View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilPapers

Representation Theorems and Radical Interpretation

Edward Elliott*

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

Abstract

This paper begins with a puzzle regarding Lewis' theory of radical interpretation. On the one hand, Lewis convincingly argued that the facts about an agent's sensory evidence and choices will always underdetermine the facts about her beliefs and desires. On the other hand, we have several representation theorems—such as those of (Ramsey 1931) and (Savage 1954)—that are widely taken to show that if an agent's choices satisfy certain constraints, then those choices can suffice to determine her beliefs and desires. In this paper, I will argue that Lewis' conclusion is correct: choices radically underdetermine beliefs and desires, and representation theorems provide us with no good reasons to think otherwise. Any tension with those theorems is merely apparent, and relates ultimately to the difference between how 'choices' are understood within Lewis' theory and the problematic way that they're represented in the context of the representation theorems. For the purposes of radical interpretation, representation theorems like Ramsey's and Savage's just aren't very relevant after all.

Introduction

Karl is an ordinary human being, with ordinary human beliefs and ordinary human desires. Our task is to work out what Karl's 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. On the other hand, we are 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 only that much as our starting point, we're to derive what we can of the facts about Karl's beliefs and desires. Call this the project of *radical interpretation*.[1]

There's no shortage of views on how we might approach this project, but I happen to think that the kind of strategy put forward by Lewis in 'Radical Interpretation' (1974) and developed in later works (1979; 1980a; 1983a; 1983b; 1986; 1994) is *basically* on the right track. I'll say more the Lewisean view in due course, but for now let's just say that it rests on two main ideas: that beliefs and desires can be characterised by reference

*E.J.R.Elliott@leeds.ac.uk. Draft of February 12, 2021.

[1] In (Lewis 1974), the project of radical interpretation also relates to meanings in language. We won't be interested in that part of the project in this paper.

to their typical roles within folk psychology, especially in relation to sensory evidence and choices; and that folk psychology is more or less *Bayesian* in character—beliefs and desires come in degrees, choices are determined by something like the rule of expected utility maximisation, and learning works by something like the rule of conditionalisation.

From the beginning, though, Lewis held that substantive constraints on the contents of our beliefs and desires would be required for any adequate functional definition of those states. The clearest argument he gives for this is in 'New Work for a Theory of Universals' (1983a: 373–5; see also 1986: 36ff, 105ff). Assume that Karl is indeed as the folk psychological theory describes him—an expected utility maximiser who conditionalises on his sensory evidence. From this it follows that, whatever the facts about Karl's life history of evidence and choices might be, there will be at least one way of assigning beliefs and desires to him that *fits* those facts perfectly. However, Lewis argues, if there are no substantive constraints on what kinds of beliefs and desires Karl can have, then there will be a great many competing interpretations as well. Mere fit with the evidence-and-choice facts radically underdetermines the belief-and-desire facts. Lewis' solution was to cut some systems of belief and desire from the running: only the most *eligible* systems of belief and desire are genuinely possible.

But now you might be wondering whether this 'eligibility' solution was really necessary. After all, don't we have a number of representation theorems for expected utility theory—such as those of Ramsey (1931) and Savage (1954)—which are widely taken to show that if an agent's choice patterns satisfy certain axioms, then there's a *unique* system of graded beliefs and desires under which those choice dispositions maximise expected utility? And if that's what these theorems really do tell us, then if Karl really is an expected utility maximiser, it follows that *if* his choices satisfy those axioms, *then* the facts about his choices alone could in principle suffice to determine his beliefs and desires. So much for Lewis' underdetermination claim!

Indeed, these theorems have long been taken to support a certain kind of approach to radical interpretation, one that emphasises the importance of choices in particular. Buchak (2013: 83ff) describes this as the 'interpretive use' of representation theorems, while Meacham and Weisberg (2011) refer to it as 'Characterisational Representationalism'. But whatever we choose to call it, the usual idea is that a person's (graded) beliefs and desires *just are* those under which her choice dispositions maximise expected utility, at least in the event that they satisfy the appropriate axioms. Ramsey and Savage developed their own theorems partly in the service of this idea, and in the years since theorems like theirs have played a starring role in several 'interpretivist' theories of graded belief (e.g., Maher 1993; see also Bermúdez 2009). Others still have suggested that representation theorems provide us with the foundations for defining graded beliefs in terms of their functional relationships with choices according to decision theory (e.g., Cozic and Hill 2015).

So we have what appears to be a conflict. Assume for the sake of argument that Karl is a rational Bayesian agent. Lewis then tells us that there will be radically distinct systems of belief and desire that equally fit the facts about Karl's evidence and about his choices, whatever those facts may be; *a fortiori*, the facts about his choice dispositions alone must underdetermine the facts about his beliefs and desires. But then there's the representation theorems, which are widely taken to show that under the right conditions the facts about Karl's choices alone can in principle be enough to determine the facts about his beliefs and desires—indeed we might not even need to consider his life history of evidence. What should we make of all this?

