final systematisation of folk psychology (cf. 1983a, p. 375; 1986, p. 30; 1994, p. 428). Fortunately, we know that some idealisations can indeed be weakened without fundamentally changing the recognisably *Bayesian* character of the resulting theory (e.g., Jeffrey 1983; Walley 1991; Weirich 2004; Bradley **?**). Furthermore, not all idealisations are inherently problematic: any general psychological theory will sacrifice some realism for greater generality and overall simplicity, and there's no reason to think the best systematisation of folk psychology will be any different. We can live with a touch of unrealism in our folk psychological theory.

More importantly, our's is supposed to be a theory of the *typical* agent. Since the majority of people will be atypical to some degree in some relevant respect at least some of the time, we should expect that we will fail to conform folk psychological expectations—often to a small degree, occasionally to a much larger degree. This applies for either sense of 'typical' noted above, though all the more so if we understand the typical agent as an archetypal ideal rather than a statistically average agent. Pobody's nerfect.

Finally, it's no requirement of analytic functionalism that our systematised folk psychology will end up being *true*, even *qua* theory of the typical agent. We do best to think of folk psychology as based around a highly simplified model of the major psychological difference-makers in relation to evidence and behaviour for typical agents in normal conditions (cf. Godfrey-Smith 2005; Maibom 2009; Lewis 1980a; 1983b; 1994, p. 416). Whatever the truth happens to be, it will be much messier than what we find in folk psychology—so if anything occupies the folk functional roles of belief and desire then it will be imperfectly. But this is just to say that we can expect folk psychology to generate false predictions even for typical cases. The analytic functionalist can (and I think should) accept that folk psychology is probably false, that there are widespread psychological phenomena that it fails to predict and explain, and that newer scientific theories designed to accommodate the empirical data might well be more accurate. These are predictions of analytic functionalism, not problems for the view.

To summarise, then, the anti-individualist analytic functionalist proposes that our folk notions of belief and desire will pick out those physical states, if any there are, which come close enough (and closer than anything else) to satisfying the roles associated with the doxastic and conative states posited by the best systematisation of the general principles underlying our shared implicit understanding of the typical agents' intentional psychology. We should expect this theory to simplify and idealise over the messy details, be rife with *ceteris paribus* clauses, and we should *not* expect the folk themselves to easily recognise it as a systematisation of their own implicit beliefs. It probably won't be the empirically most accurate theory we can come up with, but it will on average do a fairly good job at least for perfectly typical cases. And like Lewis, I think that an essentially Bayesian theory will fits the bill for systematising folk psychology quite nicely—at least 'if we omit the frills'.

1.2 The typical Bayesian

I'll now describe the typical functional roles of beliefs and desires on a relatively simplified Bayesian psychological theory. Let me first introduce *Typikarl*. We're to assume Typikarl is perfectly typical with respect to the structure of his beliefs and desires at a time; the relationship between his beliefs, desires, and choice dispositions, and how he changes over time. Karl is not Typikarl. Like the rest of us, Karl will be atypical in at least one of the above respects to at least to some degree at least some of the time—but not so our Typikarl, whose most unusual characteristic by far is that he is, always has been and always will be so uncompromisingly *typical*. Typikarl probably doesn't exist, but he's a useful myth for describing the typical functional roles of beliefs and desires.

Let me also flag from the outset that the following is going to be based on a 'Jeffreyan' conception of decision theory (see Jeffrey 1965).[2] By this I mean:

1. Propositions are subsets of a fixed set of possible worlds, Ω
2. Deliberation does not crowd out prediction: an agent has desires and preferences regarding a proposition p only if she has beliefs regarding p
3. An agent's desirability for p is her *conditional expected desirability* for p

In relation to 1, on Lewis' view Ω should be a set of *doxastic alternatives*, or *centred possible worlds* (see Lewis 1979; 1986, pp. 28–9). However, for the sake of simplicity I'll be ignoring complications arising from the egocentricity of content since they're orthogonal to the issues I want to raise.[3] For the same reason, I'll assume throughout that Ω is finite. Nothing of importance hangs on this.

Beliefs at a time

We'll start with Typikarl's beliefs at a time. Any vaguely Bayesian theory will require at minimum that beliefs are in some sense *graded*, and furthermore that they're coherent enough to permit the comparison of expected desirabilities (Lewis 1974, p. 337). The most familiar way to ensure this is to assume that Typikarl's beliefs can be represented by a probability function.

I don't want to assume that agents have beliefs regarding every proposition. (Lewis too seems to have rejected this assumption; see, e.g., 1986, pp. 104ff.) So, instead we'll say that Typikarl believes to some degree every proposition of which he has the conceptual resources to entertain. We'll assume at a minimum that if Typikarl's able to entertain p then he's able to entertain its negation; and if he's able to entertain p and q, then he's able to entertain their conjunction ($p \cap q$) and their disjunction ($p \cup q$). Furthermore, we'll say that a proposition a is an

[2] An exegetical question arises at the point: when discussing Lewis' view, should we use evidential decision theory (EDT) or causal decision theory (CDT)? I don't know. The main argument for using CDT is that it was Lewis' (1981) preferred account of rational decision-making. However, the correct theory of pragmatic rationality needn't be the best systematisation of folk psychology. On Lewis' view, the folk functional roles for belief and desire *just are* rational roles (1986, p. 36; 1994, p. 428), but Lewis only took folk psychology to be a theory of 'imperfect, near-enough rationality' (1994, p. 428); i.e., the folk could be systematically *wrong* about what's rational. Moreover, Lewis himself uses EDT in the description of his view (e.g., 1983a, p. 374), and the underdetermination argument presupposes it. I've chosen the Jeffreyan conception for my exposition because it's how Lewis described his own view, and it makes some of the technicalities easier to deal with. See §2.2 for related discussion.

[3] Ignoring egocentricity does have some strange consequences—e.g., that Typikarl always knows what time it is. Just pretend that typical agents always wear a watch.

atom whenever Typikarl has beliefs regarding a but doesn't have beliefs for any distinct non-empty propositions stronger than a. The set of atoms $\mathbf{A}$ will be a partition of Ω containing all and only the strongest propositions that Typikarl's able to entertain (including but not limited to the impossible proposition $\varnothing$). Furthermore, where $\mathcal{A}(\mathbf{A})$ is *the algebra generated by* $\mathbf{A}$,

$$\mathcal{A}(\mathbf{A}) = \{p \subseteq \Omega : p = a_1 \cup \cdots \cup a_n, \text{ where } a_1, ..., a_n \in \mathbf{A}\},$$

then $\mathcal{A}(\mathbf{A})$ is the set of propositions for which Typikarl has beliefs.

Given that, our first principle is an eligibility constraint:

$\mathcal{B}$-Eligibility. Typikarl's beliefs at any time can be represented by a function $\mathcal{B} : \mathcal{A}(\mathbf{A}) \mapsto \mathbb{R}$ such that for all p, q in $\mathcal{A}(\mathbf{A})$,

i) $\mathcal{B}(p) \geq 0$
ii) $\mathcal{B}(\Omega) = 1$
iii) if $p \cap q = \varnothing$, then $\mathcal{B}(p \cup q) = \mathcal{B}(p) + \mathcal{B}(q)$

Note that we're taking the probability function to capture the *content* of a system of beliefs as a whole. We are thus understanding 'content' somewhat more broadly than it's sometimes used—the 'content' of a system of beliefs is not a proposition, but a space of propositions and a measure of the strengths with which each is believed.

Lewis assumed $\mathcal{B}$-Eligibility frequently when characterising systems of belief (e.g., 1974, pp. 337–8; 1979, p. 534; 1980b, pp. 287–8; 1981, p. 7; 1983a, p. 374; 1986, p. 30; 1996). He recognised of course that it's overly idealising—for instance, it ignores 'limitations of logical competence' (1981, p. 7; 1983a, p. 375); it does not make room for the possibility of 'fragmented' systems of belief better represented by multiple probability functions at once (1986, pp. 29–32); and the real values of probability functions make divisions which appear sometimes too precise, and sometimes not precise enough, for the beliefs we want to represent (1986, p. 30; 1994, p. 428). With that said, Lewis clearly thought that probability functions provide a reasonably good approximate representation of beliefs for most purposes, and I'm inclined to spot him this without further argument.

Desires at a time

Next up are Typikarl's desires at a time. Although he never explicitly discussed the matter as such, Lewis often described our desires as coming in two different kinds: there are 'basic desires' or 'intrinsic values', and there are 'desires' *simpliciter*. (See, e.g., 1974, p. 336; 1983a, p. 374). We can think of an agent's intrinsic values at a time as that aspect of her overall system of desires that's independent of her beliefs at that time, while her desires will capture that aspect that's derived from her intrinsic values plus her beliefs.

Within the Jeffreyan framework, intrinsic values will be most easily represented by a distribution of real-valued *desirabilities* over the atoms in $\mathbf{A}$:

$\mathcal{V}$-Eligibility. Typikarl's intrinsic values at any time can be represented by a function $\mathcal{V} : \mathbf{A} \mapsto \mathbb{R}$

$\mathcal{V}$-Eligibility is our second basic eligibility condition—and it's not a trivial one. You might usefully think of $\mathcal{V}$ as representing the relative strength with

which TypiKarl desires that the actual world is an a-world, for each atom a. In that case it implies that TypiKarl's desirability ranking over atoms is transitive and complete. Officially, though, we commit to nothing more than $\mathcal{V}$ being a *representation* of TypiKarl's intrinsic values. This representation might be induced by a psychologically more fundamental ranking of salient properties TypiKarl cares about (see, e.g., Dietrich and List 2013).

As is standard, I'll assume that strengths of desire are measurable on nothing stronger than an interval scale. For the remainder of this paper if two desirability functions $\mathcal{V}$ and $\mathcal{V}'$ are related by an interval-preserving transformation—i.e., there's an $\mathrm{x} > 0$ and a y such that for all atoms a, $\mathcal{V}(a) = \mathcal{V}'(a)\mathrm{x} + \mathrm{y}$—then I'll presume they represent the very same values, and I won't be fussed about distinguishing between them.

We can now talk about Typikarl's desires, which will be a function of his beliefs and values. Say first that $\mathcal{B}^p$ designates $\mathcal{B}$ *conditionalised on* p; that is, for all p, q,

$$\mathcal{B}^p(q) \;=\; \mathcal{B}(q \,|\, p) \;=\; \frac{\mathcal{B}(q \cap p)}{\mathcal{B}(p)},$$

if $\mathcal{B}(p) > 0$; otherwise $\mathcal{B}^p(q)$ is undefined. In our Jeffreyan framework we can then say that the strength with which Typikarl desires p is just the $\mathcal{B}^p$-weighted average $\mathcal{V}$-desirability of the atoms that constitute p. However, all we *really* need to talk about are relative desires—that is, *preferences*. So, if we say that $p \succsim q$ just in case Typikarl desires p at least as much as he desires q, then:

$\succsim$-Coherence. If Typikarl has beliefs $\mathcal{B}$ and intrinsic values $\mathcal{V}$, then

$$p \succsim q \;\text{ iff }\; \sum_{a \in \mathbf{A}} \mathcal{B}^p(a)\mathcal{V}(a) \;\geq\; \sum_{a \in \mathbf{A}} \mathcal{B}^q(a)\mathcal{V}(a),$$

Choice dispositions

Since Typikarl's beliefs $\mathcal{B}$ and intrinsic values $\mathcal{V}$ at a time jointly determine his total system of desires at that time, from now on we'll represent Typikarl's *total belief-desire state* using pairs of the form $(\mathcal{B}, \mathcal{V})$. Our next task is to describe the sense in which Typikarl's *choice dispositions* maximise expected desirability.

If Typikarl were able to make any proposition whatsoever true, and he knew this, then he'd make it so the actual world belongs to an atom he values most. But Typikarl does not have impossible powers, and probably has no direct influence over the truth or falsity of the vast majority of propositions he's able to contemplate. So, before we can say anything specific about Typikarl's choice dispositions, we first need to fully characterise his *options*—and that's something I'm *not* going to attempt, since the matter raises tricky issues that would swallow this paper whole long before we get around to saying anything about underdetermination. A few general remarks will need to suffice.

To start with, we'll follow Lewis (1974, p. 337; 1994, pp. 416–7) in treating Typikarl's options at any time as a partition of propositions specifying how he

behaves at that time.[4] Given this, suppose we let

$$\boldsymbol{\tau} = (\tau_1, ..., \tau_n)$$

be an ordered set of (mutually excluding, jointly exhaustive) times from the beginning of Typikarl's life onwards. Then, an initial thought for how we might characterise Typikarl's options at time τ would be to use the partition

$$\mathbf{B}^{\tau} = \{b_1, ..., b_n\},$$

that captures in maximal detail each the specific ways of Typikarl behaves at τ. But this won't work, at least not necessarily, since Typikarl needn't have preferences over any of the b in $\mathbf{B}^{\tau}$. More generally, we've yet to take into consideration any relationships between Typikarl's options and his beliefs.

In the most basic version of Jeffrey's decision theory, for example, a proposition counts as an option only if it's 'actual':

> … we might call a proposition *actual* for an agent at a time if at that time he can perform an act the *direct* effect of which will be that his degree of belief in the proposition will [rationally] change to 0 or 1. Under ordinary circumstances … the proposition that *the agent blows his nose* is actual. (Jeffrey 1968, p. 170)

Something similar has been advocated by Sobel (e.g., 1983; 1986)—roughly, p's an option only if the decision-maker is certain she can make p true by an act of will—and related conditions are discussed in (Hedden 2012) and (Schwarz forthcoming). I don't know whether we should want anything this strong for our systematisation of folk psychology, but here's not the place to decide the issue. What's important is that there are plausibly *some* constraints on what counts as an option relating to what Typikarl *believes* he can choose.

I'll assume at the least that Typikarl won't be disposed to choose p unless he assigns p some positive degree of belief—he's not going to choose something that's impossible by his lights. Furthermore, any proposition about his behaviour that he *does* consider possible at τ is going to be equivalent to a disjunction of the more specific behaviour-propositions in $\mathbf{B}^{\tau}$. Consequently, we characterise *the options from Typikarl's perspective at* τ as a maximally fine-grained way of grouping together disjunctions of $b_i \in \mathbf{B}^{\tau}$,

$$\underline{\mathbf{B}}^{\tau} = \{\underline{b}_{\varnothing}, \underline{b}_1, ..., \underline{b}_n\} \subset \mathcal{A}(\mathbf{A}),$$

such that if Typikarl has beliefs $\mathcal{B}$ at τ,

1. the *null* option, $\underline{b}_{\varnothing}$, is the largest disjunction of $b_i \in \mathbf{B}^{\tau}$ so that $\mathcal{B}(\underline{b}_{\varnothing}) = 0$,
2. the *non-null* options, $\underline{b}_i \in \underline{\mathbf{B}}^{\tau} \setminus \{\underline{b}_{\varnothing}\}$, each satisfy $\mathcal{B}(\underline{b}_i) > 0$,

plus whatever further doxastic conditions we might choose to place on the non-null options—e.g., that $\underline{b}_i \in \underline{\mathbf{B}}^{\tau} \setminus \{\underline{b}_{\varnothing}\}$ only if Typikarl is certain that if he

[4] Arguably, agents choose *actions*, not *behaviours*. If so, then options should consist in a set of action-specifying propositions. Lewis occasionally described options as actions (e.g., 1981; 1980b, p. 288), but in his more careful moments he prefers the behavioural characterisation. These matters are orthogonal to the points I want to raise, however, and it would make no significant difference to the discussion if we reformulate options as actions.

were to choose $\underline{b}_i$, then $\underline{b}_i$. In brief: Typikarl's options will be (equivalent to) disjunctions of maximally specific propositions about his behaviour that satisfy some appropriate doxastic conditions, whatever those end up being.

I'll assume henceforth that there's always a unique $\underline{\mathbf{B}}^\tau$ that characterises the options from Typikarl's perspective at τ. This uniqueness assumption is implicit in some of Lewis' writings (e.g., 1981, p. 7: '… this is *the* partition of the agent's alternative options'), though it's not clear whether it will be true in general. But the assumption plays no critical role in the underdetermination argument, so I'm not especially worried about taking it for granted here. Indeed, if there can be multiple non-equivalent ways to characterise Typikarl's options at a time then we'd almost certainly expect to see *more* underdetermination than we would if there's always a unique $\underline{\mathbf{B}}^\tau$.

There's one final complication, and it's one Lewis seems to have neglected: Typikarl can presumably be wrong about what he can choose to make true. He might, say, believe he can *raise his hand* without realising that his arm has fallen asleep. So say that $\underline{b}$ is *available* just in case, were Typikarl to decide to make $\underline{b}$ true, he would make it true. Not every option from Typikarl's perspective need be available given the actual facts of the matter, and presumably when Typikarl tries and fails to raise his hand he updates his beliefs in light of the new evidence and re-evaluates his options. Before he updates, though, it's reasonable to say that if Typikarl considers $\underline{b}$ uniquely best amongst the options from his perspective, then he's disposed such that if $\underline{b}$ *were* available, then he would make it true—e.g., if he arm hadn't fallen asleep, his hand would have been raised.

We now have everything we need. Relative to a system of beliefs and desires, define the *choice function* $\mathcal{C}$ for all non-empty $\mathbf{P} \subseteq \mathcal{A}(\mathbf{A})$ as:

$$\mathcal{C}(\mathbf{P}) = \{p \in \mathbf{P} : p \succsim q, \text{ for all } q \in \mathbf{P} \text{ such that } \mathcal{B}(q) > 0\}$$

The choice function picks out whichever propositions within a set are maximally desirable relative to $(\mathcal{B}, \mathcal{V})$; thus,

$\succsim$-Maximisation. If Typikarl is conscious and has beliefs and desires $(\mathcal{B}, \mathcal{V})$ at τ, then he's disposed such that if every $\underline{b} \in \mathcal{C}(\underline{\mathbf{B}}^\tau)$ were available, he would make one of them true

As a matter of fact, Typikarl will make some $b \in \mathbf{B}^\tau$ true; and if he's not mistaken about what's available then this b will entail some $\underline{b}$ in $\mathcal{C}(\underline{\mathbf{B}}^\tau)$. But since he may (or may not) choose randomly from amongst the options he considers best, we know only that at least one of the $\underline{b}$ in $\mathcal{C}(\underline{\mathbf{B}}^\tau)$ will be made true inasmuch as Typikarl isn't mistaken about his available options.

Changes over time

$\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, and $\succsim$-Coherence tell us what Typikarl's systems of belief and desire will be like at any given time, and $\succsim$-Maximisation tells us how his beliefs and desires together explain his behavioural dispositions. What remains to be characterised is how Typikarl changes over time.

We'll assume, as Lewis usually did, that Typikarl updates by conditionalisation (cf. Lewis 1980b, p. 288; 1983a, p. 374; 1994, pp. 428–9). To make this precise, let me introduce some more notation and a couple of background assumptions. First, we'll let

$$\mathbf{E} = \{e_1, ..., e_n\}$$

designate the set containing those propositions which characterise—in all relevant detail—all possible 'streams' of sensory evidence from any time τ to any other time τ'; e.g., *there is at* τ_1 *the appearance of such-and-such shapes and colours along with such-and-such sounds and such-and-such smells, etc., and then at* τ_2 *the appearance of*…, and so on. Arguably, we cannot know the content of Typikarl's sensory evidence merely given the physical facts, not least because (a) what we perceive seems to depend in part on our expectations, and (b) the content may rest on phenomenological facts (cf. Pautz 2013). But we'll help ourselves to the evidence facts, noting the question mark for later.

We'll assume $\mathbf{E} \subseteq \mathcal{A}(\mathbf{A})$; and if we say that $\mathcal{B}$ is *consistent with* e just in case $\mathcal{B}(e) > 0$, then we'll assume that Typikarl's *initial* system of beliefs is consistent with every $e \in \mathbf{E}$—Typikarl doesn't rule out the possibility of any life history of sensory evidence prior to having had any sensory experiences whatsoever. Finally, and purely for the sake of technical convenience later on, we'll assume that $\mathbf{E}$ includes Ω—the 'trivial' evidence.

With all that in place, the main driving force for psychological change is:

Conditionalisation. Typikarl's beliefs at any time τ are given by $\mathcal{B}_i^e$, where $\mathcal{B}_i$ is his *initial* system of beliefs and e characterises his evidence up to τ

No doubt this over-idealises. The perfect conditionaliser always knows exactly what their evidence is, and never forgets. But I'll emphasise again that it's only supposed to be approximately true in perfectly typical cases. Typikarl is better *qua* agent than we are, so we can hold him to a higher standard.

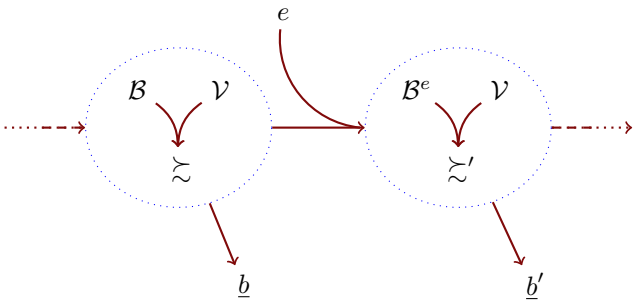
The only thing left to describe is how Typikarl's desires might change over time. Since Typikarl's beliefs will generally change as a result of incoming sensory evidence, so too will his preferences tend to shift and change about as time goes on. On the other hand, an assumption that I'll be making for now is:

Static Values. Typikarl's intrinsic values do not change over time

This assumption is implicit in a number of Lewis' works. (See 1974, pp. 336–7; 1980b, p. 288; 1983a, pp. 374–5.) In his (1983a, p. 375), Lewis claimed that it was a 'dire' over-simplification. By this I suspect he meant that commonsense psychology allows that an agent's intrinsic values *can* change, not that they *will*, and not that they will *often*. In any case, we'll hold onto the assumption for now and discuss the implications denying it in §2.2.

That's all folks

In summary: Typikarl's beliefs and intrinsic values place constraints on his desires, and his total system of beliefs and desires at τ determines his behavioural dispositions at τ. Each system of beliefs and desires is poised to give rise to a range of new systems of belief and desire by conditionalising on sensory input, and each such system might itself come about from some earlier systems conditionalised on the appropriate (possibly trivial) evidence.



Is there anything more to consider? In particular, what should we say about the functional relationships that beliefs and desires share with other folk psychological attitudes like *hopes* and *fears*? These are emphasised by some authors who adopt a broadly functionalist line (e.g., Christensen 2001, p. 361; Eriksson & Hájek 2007, p. 208), but Lewis apparently had no settled views on the matter:

> I think it an open question to what extent other states with content—doubting, wondering, fearing, pretending, …—require separate treatment, and to what extent they can be reduced to patterns in belief and desire and contentless feeling. Be that as it may, I shall ignore them here. (Lewis 1994, p. 421)

Given the consistency with which he ignored these other states, Lewis evidently thought that according to the folk an agent's beliefs and desires will typically be the *main* explanatory considerations in relation to her behaviour given her evidence. And I'm inclined to agree. If the choice and evidence roles don't exhaust the total functional role of beliefs and desires, they at least capture the most important parts when it comes to deciding matters of fit (cf. Elliott 2019).

1.3 Fitness

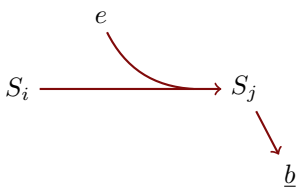
The above is a fleshed out and lightly generalised version of the psychological theory that Lewis describes at the end of 'New Work'. The only thing left is to say what a scheme of interpretation is and what it is for a scheme to maximise *fit*. As we'll see in §2.2, there are some subtle issues in relation to the underdetermination argument that rest on how 'schemes of interpretation' are being understood, so it's worth looking at this in a bit more detail.