That's the set-up, now here's where I'm going with it: Lewis was not wrong, but then neither are the representation theorems. There is no genuine conflict between Lewisean underdetermination and theorems like Ramsey's and Savage's—the appearance to the contrary is the result of a widespread misunderstanding about what those theorems can plausibly tell us about the relationships between choices, beliefs, and desires. As we will see, the key to reconciliation lies in the difference between how 'choice dispositions' are understood in Lewis' theory, and the problematic way they're treated by 'interpretive uses' of representation theorems. More importantly, once we see why Lewisean underdetermination is consistent with the representation theorems, we can learn a lot not only about Lewis' approach to radical interpretation, but also about the relevance of representation theorems to the project of radical interpretation more generally—if I'm right, then it's a mistake to think that representation theorems can be used 'interpretively'.[2]

Let me be as clear as I can: my goal here is not to defend Lewis' underdetermination claim against an objection arising from representation theorems. Lewis does not need my defence. Rather, the point of the paper is to address two questions about the relationship between representation theorems and radical interpretation:

(i) How do these theorems relate to Lewis' underdetermination result?
(ii) Are there *any* plausible theories of radical interpretation under which these theorems show that beliefs and desires are (sometimes) determined by choice dispositions?

My answer to the former is 'they don't really', my answer to the latter is 'probably not', and I come by my answer to the latter by way of considering the former.

The remainder of the paper is divided into four main parts. I will start with an outline of Lewisean functionalism in §1, and then in §2 I will present the underdetermination argument. In §3 I will describe how representation theorems are consistent with Lewis' argument. In this section I will also describe why pre-existing discussions on this matter—such as those in (Schwarz 2012; 2014) and (Williams 2016)—are insufficient. Finally, in §4 I describe why the 'interpretive use' of representation theorems is problematic, and why they are therefore ultimately of little direct relevance to the project of radical interpretation. (There is also a short Appendix, which addresses several small issues with Lewis' underdetermination argument that are tangential to the main thread of discussion.)

1. Lewisean Functionalism

I will start with some 'big picture' matters on Lewis' functionalism and its relationship to Bayesianism (§1.1), after which I'll outline in more detail what a Bayesian functionalist theory ought to look like (§1.2–§1.3).

1.1 The big picture

Our task as radical interpreters is to identify states of belief and desire with the physical states that constitute our world. We can do this by specifying a *scheme of interpretation*—that is, a function from physical states $S, S', \ldots$ (e.g., brain states) to the mental states $M, M', \ldots$ with which they're to be identified (cf. Lewis 1983b: 119).

[2] So there's no confusion: in this paper we're interested in representation theorems *for expected utility theory*, which are *like Ramsey's and Savage's*. By this I mean exactly those theorems that state sufficient conditions under which a binary relation can be given an expected utility representation, and one that is unique (or at least unique up to some non-trivial condition). It should go without saying that there might be a representation theorem *of some form or other* that could be relevant to radical interpretation—it would be quite amazing if there weren't! But, rather than saying 'theorems like Ramsey's and Savage's' over and again, from now on I'll just say 'representation theorems'.

The functionalist about beliefs and desires has a certain methodology for arriving at the correct scheme(s) of interpretation. She starts with some theory of intentional psychology, call it $\mathcal{T}$, which tells us

i) what kinds of contents it is possible to believe and/or desire; and
ii) how states of belief and desire relate to one another, other relevant mental states, and the world (e.g., via the senses and behaviour).

The idea is to use this theory $\mathcal{T}$ to determine the best scheme (or schemes) of interpretation. To the extent possible, such a scheme

i) should assign mental states that are *eligible* for assignment according to $\mathcal{T}$; and
ii) it should *fit* in the sense that if $S, S', \ldots$ are assigned to $M, M', \ldots$ respectively, then the $S, S', \ldots$ relate to one another and to the wider world in the same way that the $M, M', \ldots$ relate to one another and to the wider world according to $\mathcal{T}$.

A scheme is considered *better* to the extent that it assigns eligible interpretations and maximises fit, and we say that the *correct* scheme is whichever is *best*. If $\mathcal{T}$ is true, then at least one scheme is guaranteed to assign eligible contents and to have perfect fit; otherwise, we make do with what we can get. If there are multiple schemes tied for equal best, then the usual thing is to say that truth is indeterminate between them (e.g., Lewis 1983b: 120).

Given this very schematic characterisation, we can distinguish several more specific varieties of functionalism by reference to characteristics of the underlying theory $\mathcal{T}$. The first 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, it 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. The idea, in other words, is to define beliefs and desires in terms of their functional roles within the best systematisation of folk psychology properly so-called. (Cf. Jackson 1998 on functionally defining moral concepts using hypothetical future systematisations of folk moral theories.)