Here's what Lewis has to say:

> $\mathcal{B}$ is … regarded as encapsulating the subject's dispositions to form beliefs under the impact of sensory evidence: if a stream of evidence specified by proposition e would put the subject into a total state S—for short, e *yields* S—we interpret S to consist in part of the belief system given by the probability distribution $\mathcal{B}^e$ […] $\mathcal{V}$ is a function from worlds to numerical desirability scores, regarded as encapsulating the subject's intrinsic values: if e yields S, we interpret S to consist in part of the system of desires given by the $\mathcal{B}^e$-expectations of $\mathcal{V}$. Say that $\mathcal{B}$ and $\mathcal{V}$ *rationalise* behaviour b after evidence e iff the system of desires given by the $\mathcal{B}^e$-expectations of $\mathcal{V}$ ranks b at least as high as any alternative behaviour. Say that $\mathcal{B}$ and $\mathcal{V}$ fit iff,

> for any evidence-specifying e, e yields a state S that would cause behaviour rationalised by $\mathcal{B}$ and $\mathcal{V}$ after e. That is our only constraining principle of fit. (Where did the others go? – We built them into the definitions whereby $\mathcal{B}$ and $\mathcal{V}$ encapsulate an assignment of content to states.) (1983a, p. 374, notation altered for consistency)

There's a lot here, and not all of it is especially clear. So, to help spell it all out I'll start with something that we'll call *matching*: a total momentary physical state S_i *matches* an interpretation $(\mathcal{B}, \mathcal{V})$ just in case, for any e consistent with $\mathcal{B}$, the typical agent in S_i given evidence e will come to be in some state S_j that disposes them such that if $\underline{b}$ true, where $\underline{b}$ and maximises expected desirability with respect to $(\mathcal{B}^e, \mathcal{V})$ (assuming $\underline{b}$ is available). In visual form, the total momentary physical state S_i matches $(\mathcal{B}, \mathcal{V})$ whenever, for all relevant evidence-specifying e,



where $\underline{b}$ is available and maximises expected desirability with respect to $(\mathcal{B}^e, \mathcal{V})$.

Since we've assumed that $\mathbf{E}$ includes Ω, S_i will match $(\mathcal{B}, \mathcal{V})$ only if S_i itself gives rise to behaviour that maximises expected desirability with respect to $(\mathcal{B}, \mathcal{V})$. However, to say that S_i *matches* $(\mathcal{B}, \mathcal{V})$ does not imply that S_i will be assigned $(\mathcal{B}, \mathcal{V})$ by the most *fitting* scheme(s) of interpretation. For instance, S_i might match $(\mathcal{B}, \mathcal{V})$ because given e it yields S_j which produces behaviour that maximises expected desirability with respect to $(\mathcal{B}^e, \mathcal{V})$, but for that implies S_j itself might not match $(\mathcal{B}^e, \mathcal{V})$. Nor would it be enough to just say that a fitting scheme of interpretation will be one where every state is assigned an interpretation that it matches, since a state might match more than one system of beliefs and desires. We need instead that the interpretations are *aligned* in the appropriate way in light of the causal relationships between the states to which they're assigned. This I take it is the intended upshot of 'the definitions whereby $\mathcal{B}$ and $\mathcal{V}$ encapsulate an assignment of content to states'.

To make this precise, let a *coarse-grained scheme of interpretation* be any function $\mathcal{I}$ from a set of (mutually exclusive and jointly exhaustive) types of *total* momentary physical states to *total* systems of belief and desire $(\mathcal{B}, \mathcal{V})$. This is how Lewis briefly characterises schemes of interpretation in (1983b, p. 119), and it's evident from how Lewis defines his 'only constraining principle of fit' (1983a, p. 374) that he was thinking about them this way in 'New Work' as well. Given this, Lewis' 'only constraining principle of fit' in conjunction with 'the definitions whereby $\mathcal{B}$ and $\mathcal{V}$ encapsulate an assignment of content to states' amounts to:

Fitness (Lewis). A scheme $\mathcal{I}$ *fits* iff, if $\mathcal{I}(S_i) = (\mathcal{B}, \mathcal{V})$, then

i) S_i matches $(\mathcal{B}, \mathcal{V})$
ii) for any e consistent with $\mathcal{B}$, if S_i given e yields S_j, then $\mathcal{I}(S_j) = (\mathcal{B}^e, \mathcal{V})$

Now Lewis seems to have made a small error here. The definition of fitness essentially says that if you're going to interpret a total physical state S as a certain belief-desire state $(\mathcal{B}, \mathcal{V})$, then you'd better make sure that any total physical states that are causally downstream of S can also be given appropriate interpretations in line with that initial assignment of $(\mathcal{B}, \mathcal{V})$ to S. And that's *almost* enough—or rather, it *would* be enough, *if* we were allowed to assume a priori that our total physical states will slot neatly into the network of causal relations between whole belief-desire states that Lewis was imagining they have. But it is at least conceptually possible to have a sequence of causally-related physical states that *start off* behaving in a very non-Bayesian way, and then at some point down the track start behaving in a Bayesian way and continue to do so forever onwards. We don't want to assign an interpretation $(\mathcal{B}, \mathcal{V})$ to a state S just because everything causally *downstream* of S behaves in the way we'd expect from that interpretation, if there might be states *upstream* of S that don't play nicely with that interpretation.

So in short the problem is that Lewis forgot to look backwards, and a more complete characterisation of fitness should instead be:

Fitness. A scheme $\mathcal{I}$ *fits* iff, if $\mathcal{I}(S_i) = (\mathcal{B}, \mathcal{V})$, then

i) S_i matches $(\mathcal{B}, \mathcal{V})$
ii) for any e consistent with $\mathcal{B}$, if S_i given e yields S_j, then $\mathcal{I}(S_j) = (\mathcal{B}^e, \mathcal{V})$
iii) if S_k given e yields S_i, then $\mathcal{I}(S_k) = (\mathcal{B}_i, \mathcal{V})$ for some $\mathcal{B}_i$ where $\mathcal{B}_i^e = \mathcal{B}$

Fitness gives us an ideal. Imperfect fit can then be cashed out in terms of how close overall an interpretation comes to satisfying this ideal. Exactly *how* we rank interpretations by closeness to the ideal is not an easy problem to address, and I'm not going to attempt to solve it. As far as the underdetermination argument is concerned, we assume that typical agents satisfy the Bayesian theory I've outlined, so at least one fitting scheme of interpretation exists.

2. Radical Indeterminacy

So my cards are on the table: with the exception of Static Values, I think that the kind of functionalism that I've described is *basically* correct. Some tweaks are still needed, some idealisations need to be dropped, and quite a few details need to be worked out. We've still got to deal with logical fallibility, essentially indexical content, imprecise attitudes, and change in intrinsic values. If we want to spell the theory out in thorough-going physicalist terms, we'll also need to say a lot more about what the evidence propositions are and how options ought to be characterised in relation to beliefs. But I doubt the main outlines are going to change much. Indeed, I take it as a constraint on any good systematisation of folk psychology that it doesn't depart too far from what's been described. Whatever we end up with after crossing all the t's and dotting all the i's should look a lot like the kind of theory characterised by $\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, $\succsim$-Coherence, $\succsim$-Maximisation, and Conditionalisation.

That brings us to the underdetermination argument. According to Lewis, if one scheme of interpretation maximises *fit* then there will be others—that is, *unless* we impose eligibility constraints on the contents of beliefs and desires. Lewis

only *very* briefly sketches the argument for this in 'New Work', though the neglected details have more recently been filled in by Schwarz (2012), Weatherson (2012, p. 5), and Williams (2016). Here's how it goes.[5]

2.1 The underdetermination argument

We assume to start with that the six principles describing Typikarl's psychology from §1.2 are all true, and that $\mathcal{B}$-Eligibility and $\mathcal{V}$-Eligibility exhaust the eligibility constraints—i.e., every $(\mathcal{B}, \mathcal{V})$ represents a distinct belief-desire state.

Next, we let

$$\mathbf{G} = \{g_1, ..., g_n\}$$

designate the smallest partition of Ω such that any b in $\bigcup(\mathbf{B}^\tau)_{\tau\in\mathfrak{T}}$ (i.e., all of the *specific* behaviour-propositions, across all times τ) plus any e in $\mathbf{E}$ is expressible as the union of cells in $\mathbf{G}$ (i.e., is a member of $\mathcal{A}(\mathbf{G})$, the algebra generated by $\mathbf{G}$). That is, we want the smallest partition $\mathbf{G}$ such that:

$$\Big(\bigcup_{\tau\in\mathfrak{T}}(\mathbf{B}^\tau)\cup\mathbf{E}\Big)\subseteq\mathcal{A}(\mathbf{G})$$

Given this, the key assumption needed for the underdetermination result is that $\mathcal{A}(\mathbf{A}) \not\subseteq \mathcal{A}(\mathbf{G})$.

An example will help. Suppose that for some g in $\mathbf{G}$, $g = \{a_1, a_2, a_3\}$. Start with some 'decent, reasonable' system of initial belief and intrinsic value, $\mathcal{B}_1$ and $\mathcal{V}_1$. As above, we assume $\mathcal{B}_1$ assigns a non-zero value to each e in $\mathbf{E}$.[6] The specific numbers don't really matter, so suppose:

$$\mathcal{B}_1(a_i) = \begin{cases} 0.1, & \text{if } i = 1 \\ 0.2, & \text{if } i = 2 \\ 0.3, & \text{if } i = 3 \end{cases} \qquad \mathcal{V}_1(a_i) = \begin{cases} 3, & \text{if } i = 1 \\ 6, & \text{if } i = 2 \\ 9, & \text{if } i = 3 \end{cases}$$

We now 'twist' $\mathcal{B}_1$ into a new probability distribution that assigns the same values to every p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$; and we twist $\mathcal{V}_1$ 'in a countervailing way', so as to end up with the same expected desirabilities for every p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$ when those values are combined with the twisted system of beliefs. An obvious way to do this is to simply make parallel rearrangements the probability and

[5] My presentation most closely follows that of Williams (2016), though I've made modifications to reflect some differences between how Williams and I characterise Typikarl's evidence, his options, and state of awareness. The modifications make no important difference to the main steps of the argument, however, and hence I've omitted proofs for those steps.

[6] The point of this assumption is to ensure that there's no evidence-specifying e, other than the trivial evidence Ω, such that for some $\mathcal{B}_0$, $\mathcal{B}_0^e = \mathcal{B}_1$. Consequently, any *fitting* scheme of interpretation can assign interpretations that include $\mathcal{B}_1$ only to what we might call 'initial' states—i.e., states S for which there are no distinct S' which given some e yield S, and as such always sit at the start of any sequence of states that share possible relations via evidence. Note, though, that we do *not* have to assume that actual agents always begin their lives assigning non-zero probability to every evidence proposition. The claim here is about *possible* causal antecedents to the states that an agent might be in. For all I've said, an actual agent might begin their life in any state, including a state S which could (in someone else) have been brought about by an earlier state S' given some evidence e.

desirability values that the atoms making up g are assigned by $\mathcal{B}_1$ and $\mathcal{V}_1$:

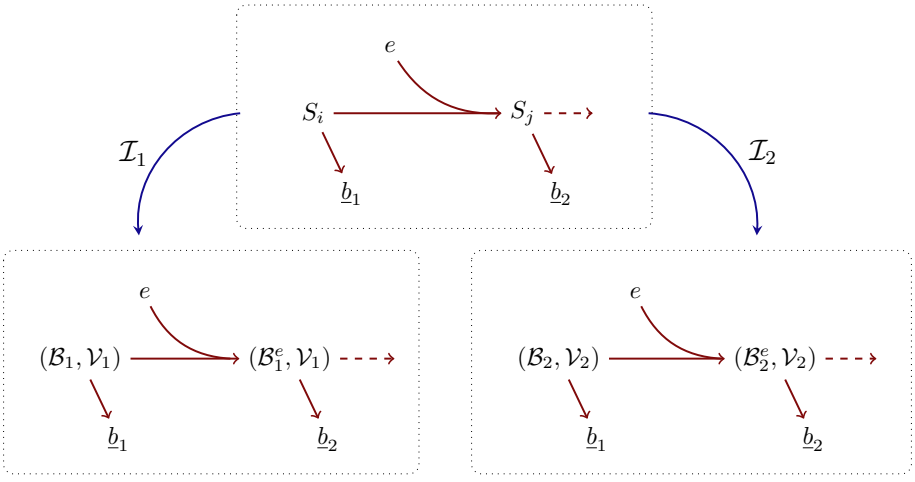
$$\mathcal{B}_2(a_i) = \begin{cases} 0.3, & \text{if } i = 1 \\ 0.2, & \text{if } i = 2 \\ 0.1, & \text{if } i = 3 \\ \mathcal{B}_1(a_i) & \text{otherwise} \end{cases} \qquad \mathcal{V}_2(a_i) = \begin{cases} 9, & \text{if } i = 1 \\ 6, & \text{if } i = 2 \\ 3, & \text{if } i = 3 \\ \mathcal{V}_1(a_i) & \text{otherwise} \end{cases}$$

This kind of 'parallel permutation' will always determine the same strengths of belief and desire for all p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$. As Williams (2016, p. 429) notes, though, there are countless other ways to get the same result. Consider:

$$\mathcal{B}_3(a_i) = \begin{cases} 0.3, & \text{if } i = 1 \\ 0.3, & \text{if } i = 2 \\ 0, & \text{if } i = 3 \\ \mathcal{B}_1(a_i) & \text{otherwise} \end{cases} \qquad \mathcal{V}_3(a_i) = \begin{cases} 7, & \text{if } i = 1 \\ 7, & \text{if } i = 2 \\ 5, & \text{if } i = 3 \\ \mathcal{V}_1(a_i) & \text{otherwise} \end{cases}$$

Now note two things. First, $(\mathcal{B}_1, \mathcal{V}_1)$, $(\mathcal{B}_2, \mathcal{V}_2)$ and $(\mathcal{B}_3, \mathcal{V}_3)$ determine the same expected desirabilities for any disjunction of behaviour propositions in $\mathcal{A}(\mathbf{A}) \cap \bigcup(\mathbf{B}^t)_{t \in \mathfrak{T}}$, and consequently they determine the very same preferences over *options* at a given time, whatever those options may end up being. And second, the same holds even after conditionalising on any evidence-specifying proposition in $\mathbf{E}$—so for any such e, $(\mathcal{B}_1^e, \mathcal{V}_1)$, $(\mathcal{B}_2^e, \mathcal{V}_2)$ and $(\mathcal{B}_3^e, \mathcal{V}_3)$ will *also* determine the same preferences over options.

These two facts entail that for all e, S_i and S_j *match* the interpretations $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_1^e, \mathcal{V}_1)$ respectively just in case they match $(\mathcal{B}_2, \mathcal{V}_2)$ and $(\mathcal{B}_2^e, \mathcal{V}_2)$ respectively, and likewise for $(\mathcal{B}_3, \mathcal{V}_3)$ and $(\mathcal{B}_3^e, \mathcal{V}_3)$. Given the definition of FITNESS, then, it follows that there will be multiple equally-fitting schemes of interpretation:



This establishes *underdetermination* in relation to any propositions in $\mathcal{A}(\mathbf{A})$ that aren't in $\mathcal{A}(\mathbf{G})$. To get *radical* underdetermination, need now only argue that there will be *many* propositions in $\mathcal{A}(\mathbf{A})$ that are not in $\mathcal{A}(\mathbf{G})$.

And this is entirely plausible. The cells of $\mathbf{G}$ need agree only on what behaviours Typikarl performs at what times, and on what sensory evidence he

receives in what sequence. It should be intuitively obvious that we *can* (and almost certainly *does*) have many beliefs and desires regarding matters that aren't equivalent to disjunctions about how we behave and what evidence we get in what order. Lewis' own examples are that *emeralds are green* and that *emeralds are grue*, but hardly the extent of it. The vast majority of the things I take myself to have beliefs and desires about aren't *about* my behaviour and sequences of sensory evidence, and I'm hardly atypical in this respect. More importantly, it's clearly part of *folk psychology* that we can have beliefs regarding propositions that aren't about our own behaviour and/or sensory evidence.

2.2 Generalisations and limitations

The argument I've given makes use of some idiosyncratic features of the 'Jeffreyan' conception of decision theory. In particular, it use the fact that Typikarl's desirability for an option $\underline{b}$ is a function of his beliefs and desires over the atoms that constitute $\underline{b}$. This is a consequence of $\succsim$-Coherence, but it need not be true for non-Jeffreyan decision theories—e.g., it's not true in Lewis' own version of causal decision theory (Lewis 1981), nor does it hold for decision theories spelled out within a 'Savagean' framework.

With that said, it clearly wasn't Lewis' intent for his conclusion to depend on the specifics of Jeffrey's theory, and we have some good reasons to think it doesn't. Williams (2016) shows how to get a similar underdetermination result for causal decision theory, while (Elliott 2017) contains much the same style of argument applied to a wide class of 'Savagean' theories—including generalisations of expected utility theory. In both cases, the main result is reminiscent of Lewis': an agent's choices over options might *at most* determine her beliefs and desires up to a limited subset of $\mathcal{A}(\mathbf{A})$. So Lewis' conclusion seems to be quite robust against variations to the decision rule.

It is, furthermore, robust against obvious ways of 'de-idealising' the Bayesian theory laid out in §1.2. It doesn't hinge on $\mathcal{B}$ and $\mathcal{V}$ assigning *precise* numerical values, for example, nor does it require that either $\mathcal{B}$ or $\mathcal{V}$ be *coherent* beyond what's strictly necessary to make sense of the decision and updating rules.If anything, allowing for more variety in the kinds of beliefs and desires considered possible will only make the underdetermination worse. Likewise, a consequence of $\succsim$-Coherence, $\succsim$-Maximisation, and Conditionalisation is that Typikarl choices over time are quite rigidly determined by his initial beliefs, values, and evidence. Strengthening these constraints might help, but weakening them isn't likely to at all.

But there are two limitations to Lewis' argument that are worth flagging.[7] The first concerns Static Values. Supposing we want to deny Static Values, *how* we go about denying it can make a difference. If we simply say that Typikarl's intrinsic values may randomly change sometimes, then the result will be fewer constraints on transitions between states, and hence more underdetermination. On the other hand, suppose Typikarl's values might systematically change in a way that depends in part on his beliefs and/or desires. Lewis' argument rests on the idea that for any $(\mathcal{B}_1, \mathcal{V}_1)$ there will be a $(\mathcal{B}_2, \mathcal{V}_2)$ that not only generates the same preferences over options, but *also* leads to the same

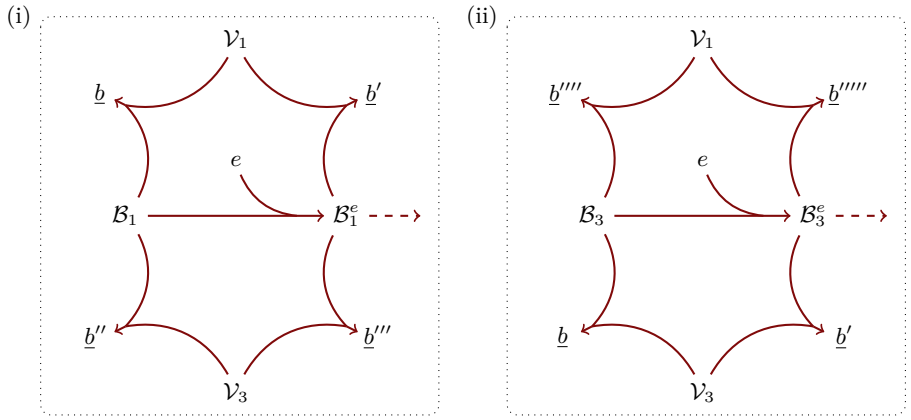
[7] The short version of what follows is that while there are some ways to decrease the severity of the underdetermination Lewis' argument predicts, these aren't likely to fully undermine his conclusion. Readers not interested in the details can skip to §2.3 without loss of comprehension.

preferences over future options for any sequence of evidence. However, if $\mathcal{V}_1$ and $\mathcal{V}_2$ might evolve in different and predictable ways due to differences in $\mathcal{B}_1$ and $\mathcal{B}_2$, then they might lead to divergent predictions about choices and we'd be able to tell $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ apart.

As a rough example, suppose that if Typikarl is confident that $p =$ *Typikarl is a loving relationship with Karl*, then over time he'll assign increasingly more value to those worlds where Karl is well-off; if on the other hand he's not confident that p, then his values for those worlds won't change. At the start of his relationship with Karl there might be two systems of belief and desire that rationalise Typikarl's choices: a 'reasonable' one according to which he's confident that p, and a 'deviant' counter-inductive interpretation according to which the more time he spends with Karl the less confident that p he becomes. The 'reasonable' and 'deviant' interpretations will then diverge in the choices they predict at later times due to the systematic change in values.

The second limitation concerns how *schemes of interpretation* are understood. As Lewis (1983b, p. 119) notes, the correct scheme of interpretation should specify an agent's beliefs and desires as a function of her momentary total physical state. But there are different ways this scheme might work. One involves what we might call *coarse-grained* schemes, which assign whole belief-desire systems $(\mathcal{B}, \mathcal{V})$ directly to total physical states depending on how well those states fit the functional role of $(\mathcal{B}, \mathcal{V})$ considered as a whole. An alternative would be to use a *fine-grained* scheme that assigns sub-total mental states to partial physical states. For example, the fine-grained scheme $\mathcal{I}$ might have us identify the total state S with $(\mathcal{B}, \mathcal{V})$ not *directly* because S satisfies the functional role of $(\mathcal{B}, \mathcal{V})$, but because $S = S_1 \sqcup S_2$, and S_1 satisfies the role of $\mathcal{B}$ while S_2 satisfies the role of $\mathcal{V}$. And this distinction matters because coarse-grained schemes of interpretation ignore causal and counterfactual relationships between sub-total mental states.

Here's an example. $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_3, \mathcal{V}_3)$ are as they were described in §2.1.[8] Then, their functional roles in relation to one another can be represented thus:



We know that $\underline{b}$ maximises expected desirability relative to $(\mathcal{B}_1, \mathcal{V}_1)$, and $\underline{b}'$ maximises expected desirability relative to $(\mathcal{B}_1^e, \mathcal{V}_1)$, just in case $\underline{b}$ and $\underline{b}'$ also

[8] I've skipped $(\mathcal{B}_2, \mathcal{V}_2)$ for a reason—we'll come back to it shortly. A similar example is discussed in more detail in (Elliott forthcoming).

maximise expected desirability relative to $(\mathcal{B}_3, \mathcal{V}_3)$ and $(\mathcal{B}_3^e, \mathcal{V}_3)$ respectively. Because of this, the functional role of the *total* state $(\mathcal{B}_1, \mathcal{V}_1)$ is isomorphic to the functional role of $(\mathcal{B}_3, \mathcal{V}_3)$—that's what Lewis' argument establishes, and it's why the best fitting *coarse-grained* schemes will underdetermine whether a total physical state S should be assigned $(\mathcal{B}_1, \mathcal{V}_1)$ or $(\mathcal{B}_3, \mathcal{V}_3)$.

However, for all we've said so far the functional role of $\mathcal{B}_1$ need *not* be isomorphic to the functional role of $\mathcal{B}_3$; nor need it be the case that the functional role of $\mathcal{V}_1$ is isomorphic to the functional role of $\mathcal{V}_3$. To recall,

$$\sum_{a \in \mathbf{A}} \mathcal{B}_1^g(a)\mathcal{V}_1(a) \;=\; \sum_{a \in \mathbf{A}} \mathcal{B}_3^g(a)\mathcal{V}_3(a) \;=\; 7,$$

thus $(\mathcal{B}_1, \mathcal{V}_1)$ generates the same expected desirabilities over options as $(\mathcal{B}_3, \mathcal{V}_3)$. But these generate different expected desirabilities than $(\mathcal{B}_1, \mathcal{V}_3)$ and $(\mathcal{B}_3, \mathcal{V}_1)$:

$$\sum_{a \in \mathbf{A}} \mathcal{B}_1^g(a)\mathcal{V}_3(a) \;=\; 6 \qquad \sum_{a \in \mathbf{A}} \mathcal{B}_3^g(a)\mathcal{V}_1(a) \;=\; 4.5$$

So under recombinations with the values $\mathcal{V}_1$ and $\mathcal{V}_3$, $\mathcal{B}_1$ and $\mathcal{B}_3$ generate distinct expected desirabilities for at least some p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$, and hence at least in principle might end up generating different patterns of choice dispositions when combined with the same systems of intrinsic value.

More generally, say that $\mathcal{B}'$ is a *permutation* of $\mathcal{B}$ just in case $\mathcal{B}'$ rearranges the values $\mathcal{B}$ assigns to atoms that are subsets of the $g \in \mathbf{G}$, and otherwise assigns the same values to all atoms that cross-cut $\mathbf{G}$'s cells. Define permutations for desirability functions similarly. Now let the *recombination function* $\mathcal{R}_{\mathcal{V}}^{\mathcal{B}}$ select for each p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$ the *set* of expected desirabilities assigned to p by those $(\mathcal{B}', \mathcal{V}')$ such that $\mathcal{B}'$ is a permutation of $\mathcal{B}$ and $\mathcal{V}'$ is a permutation of $\mathcal{V}$. With a finite $\mathbf{A}$ it turns out that if $\mathcal{B}_i$ isn't a permutation of $\mathcal{B}_j$ *or* $\mathcal{V}_i$ isn't a permutation of $\mathcal{V}_j$, then

$$\mathcal{R}_{\mathcal{V}_i}^{\mathcal{B}_i} \neq \mathcal{R}_{\mathcal{V}_j}^{\mathcal{B}_j}$$

That is: any two systems of beliefs (or intrinsic values) that *aren't* permutations of one another will generate a distinctive pattern of expected desirabilities for the p in $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$ when combined with different systems of value (beliefs). And that's interesting, because *most* of the underdetermination that's established by the argument in §2.1 isn't between parallel permutations.[9]

But I don't think these facts are a cure-all to the underdetermination argument. Regarding the first limitation, my guess is that any plausible theory of typical changes in intrinsic values won't affect the conclusion too much. To the extent that it's typical for intrinsic values to change over time, those changes probably aren't wholly *random*—but at the same time it strikes me as unlikely that they change with the required kind of *systematicity* needed to completely undermine Lewis' conclusion. And regarding the second limitation, that $\mathcal{B}_1$ and

[9] Lewis expresses a preference for fine-grained schemes in (1983b): 'an interpretation [must follow from] a scheme of interpretation specifying the attitudes and meanings as a function of the momentary total physical state. On the basis of such states, the scheme assigns interpretations to individuals at times. (Indeed it might—and should, I think—do this simply by identifying certain attitudes with certain (partial) physical states.)' (p. 119). Given this, and since fine-grained schemes of interpretation cannot help us to distinguish between permutations, it's entirely possible that Lewis made used coarse-grained schemes merely to simplify the discussion in 'New Work'.

$\mathcal{B}_3$ generate different patterns of *expected desirabilities for the p in* $\mathcal{A}(\mathbf{A}) \cap \mathcal{A}(\mathbf{G})$ under recombination doesn't yet entail they generate different *preferences over options*; still less does it entail different *choice dispositions*. In most cases the latter will also be true, but not always. More importantly, we cannot use recombinations to distinguish between permutations—it turns out that if $\mathcal{B}$ and $\mathcal{B}'$ *are* permutations of one another, then the functional role of $\mathcal{B}$ *is* isomorphic to the role of $\mathcal{B}'$. So while we *might* be able to use recombinations to tell $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_3, \mathcal{V}_3)$ apart, they won't help us to distinguish between $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ (or any other parallel permutations).

2.3 Lewis' eligibility solution

So how can we fix this situation? Lewis' proposal was, in effect, to impose stronger eligibility constraints than what's entailed by $\mathcal{B}$-Eligibility and $\mathcal{V}$-Eligibility alone. Not every $(\mathcal{B}, \mathcal{V})$ pair represents a genuinely possible belief-desire state:

> 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 fit. They impose *a priori*—albeit defeasible—presumptions about… what dispositions to develop beliefs and desires, what inductive biases and intrinsic values, someone may rightly be interpreted to have. (1983a, p. 375)

To continue the example, we might suppose that $(\mathcal{B}_1, \mathcal{V}_1)$ is overall more reasonable than $(\mathcal{B}_2, \mathcal{V}_2)$ does. An initial state S matches the interpretation $(\mathcal{B}_1, \mathcal{V}_1)$ just when it matches $(\mathcal{B}_2, \mathcal{V}_2)$. Consequently, according to Lewis' suggestion a scheme of interpretation that assigns $(\mathcal{B}_1, \mathcal{V}_1)$ to all such states (and assigns the appropriate interpretations for all downstream states) will be better overall, *ceteris paribus*, than any scheme which assigns $(\mathcal{B}_2, \mathcal{V}_2)$ to any one of these states. The upshot is that $(\mathcal{B}_2, \mathcal{V}_2)$ will never be assigned to *any* state by the *best* scheme, and if there's something that's even more reasonable than $(\mathcal{B}_1, \mathcal{V}_1)$, then $(\mathcal{B}_1, \mathcal{V}_1)$ will be removed from consideration as well. Some ways of having beliefs and desires are rendered 'ineligible', and so we avoid one of the key assumptions needed to get the underdetermination argument off the ground.

Lewis evidently also thought it likely that the strengthened eligibility constraints would be enough to determine a unique best scheme of interpretation whenever more than one has equal fit. He never provided an argument for this; nor for that matter did he say very much about what the eligibility requirements were supposed to be aside from some vague connections with reasonableness. But that's work for another paper. Let's turn now to representation theorems, and whether these have any implications for Bayesian functionalism.

3. The Irrelevance of Representation Theorems

A representation theorem for expected utility theory is often taken to supply us with conditions under which we can derive the facts about an agent's beliefs and desires if we have enough information about her choice dispositions. There's

no small number of these theorems—Peter Fishburn's (1981) well-known review covers 28 of them, and there's been plenty more published in the decades since. I couldn't hope to cover all of them in any detail here.

So here's what I'm going to do. I'll start by describing the stereotypical representation theorem (§3.1). Then I'll consider the suggestion that Lewis would have rejected some of the conditions these theorems rely on (§3.2). And then, finally, I'll discuss the deeper reason why there's no genuine conflict between the underdetermination argument and the representation theorems (§3.3), and indeed why the latter just aren't very relevant for radical interpretation (§3.4).

3.1 The stereotypical representation theorem

To start, note that any precisification of expected utility theory will do two things. First, the theory will say that an agent's system of preferences $\succsim$ over some pre-specified domain $\mathbf{D}$ will be determined by her beliefs and intrinsic values according to some precisification of the expected utility rule. Depending on the theory in question, the domain $\mathbf{D}$ might be a set of act-specifying propositions (as in Savage 1954); or it might be a more limited set of 'worlds' and 'gambles' (as in Ramsey 1931); or it might be an arbitrary algebra of propositions (as in Jeffrey 1965). Second, the theory will also impose at least some minimal restrictions conditions on beliefs and desires—for example, that the agent's beliefs be coherent enough to be representable by a probability distribution.

Given this, suppose we represent the content of the theory as a function, $\mathcal{T}_{\text{EU}}$, from eligible systems of belief and desire to systems of preference; i.e.,

$$\mathcal{T}_{\text{EU}} : (\boldsymbol{\beta} \times \boldsymbol{\delta}) \mapsto \boldsymbol{\pi},$$

where:

$\boldsymbol{\beta} = \{\mathcal{B}_1, \mathcal{B}_2, ...\}$ = the eligible systems of belief
$\boldsymbol{\delta} = \{\mathcal{V}_1, \mathcal{V}_2, ...\}$ = the eligible systems of intrinsic values
$\boldsymbol{\pi} = \{\succsim, \succsim', ...\}$ = the possible systems of preference over $\mathbf{D}$

The stereotypical representation theorem for the theory $\mathcal{T}_{\text{EU}}$ will then be:

Representation Theorem. If a system of preferences $\succsim$ satisfies $c_1, ..., c_n$, then $\mathcal{T}_{\text{EU}}(\mathcal{B}_i, \mathcal{V}_i) = \mathcal{T}_{\text{EU}}(\mathcal{B}_j, \mathcal{V}_j) = \succsim$ iff $(\mathcal{B}_i, \mathcal{V}_i) = (\mathcal{B}_j, \mathcal{V}_j)$

The $c_1, ..., c_n$ will include some domain-independent conditions, such as that $\succsim$ must be transitive and complete. The conditions will also, whether directly or indirectly, impose at least some restrictions on $\mathbf{D}$—e.g., that it has a certain logical structure. I'll talk about the latter kind of condition more in §3.2, but for now we can safely ignore the specifics of what $c_1, ..., c_n$ actually say.

Ramsey's theorem takes the above form, and Savage's is often thought to do so as well. Because of this, they're frequently taken to show that it's possible to fix an agent's beliefs and desires given enough information about her *choice dispositions*. For this to be true, though, we first need an appropriate theoretical connection between preferences and choice dispositions. Something like $\succsim$-Maximisation will tell us that the typical agent at time τ will choose whichever of her options she considers best at τ, but that's not going to be enough to derive her preferences at τ—more is needed.

Well, we know how to solve this one, right? Counterfactual choices reveal preferences. Specifically, (typical) agents' preferences correspond directly to their choice dispositions under hypotheses about available options:

Revealed Preference. If Typikarl has beliefs and desires $(\mathcal{B}, \mathcal{V})$, then if his available options were $\mathbf{D}^\star \subseteq \mathbf{D}$, then he would be equally disposed between making any of the $p \in \mathcal{C}(\mathbf{D}^\star)$ true

In other words, the hypothesis is that the choice function $\mathcal{C}$ describes Typikarl's dispositions to choose under counterfactual hypotheses about what options are available. Next, say that $p \succsim^c q$ just in case, if p and q were Typikarl's only options, then he'd be disposed to choose either p alone or equally disposed between p and q. Following Hausman (2012, pp. 1–2), we'll call $\succsim^c$ a *choice ranking*. If Revealed Preference is true then it provides the required theoretical link between preferences and choice dispositions represented as a choice ranking.[10]

Suppose that we can make sense of any two propositions p and q in the relevant domain $\mathbf{D}$ as being the only potential objects of choice in some counterfactual choice situation. Then, from $p \succsim q$ we know that $\mathcal{C}(\{p, q\})$ will equal either $\{p\}$ or $\{p, q\}$, and by Revealed Preference this entails $p \succsim^c q$. So $p \succsim q$ implies $p \succsim^c q$. The other direction is a little more tricky, as there will be some propositions that sit outside of the preference ordering. However, at least in Typikarl's case this is easily resolved: his preference ordering $\succsim$ is transitive and complete over those p that are assigned some positive probability; for all the rest, Revealed Preference entails they'll sit together at the bottom of the choice ranking. Hence, $p \succsim^c q$ implies $p \succsim q$ just in case neither p nor q is bottom $\succsim^c$-ranked; otherwise p and q aren't $\succsim$-related to one another. In sum, then, $\succsim$ *just is* $\succsim^c$ minus the bottom rung.

So here's where we're at: *if* there's an appropriately tight connection between choice dispositions and preferences, then it's plausible enough to say that a representation theorem like Ramsey's or Savage's describes sufficient conditions for when an agent's choice dispositions uniquely determine her beliefs and desires, under the assumption she maximises expected utility. And this is just the kind of thing you might *think* a Bayesian functionalist ought to be very interested in, especially in light of the underdetermination argument. So why did Lewis never mention any of these theorems in his published works?

3.2 Domain-richness conditions

There's no reasonable doubt that Lewis would have been aware of Ramsey's and Savage's theorems at the time of writing 'New Work'. In his (2014), Schwarz states that 'Lewis did not trust these results' (pp. 21–2), and pins that distrust on the fact that amongst the conditions $c_1, \ldots, c_n$ that Savage, Ramsey, *et al.*, use to prove their uniqueness results there will usually be some strong (and controversial) requirements relating to the 'richness' of the domain $\mathbf{D}$ over which the agent's preferences are defined (see also (Schwarz 2012; 2015, p. 513). Williams (2016, p. 430) argues something similar:

[10] Here, the choice function $\mathcal{C}$ is defined using $\succsim$, and it's hypothesised that $\mathcal{C}$ describes $\succsim^c$. An alternative would be to define a choice function using $\succsim^c$, and then hypothesise that it 'reveals' $\succsim$. It makes no difference which we go with. On the other hand, some behaviourists have supposed that Revealed Preference is *analytic*, so $\succsim$ and $\succsim^c$ are interchangeable a priori. If this were true then it would be problematic for my argument. Isn't it lucky, then, that Revealed Preference isn't analytic (barring a change of topic).

> In a sense [Lewis' argument] is the limitative counterpart of the representation theorems [...] They showed that if you assume that choices across a range of options are sufficiently rich, the degree of belief and desirability can be pinned down. To get that result, you need to have information about the choices an agent is disposed to make between arbitrary propositions. But the sphere of action of ordinary agents like us is limited—we cannot choose whether or not a fly on Alpha Centauri keels over or not. (p. 430)

If this is right, then you can see one way to reconcile Lewis' argument with Savage's and Ramsey's results: if the conditions $c_1, \ldots, c_n$ are never jointly satisfied because the domain-richness conditions fail, then choice dispositions never uniquely determine beliefs and desires.

But I don't think this gets us to the heart of the matter. Lewis certainly *might* have rejected these domain-richness conditions—a great many others have!—however it's not obvious that he would have in all relevant cases, and more importantly there was no *need* for him to do so. Ultimately, the reason for this is that even if we take an agent's space of options to be arbitrarily rich, there would *still* be no direct conflict between Lewis' conclusion and theorems like Ramsey's and Savage's. The key to reconciliation relates to the difference between how 'choice dispositions' are understood in the context of Lewis' theory, and how they're understood in relation to representation theorems. This is what I'll argue below. Right now, let me first say why it's not obvious that Lewis would have rejected these domain-richness assumptions.[11]

We'll focus on Savage, since that's where most of the controversy lies. Savage's formalisation of decision theory starts off with two partitions: a set of *states*, and a set of *outcomes* that specify ways the world might be in as much detail as makes a difference to what we care about—think of the outcomes as sets of equally-desirable worlds. From there, Savage defines his preference relation over total functions from states to outcomes, each intended to represent a distinct *act*. If the act-function $\mathcal{F}$ takes us from state s_1 to outcome o_1, s_2 to o_2, and so on, then $\mathcal{F}$ represents the act such that if it's chosen and s_1, then o_1 results; if it's chosen and s_2, o_2 results, and so on. So far so good. But one of the assumptions Savage uses to prove his uniqueness result is that *every* act-function represents a distinct act. This implies, amongst other things, the existence of so-called *constant acts*—acts with the same outcome regardless of what state happens to be true—which has long been a focal point of criticism for Savage's theorem. (See, e.g., Fishburn 1981, p. 162; Maher 1993, pp. 182–5; Joyce 1999, pp. 107–8.) Would Lewis have rejected Savage's assumption too?

Before we answer that, we should get clear on what the problem with constant acts is supposed to *be*. First, if Savage was assuming that every act-function represents an *available* act, then it's implausible that there's an act available to me that will bring about any level of value I so care to choose. But it's unlikely that this is what Savage had in mind. If every constant act were available then any minimally rational agent would just choose the constant act that guarantees the best outcome come what may (Joyce 1999, p. 67). Savage's theorem would then be relevant only to those who can in fact choose any outcome they like at will, and this is not how Savage understood the import of his work. More

[11] Readers not interested in the details should feel free to skip to §3.3.

importantly, Revealed Preference doesn't *require* us to only consider available acts. After all, if we're going to be considering what Typikarl would do under counterfactual hypotheses about his option set, why restrict ourselves to his actually available acts? When drawing up our decision tables we'll focus on the acts we consider available, of course, but that doesn't mean we don't have preferences over possible-but-not-available acts and dispositions to choose between them if they were to somehow become available.

Ok—but isn't the problem with constant acts that they might not be *possible*? Well, that depends a lot on what the outcomes look like. On the one hand, imagine that Typikarl has an extremely opinionated system of intrinsic values: he assigns a distinct value to each and every possible world. There would then almost certainly be at least one world ω such that each act Typikarl performs within that world is also performed within some other world. It follows immediately that there are no *possible* acts that are guaranteed to result in something exactly as valuable as ω. So for *some* ways of carving up the outcomes, corresponding to *some* ways an agent's intrinsic values might be, Savage's domain-richness assumption does seem to imply the existence of impossible acts. On the other hand, suppose Typikarl doesn't care about very much at all—say, there are only two or three relevantly distinct outcomes as far as he's concerned. Then it's very plausible that there's going to be some way of carving up the states such that for each outcome there's at least one *possible* act that guarantees that outcome regardless of which state happens to be true.

As a general rule of thumb, the more fine-grained the outcome-partition, the less likely it will be that there exists a set of states such that every function from states to outcomes corresponds to a possible act. But the flip-side is that the more coarse-grained the outcomes are, the more plausible Savage's domain-richness assumption becomes. Now, because we usually do care about quite a lot of things, Savage's assumption is problematic for most *actual* systems of intrinsic value. But, crucially, the conditions $c_1, \ldots, c_n$ of Savage's theorem don't require the outcome-partition to be extremely fine-grained, and neither does Lewis' argument at any point require any assumptions about how opinionated Typikarl's intrinsic values can be. So there are cases where Lewis' argument applies for which we've no special reason to reject Savage's domain-richness assumption. Whatever's going on between Lewis and Savage, it's not—or not only—a disagreement over the structure of the act-space. We're going to need something more general than this if we're going to reconcile the two.[12]

3.3 Evidence-counterfactual choices

Thus far I've been going along with the idea that the facts about Typikarl's choice dispositions, *in the sense that Lewis understood them*, can be appropriately represented by a choice ranking $\succsim^c$ that 'encodes' his preferences. But at no point is this how Lewis himself describes things.

Here's a typical instance of what Lewis says about the kind of information relating to choice and behaviour we have for radical interpretation:

[12] In addition to the points above, it's worth noting that while Savage *used* the assumption that every act-function represents a possible act, it appears that this assumption can be very significantly weakened. More recent theorems which closely follow the structure of Savage's own manage to establish conditions sufficient for unique rationalisability without assuming anything as strong (e.g., Abdellaoui and Wakker 2005; Gaifman and Liu 2018).

> Thus if [the physical facts entail] that Karl's arm goes up at a certain time, [we] should ascribe beliefs and desires according to which it is a good thing for his arm to go up then. I would hope to spell this out in decision-theoretic terms, as follows. Take a suitable set of mutually exclusive and jointly exhaustive propositions about Karl's behaviour at any given time; of these alternatives, *the one that comes true according to* $\boldsymbol{P}$ *should be the one (or: one of the ones) with maximum expected utility* according to the total system of beliefs and desires ascribed to Karl at that time… (1974, p. 337, emphasis added; see also 1983a, p. 374; 1986, pp. 36ff)

That is, not a *ranking* of Karl's possible options, but a proposition about his behaviour that entails (one of) the option(s) that's maximal relative to his beliefs and desires (i.e., assuming he's not mistaken about what he can do). And in (1980b), Lewis adds a bit more:

> … what makes it be so that a certain reasonable initial credence function and a certain reasonable system of basic intrinsic values are both yours is that you are disposed to act in more or less the ways that are rationalized by the pair of them together, taking into account the modification of credence by conditionalizing on total evidence; and further, *you would have been likewise disposed if your life history of experience, and consequent modification of credence, had been different*… (pp. 287–8, emphasis added)

So, for each time τ, we know :

i) which $b \in \mathbf{B}^\tau$ Typikarl makes true at τ
ii) which $b \in \mathbf{B}^\tau$ he *would* have made true if he were to have had this or that sensory evidence up to τ

Note that I've been saying 'behaviours' here for a reason—not 'options'. Since Typikarl's *options* depend in part on his beliefs, these are not something to which we can assume access. The physical information tells us how Typikarl *behaves*, in complete detail. If Typikarl's not mistaken about what's available then the specific behaviour-proposition he makes true will entail an option that's amongst those he desires most. But in general there will be a range of hypotheses about what Typikarl's options are at a time each consistent with the facts about his behaviour at that time, and hence we need to treat the former as a variable to be solved for in radical interpretation.

So let's make a distinction. On the one hand there's the matter of what options Typikarl would be disposed to choose if his available options were thus-and-so. We'll call these *option-counterfactual choice dispositions*. If Revealed Preference is correct, then Typikarl's option-counterfactual choice dispositions will be described by his choice function $\mathcal{C}$—and from this we can extract a choice ranking $\succsim^c$ that directly 'encodes' his preferences. On the other hand, there's how Typikarl would be disposed to behave if his evidence were thus-and-so. We'll call these *evidence-counterfactual behavioural dispositions*. The question for us now is whether Typikarl's evidence-counterfactual behavioural dispositions suffice to determine his preferences. (Spoiler alert: they don't.)

Here's what we know. If Typikarl starts off at time τ_1 with beliefs and desires $(\mathcal{B}, \mathcal{V})$, then if he were to receive the sequence of evidence e he would be

disposed to choose any one of the options that have maximal desirability relative to $(\mathcal{B}^e, \mathcal{V})$, from amongst the options $\underline{\mathcal{B}}^{\tau_{1+e}}$ available from his perspective at τ_{1+e}, where τ_{1+e} is τ_1 plus the duration the stream of evidence e. Thus, we can define the *evidential choice function* $\mathcal{E}$ relative to the initial system of beliefs and desires $(\mathcal{B}, \mathcal{V})$ as follows: for each $e \in \mathbf{E}$, $\mathcal{E}(e)$ picks out the set of $\underline{b}_i \in \underline{\mathbf{B}}^{\tau_{1+e}}$ such that for all $\underline{b}_j \in \underline{\mathbf{B}}^{\tau_{1+e}}$ where $\mathcal{B}^e(\underline{b}_j) > 0$,

$$\sum_{a \in \mathbf{A}} \mathcal{B}^e(a|\underline{b}_i)\mathcal{V}(a) \geq \sum_{a \in \mathbf{A}} \mathcal{B}^e(a|\underline{b}_j)\mathcal{V}(a),$$

In other words, $\mathcal{E}$ represents what *options* Typikarl would most prefer, if he were to have evidence e, given that he starts with beliefs and desires $(\mathcal{B}, \mathcal{V})$. In many ways this is very similar to how we defined the choice function $\mathcal{C}$, so you might think that we could use $\mathcal{E}$ to extract Typikarl's preferences over the relevant domain $\mathbf{D}$ out of his evidence-counterfactual behavioural dispositions similar to how we used Revealed Preference to extract his preferences over $\mathbf{D}$ out of $\mathcal{C}$.

There's a few general reasons to think that this isn't so. Let's get a couple of the obvious points out of the way first. To start with, $\mathcal{E}$ only tells us what *options* Typikarl most prefers, from amongst those he believes are available. But that's not the main concern here, so let's assume that every $b \in \mathbf{B}^\tau$ is an option at τ from Typikarl's perspective, and he's never wrong about what he can do—so he makes b true iff b maximises expected desirability. Next, it need not be the case that every p in the relevant domain $\mathbf{D}$ is an option at some time. Indeed, the problem is a little worse than that—$\mathcal{E}$ won't help us draw any comparisons between options available *at different times*, and it need not be that every p in $\mathbf{D}$ is an option *at the same time*. Ok—so let's also assume that every sequence of evidence is the same duration, and $\mathbf{D} = \mathbf{B}^{\tau_{1+e}}$.

These are recklessly implausible assumptions, especially given that representation theorems will usually require $\mathbf{D}$ to be a richly structured set of act-specifying propositions or an algebra over Ω. But even those assumptions are not going to be enough. Imagine that

$$\mathbf{D} = \{b_1, b_2, b_3, b_4\}$$

and Typikarl's preferences are:

$$b_1 \succ b_2 \succ b_3 \succ b_4$$

In order to 'extract' these preferences from $\mathcal{E}$, we need that there are evidence-propositions that 'line up' nicely with the appropriate restrictions on $\mathbf{D}$. For example, we know that $\mathcal{E}(\Omega) = \{b_1\}$, so Typikarl prefers $\underline{b}_1$ over the others. Now we need to know how Typikarl ranks the sub-maximal options. There's two ways we can do this: we can either consider his behaviour after conditionalising on $b_2 \cup b_3 \cup b_4$ and then $b_3 \cup b_4$; or we can consider his behaviour after conditionalising on $b_2 \cup b_3$ and then $b_3 \cup b_4$. If we conditionalise on anything else, then we run the risk of distorting the results. And that's a problem because we're not going to find *any* of $\underline{b}_2 \cup \underline{b}_3 \cup \underline{b}_4$, $\underline{b}_2 \cup \underline{b}_3$, or $\underline{b}_3 \cup \underline{b}_4$ in $\mathbf{E}$. Each evidence-specifying proposition characterises *in full detail* the content of Typikarl's sensory evidence over some period of time. So e will tell us what sights and sounds and so on he experiences in what sequence, and if he *doesn't* experience any sights or sounds at some

point, then that will be entailed by e as well—and no proposition like this is going to equivalent to *Typikarl behaves in way* b_3 *or way* b_4.

Note that this problem arises even if the space of available options is *arbitrarily* rich. Indeed, the richer the space of options we posit, the more evidence-specifying propositions we'll need to find in $\mathbf{E}$ in order to extract Typikarl's preferences out of the evidential-choice function $\mathcal{E}$—and we know already that Lewis' underdetermination argument fails if $\mathbf{E}$ is made too rich.

And there's a further problem still if we assume not only that every proposition is an option, but also that every proposition can can count as evidence. For note that if every proposition is an option, TypiKarl will make true that atom (or disjunction of atoms) a with maximal desirability; so $\mathcal{E}(\Omega) = a$. But now consider: for some q that's independent of a, how are we going to determine TypiKarl's preferences between q and $\neg q$? We obviously can't consider $\mathcal{E}(q \cup \neg q)$, since we know already that$\mathcal{E}(\Omega) = a$ and that's of no use. Likewise, suppose $a \subset (q \cup r)$. How do we use $\mathcal{E}$ to determine TypiKarl's preferences between q and r, when $\mathcal{E}(q \cup r) = a$? The fundamental problem here is that, whereas the choice function $\mathcal{C}$ fixes which *set* of propositions TypiKarl has to choose between—which allows us to consider arbitrary sets of options like $\mathcal{A}(\mathbf{A})$, $\{q, \neg q\}$ and $\{q, r\}$ separately—the evidential-choice function $\mathcal{E}$ merely takes as input a single proposition e. If we assume that every proposition is available for choice, then $\mathcal{E}(e)$ just tells us the *atom* within e that's maximally ranked. With enough evidence propositions in $\mathbf{E}$, we can therefore use $\mathcal{E}$ to determine TypiKarl's preferences over the set of *atoms*. But a ranking on atoms will in most cases significantly underdetermine *preferences*.

In summary, then, given plausible assumptions about the option-space and the evidence-space, Typikarl's dispositions to choose given this or that evidence as encapsulated by the function $\mathcal{E}$ will not suffice to determine his preferences. We can determine bits and pieces of Typikarl's preferences over options out of $\mathcal{E}$, if we know what those options are—but in general most of his preference ranking will not be 'revealed' by his evidence-counterfactual behavioural dispositions.

So here, I think, is the deeper reason why there's no conflict. A representation theorem like Ramsey's or Savage's entails that under the right conditions $c_1, ..., c_n$, Typikarl's preferences $\succsim$ will be uniquely determined relative to $\mathcal{T}_{\text{EU}}$ by some system of beliefs and desires $(\mathcal{B}, \mathcal{V})$. If we combine this with Revealed Preference, then $\succsim$ is (more or less) just $\succsim^c$, and we can use option-counterfactual choice dispositions to determine Typikarl's beliefs and desires. But nothing in Lewis' writings suggests he would have been happy to include Revealed Preference as an accepted part of the best systematisation of folk psychology. *His* understanding of folk psychology includes $\mathcal{T}_{\text{EU}}$ or something like it as a proper part, but it's combined with a theory of choice and a theory of how beliefs change over time in response to sensory evidence. It says that every system of eligible beliefs and intrinsic values, plus a time, determines a disjunction of propositions describing how Typikarl would behave *if* given this or that sequence of evidence (including the trivial evidence Ω). Thus,

$$\mathcal{T}_{\text{FP}} : (\boldsymbol{\beta} \times \boldsymbol{\delta} \times \boldsymbol{\mathfrak{T}}) \mapsto \{\mathcal{E} : \mathbf{E} \twoheadrightarrow \bigcup_{\tau \in \mathfrak{T}} \left(\wp(\underline{\mathbf{B}}^\tau)\right)\}$$

Specifically, $\mathcal{T}_{\text{FP}}(\mathcal{B}, \mathcal{V}, \tau)$ picks out a function $\mathcal{E}$ from sequences of evidence e compatible with $\mathcal{B}$ to the disjunctions of options available at a later time that

maximise desirability relative to $(\mathcal{B}^e, \mathcal{V})$. And even if a result like Representation Theorem entails that there exists some $\succsim$ such that

$$\mathcal{T}_{\text{EU}}(\mathcal{B}_i, \mathcal{V}_i) = \mathcal{T}_{\text{EU}}(\mathcal{B}_j, \mathcal{V}_j) = \succsim \quad \rightarrow \quad (\mathcal{B}_i, \mathcal{V}_i) = (\mathcal{B}_j, \mathcal{V}_j)$$

this does not entail that there will be any $\mathcal{E}$ such that

$$\mathcal{T}_{\text{FP}}(\mathcal{B}_i, \mathcal{V}_i, \tau) = \mathcal{T}_{\text{EU}}(\mathcal{B}_j, \mathcal{V}_j, \tau) = \mathcal{E} \quad \rightarrow \quad (\mathcal{B}_i, \mathcal{V}_i) = (\mathcal{B}_j, \mathcal{V}_j)$$

Or to put that another way: even if it's possible to determine Typikarl's beliefs and desires given enough knowledge about his preferences, that's not going to be of much use if we cannot derive his preferences from his evidence-counterfactual behavioural dispositions.

3.4 Option-counterfactual choices

Finally, we should talk about the elephant in the room. I've said that Lewis understood *choice dispositions* in terms of how an agent would behave if her evidence were thus and so, not in terms of how the agent would choose if her options were thus and so. But nothing I've said so far entails that we *can't* use option-counterfactual choice dispositions to pin down an agent's preferences and hence, given the right representation theorem, her beliefs and desires. After all, if we can help ourselves to one kind of counterfactual, then why not help ourselves to the other kind as well?

Well, not so fast. Deriving preferences from option-counterfactual choices requires something like Revealed Preference to mediate the inference, so first we need to ask whether that hypothesis belongs in any decent systematisation of folk psychology. I don't think it does. Indeed, Revealed Preference doesn't play very nicely at all with the kind of psychological theory described in §1.2. If we like the latter, we should reject the former.[13]

Let me start by briefly noting and then setting aside three commonly cited issues that I'm *not* going to be worried about, before we get on to discussing the real problem.

1. There are long-standing concerns about whether it's possible to use choice dispositions to distinguish between *strict preference* and *indifference*, since an agent might (be disposed to) choose one option rather than another despite being indifferent between the two (Savage 1954, p. 17; see also Maher 1993, pp. 12–4); or to distinguish between *indifference* and a *lack of preference* that's due to, e.g., incommensurability (Sen 1971; 1973; Joyce 1999, pp. 99–100). As noted in §3, the latter concern doesn't arise given the assumptions made in §1.2. Moreover, both types of ambiguity will usually be resolved by comparing choice dispositions under arbitrarily small 'sweetenings' of the options. Preference and lack-of-preference are in most cases robust under sufficiently small improvements to the options; indifference isn't. This solution requires us to know what constitutes an improvement, of course, and that's problematic *if* you're operating under complete ignorance about the kinds of things an agent values. That's the reason Savage didn't much care for the 'sweetening' solution (1954, p. 17). But for Lewis,

[13] To be clear: I'm *not* saying that Lewis would have agreed with any of the following points. The issue her concerns what *we* should say about the Revealed Preference hypothesis.

intrinsic values adhere to substantive reasonableness constraints (cf. Lewis 1986, pp. 38, 107; 1994, p. 427; 1996, p. 306), so *plausibly* we could use these considerations to determine what would *reasonably* be considered a small improvement to the agent's options.

2. As mentioned earlier, it may not make sense to say that for any non-empty set $\mathbf{D}^\star \subseteq \mathbf{D}$, an agent can be in a situation where $\mathbf{D}^\star$ exactly characterises her full range of options. This looks especially implausible where $\mathbf{D}^\star$ consists in two randomly chosen propositions. But for the sake of argument I'm happy to pretend that this always makes sense, regardless of what $\mathbf{D}^\star$ happens to be.

3. Revealed Preference requires us to know what an agent's *options* are, but an agent's options depend at least in part on her epistemic situation and how she conceives of her choices. This point is made especially forcefully in (Hausman 2000; 2012, pp. 27ff), and it has received some recent attention (e.g., Thoma forthcoming). It's an important problem—but it's also problem that Lewis has to face up to anyway. The purely physical facts tell us directly only how TypiKarl *behaves* in this or that counterfactual circumstance; how he *chooses* depends on his options, and we won't know that until we know his beliefs. So let's engage in a bit more make-believe, and pretend that we can know what TypiKarl's options are without knowing his beliefs.

Now we're ready for the real problem. In any realistic decision situation, there's a very wide range of available options an agent might choose between. We'll ignore most of these when drawing up a decision table, but instead of just *going out for Thai food* versus *going out for pizza*, one could for example *dance around like a chicken* or *sing a Springsteen power ballad*. So, imagine that the options from which Typikarl can actually choose are given by

$$\underline{\mathbf{B}} = \{\underline{b}_1, ..., \underline{b}_{100}, \underline{b}_\varnothing\},$$

with

$$\underline{b}_1 \succ \underline{b}_2 \succ ... \succ \underline{b}_{99} \succ \underline{b}_{100}$$

$\succsim$-Maximisation and Revealed Preference alike both predict that in this case, Typikarl will be disposed to choose $\underline{b}_1$. But let's now go to the counterfactual scenario where his options are given by, let's say,

$$\underline{\mathbf{B}}^\star = \{\underline{b}_{83}, \underline{b}_{84}\}$$

$\succsim$-Maximisation makes no predictions about what Typikarl will do in this case—that principle tells us how Typikarl will choose amongst his options given his preferences, *not* how Typikarl *would* choose if his options were thus and so given that his preferences *actually* thus and so. Revealed Preference, on the other hand, links option-counterfactual choices directly to actual preferences; hence it predicts that Typikarl will choose $\underline{b}_{83}$. So now consider: *are Typikarl's beliefs and values the same in counterfactual as they are in the actual scenario*?

If his beliefs and desires remain the same across the two scenarios, then whence the change in behaviour? In the actual scenario he chooses $\underline{b}_1$ on the basis of his beliefs and desires; in the counterfactual he purportedly chooses $\underline{b}_{83}$. What could plausibly explain the difference aside from a change in beliefs about what's available for choice? Surely we don't want to posit that Typikarl

has some magical direct access to what options are available, which informs his choices *without* in any way affecting his beliefs (cf. Hausman 2012, pp. 31–3).

No: if there's a difference in Typikarl's behaviour across the different scenarios, then the most plausible explanation will involve some difference in his beliefs. But now it matters a great deal for the plausibility of Revealed Preference exactly *how* Typikarl's beliefs differ in the counterfactual scenario. After all, imagine coming to believe that the (usually very large) range of options that you thought you had to choose between has been reduced down to exactly two. I'd imagine this would involve quite a large change in your beliefs about the world around you, since the nearest possible world where anything like that could be true—if there even *are* any—would be quite far off indeed. So, presumably, Typikarl's beliefs in these strange counterfactual circumstances are not much like his beliefs in the actual world. And if they're not, then why think his option-counterfactual choices constitute any kind of infallible evidence about what's going on inside his head in the actual world?

Here's a more precise way to put that point. Assume Typikarl's actual beliefs are given by $\mathcal{B}$. Then I can think of two cases where Revealed Preference might be considered plausible. The first would be if we have good reasons to hold that Typikarl's beliefs in the counterfactual scenario will be given by $\mathcal{B}$ conditionalised on $(\underline{b}_{83} \cup \underline{b}_{84})$. Supposing there's no difference in his intrinsic values, then $\succsim$-Maximisation and Revealed Preference would *both* predict that $\underline{b}_{83}$ is chosen in the counterfactual scenario. But what Typikarl knows in the counterfactual is not (only) that he will choose $\underline{b}_{83}$ or $\underline{b}_{84}$; rather, it's the much stronger proposition $\{\underline{b}_{83}, \underline{b}_{84}\}$ *are the only available options*. And there's a very big difference between coming to believe $\underline{b}_{83} \cup \underline{b}_{84}$ under the supposition that he could have chosen any of $\underline{b}_1$ through to $\underline{b}_{100}$, versus coming to believe that $\underline{b}_{83}$ and $\underline{b}_{84}$ exhaust his choices. The latter presumably requires a major rethinking of the causal structure of the world, so his counterfactual beliefs aren't likely to be very similar to $\mathcal{B}$ conditionalised on $\underline{b}_{83} \cup \underline{b}_{84}$ at all—in which case we've no compelling reason to think that how Typikarl chooses between $\underline{b}_{83}$ and $\underline{b}_{84}$ in the counterfactual will necessarily reflect his preferences in the actual world.

(An exception: if Typikarl is certain that $\underline{b}_{83} \cup \underline{b}_{84}$ *just in case* $\{b_{83}, b_{84}\}$ *are the only available options*, then conditionalising on $\{b_{83}, b_{84}\}$ *are the only available options* will give the same result as conditionalising on $\underline{b}_{83} \cup \underline{b}_{84}$. But it's no implication of any of the principles in §1.2 that Typikarl would believe any such thing; nor for the reasons given is it plausible to say that he should. And that matters because if Revealed Preference is plausible only for *some* systems of belief, then we'd need to know that Typikarl's beliefs are *already* before we knew whether his option-counterfactual choices 'reveal' his actual system of preferences. For the same reason, we cannot just stipulate as part of the scenario that Typikarl's beliefs are given by $\mathcal{B}$ conditionalised on $(\underline{b}_{83} \cup \underline{b}_{84})$ without rendering the method useless for the project of radical interpretation—for how do we know what physical states constitute *that* scenario without having fixed the correct scheme of interpretation?)

A second possibility would be to posit that, when making decisions, Typikarl doesn't just decide what option is best and then try to make it true—rather, he formulates a *plan* based on his ranking of the options and their availability. That is, where he starts off believing his options are $\{\underline{b}_1, \ldots, \underline{b}_{100}\}$, what he really decides is not *do* $\underline{b}_1$, but to follow this plan:

if $\underline{b}_1$ is available, then I'll do $\underline{b}_1$; and
if $\underline{b}_1$ is not available and $\underline{b}_2$ is, then I'll do $\underline{b}_2$; and
if neither $\underline{b}_1$ nor $\underline{b}_2$ are available and $\underline{b}_3$ is, then I'll do $\underline{b}_3$; and…

This would entail $\succsim$-Maximisation, and moreover it entails something very close to Revealed Preference. Specifically, it entails that we consider the counterfactual scenario where $\{\underline{b}_{83}, \underline{b}_{84}\}$ are Typikarl's only options *and* he's in the same total physical state S that he's actually in (and hence has the same beliefs and desires, whatever they may be), then *assuming he sticks with his plans* he'll first try $\underline{b}_1$ and find that that's unavailable, then try $\underline{b}_2$, and so on, until eventually he succeeds with $\underline{b}_{83}$. This would avoid the problems above—there's no need to posit magical access to which options are available, and it accommodates belief change without prior knowledge of Typikarl's beliefs. Nevertheless, all we've done here is trade one implausible hypothesis for two.

The first problem is that we clearly *don't* formulate plans that cover every possible way an option set could be constructed out of the relevant domain **D**. If I think my *available* options are given by $\{\underline{b}_1, \ldots, \underline{b}_{100}, \underline{b}_\varnothing\}$, then I might *at most* formulate the plan to do one of my top three or four of those depending on what's available. The probability that *each* of $\underline{b}_1$ through $\underline{b}_{82}$ are unavailable is vanishingly small, so there's no reason to come up with a plan that covers that eventuality—and still less is there any reason to make a plan for the eventuality wherein my options are given by two propositions in **D** that I *don't* believe I can make true. The second problem is that if I *were* to find out the vast majority of options I *thought* were available really aren't, and so I was terribly wrong about what kind of world I'm in, then it's highly unlikely I'd stick to my original plans!

So one way or another, Revealed Preference and its nearby relatives face a problem: to make sense of the change in behaviour in counterfactual circumstances we need to posit a change in beliefs, but the extent of the change required obviates any immediate connection between the counterfactual choices made and actual preferences.

4. Conclusion

Representation theorems have long been taken to support an approach to the project of radical interpretation that's closely related to, but distinct from, Lewis' own—one according to which it's possible to characterise what it is for someone like Karl to have (graded) beliefs and desires wholly in terms of his choices and choice dispositions. Ramsey and Savage developed their own theorems partly in service of this idea, as many others have also done. More recently, representation theorems have played a starring role in some interpretivist theories (e.g., Maher 1993), where they're taken to show that it's possible to assign belief-desire interpretations to agents by rationalising their choice dispositions without fear of radical underdetermination. Others still have suggested that the theorems provide the foundations for a functionalist theory of graded beliefs in terms of their relationship with choices within contemporary decision theory (e.g., Cozic and Hill 2015).

In all the above, representation theorems serve as a kind of proof of the principle that it's possible to determine an agent's beliefs and desires if you know enough about her choice dispositions. I've argued that this approach has no legs. If we understand 'choice dispositions' in terms of evidence-counterfactuals—as

I think we should—then they significantly underdetermine preferences; and if we understand them in terms of option-counterfactuals, then they determine preferences only under clearly implausible assumptions.

It's not all doom and gloom for the representation theorems. If your philosophical project is to reduce some intentional states to other intentional states, then a result like Representation Theorem might be quite useful indeed. For the same reason, functionalists might find them interesting for what they tell us about the internal psychological relationships between beliefs, intrinsic values, and preferences. But the target of our discussion has been the relevance of representation theorems for Lewis' underdetermination argument, and on that front things look less optimistic. Theorems like Ramsey's and Savage's simply don't tell us how it's possible to determine an agent's beliefs and desires given her choice dispositions under any plausible functionalist theory, so they won't save the Bayesian functionalist from Lewis' underdetermination argument.

References

Abdellaoui, M. and P. Wakker (2005). The Likelihood Method for Decision under Uncertainty. *Theory and Decision 58*(1), 3–76.

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

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. (1973). Radical interpretation. *Dialectica 27*(3-4), 313–328.

Dietrich, F. and C. List (2013). Where do preferences come from? *International Journal of Game Theory 42*(3), 613.

Elliott, E. (2017). Probabilism, Representation Theorems, and Whether Deliberation Crowds out Prediction. *Erkenntnis 82*(2), 379–399.

Elliott, E. (2019). Betting Against the Zen Monk: On Preferences and Partial Belief. *Synthese*, 1–26.

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

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision 13*(2), 139–199.

Gaifman, H. and Y. Liu (2018). A simpler and more realistic subjective decision theory. *Synthese 195*(10), 4205–4241.

Godfrey-Smith, P. (2005). Folk Psychology as a Model. *Philosopher's Imprint 5*(6), 1–16.

Hausman, D. M. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy 16*(01), 99–115.

Hausman, D. M. (2012). *Preference, Value, Choice, and Welfare*. New York: Cambridge University Press.

Hedden, B. (2012). Options and the subjective ought. *Philosophical Studies 158*, 343–360.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford University Press.

Jackson, F. and P. Pettit (1990). In Defence of Folk Psychology. *Philosophical Studies 59*, 31–54.

Jeffrey, R. C. (1965). *The Logic of Decision*. Chicago: University of Chicago Press.

Jeffrey, R. C. (1968). Probable knowledge. *Studies in Logic and the Foundations of Mathematics 51*, 166–190.

Jeffrey, R. C. (1983). Bayesianism with a human face. *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science 10*, 133–56.

Joyce, J. M. (1999). *The foundations of causal decision theory*. New York: Cambridge University Press.

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. (1980a). Mad Pain and Martian Pain. In *Philosophical papers*, Volume 1, pp. 122–130. New York: Oxford University Press.

Lewis, D. (1980b). A Subjectivist's Guide to Objective Chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, pp. 263–293. Berkeley: University of California Press.

Lewis, D. (1981). Causal Decision Theory. *Australasian Journal of Philosophy 59*(1), 5–30.

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'. In *Philosophical Papers: Volume 1*, pp. 119–121. New York: Oxford University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Cambridge University Press.

Lewis, D. (1994). Reduction of Mind. In S. Guttenplan (Ed.), *Companion to the Philosophy of Mind*, pp. 412–431. Blackwell.

Lewis, D. (1996). Desire as Belief II. *Mind 105*(418), 303–313.

Maher, P. (1993). *Betting on Theories*. Cambridge: Cambridge University Press.

Maibom, H. (2009). In defence of (model) theory theory. *Journal of Consciousness Studies 166*(8), 360–378.

Pautz, A. (2013). Does Phenomenology Ground Mental Content? In U. Kriegel (Ed.), *Phenomenal Intentionality*, pp. 194–234. Oxford: Oxford.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and Other Logical Essays*, pp. 156–198. London: Routledge.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Dover.

Schwarz, W. (2012). Representation theorems and the indeterminacy of mental content. https://www.umsu.de/blog/2012/580 [Accessed: 20/06/20].

Schwarz, W. (2014). Against Magnetism. *Australasian Journal of Philosophy 92*(1), 17–36.

Schwarz, W. (2015). Analytic Functionalism. In *A Companion to David Lewis*, pp. 504–518. John Wiley & Sons.

Schwarz, W. (forthcoming). Objects of Choice. *Mind*.

Sen, A. (1971). Choice Functions and Revealed Preference. *The Review of Economic Studies 38*(3), 307–317.

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica 40*(159), 241–259.

Sobel, J. H. (1983). Expected utilities and rational actions and choices. *Theoria 49*, 159–183.

Sobel, J. H. (1986). Notes on decision theory: Old wine in new bottles. *Australasian Journal of Philosophy 64*(4), 407–437.

Thoma, J. (Forthcoming). In Defence of Revealed Preference Theory. *Economics & Philosophy*.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman & Hall.

Weatherson, B. (2012). The Role of Naturalness in Lewis's Theory of Meaning. *Journal for the History of Analytic Philosophy 1*(10), 1–19.

Weirich, P. (2004). *Realistic Decision Theory*. Oxford: Oxford University Press.

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