Next up is a style of functionalism that Lewis advocated from at least (1980a) onwards (see also 1983b; 1986: 39–40):

Anti-Individualist Functionalism: $\mathcal{T}$ is about typical agents.

That is: the theory places no specific constraints on properties and relations that any individual's beliefs and desires will have, but rather on how we might expect beliefs and desires to typically be.[3] Functional roles are typical roles, so a scheme of interpretation identifies mental states with physical states on the basis of how the latter typically behave. 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 in Karl's particular case S behaves anything at all like M typically behaves.

Finally, we have:

[3] In (Lewis 1980a; 1983b; 1986: 39–40), 'typicality' is understood statistically. So, roughly, a state S will be interpreted as a certain system of beliefs if, across the actual and possible individuals in which S recurs, S tends to have the kinds of causes and effects we associate with that system of beliefs. I'm not sure this is the best way to understand the relevant sense of 'typicality', but what I have to say won't depend very much on the matter so I'm going to leave it ambiguous.

Bayesian Functionalism: $\mathcal{T}$ is more or less Bayesian in character.

In Lewis' case, Bayesian functionalism is a consequence of his commitment to analytic functionalism, given that any good systematisation of folk psychology 'should look a lot like Bayesian decision theory' (1979: 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: 337–8)

Not exactly like orthodox Bayesian theory, mind you, but not too far from it either. Lewis clearly thought that the more extreme idealisations of Bayesianism would be absent from any final systematisation of folk psychology (cf. 1983a: 375; 1986: 30; 1994: 428), and we know that some of these idealisations can be weakened without fundamentally changing the recognisably *Bayesian* character of the resulting theory (e.g., Jeffrey 1983; Walley 1991; Weirich 2004; Bradley 2017; Elliott 2017b). With that said, three potential misunderstandings are worth briefly addressing.

First, ours is a theory of the *typical* agent. Like the rest of us, Karl will be atypical to some degree in some relevant respect at least some of the time. So we should expect that he will deviate from Bayesian norms to some degree—perhaps even to a very significant degree (Lewis 1983b; 1994: 428).

Second, not all idealisations are inherently problematic: *any* psychological theory will sacrifice some realism for greater generality and overall simplicity, and there's no good reason to think that the best systematisation of folk psychology would be any different. The *truth* of Bayesianism (*qua* theory of the typical agent) is not a precondition for the success of Bayesian functionalism. Close enough is good enough. So we can live with a bit of unrealism, especially if that's the price we pay for greater generality and simplicity.

Finally, you might worry that folk psychology cannot be much like Bayesianism, since Bayesianism is obviously not 'platitudinous'. But we do not require that our systematisation of folk psychology be constituted from platitudes that members of the folk would themselves spontaneously *assert*, or even *readily recognise* as things that they themselves tacitly believe. For one thing, the expression of a theory may be couched in technical jargon with which average members of the folk may be unfamiliar (Lewis 1974: 338). More importantly, the folk's tacit understanding of intentional psychology can be likened to our tacit understanding of grammar, involving complicated rules and principles that need not be apparent even to those who unfailingly adhere to them (Lewis 1994: 416; cf. Jackson and Pettit 1990: 33–6).

To summarise, then, to be a *Lewisean functionalist* about beliefs and desires is to be an anti-individualist analytic Bayesian functionalist. The Lewisean functionalist proposes that our folk notions of belief and desire will refer to those states, if any there are, which come close enough (and closer than anything else) to satisfying the functional roles associated with the doxastic and conative mental states posited by the best systematisation of our shared implicit understanding of typical intentional psychology; and, furthermore, she'll think that this systematisation will look a lot like contemporary Bayesianism. The Bayesian theory we use to characterise the functional roles of our beliefs and desires need not be perfectly accurate, and indeed we ought to expect it will simplify and idealise over many of the messy details, but it is assumed that it will do a fairly good job on average at least for typical agents.

1.2 A model Bayesian

Our topic to start with concerns what the Lewisean functionalist ought to say about underdetermination and representation theorems. It will help if we have a specific model of a typical agent to serve as a fixed point of reference for the ensuing discussion.

With that in mind, let me now introduce *Typikarl*. Typikarl is perfectly typical with respect to (i) the structure of his beliefs at a time, (ii) the structure of his desires at a time; (iii) how his beliefs and desires change over time; and (iv) the relationship between his beliefs, desires, and choices. Karl is *not* Typikarl. Like the rest of us, Karl will be atypical in at least some of these 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, has been and always will be so uncompromisingly *typical*. Typikarl is a fiction, but he will be a useful fiction for describing the typical functional roles of belief and desire.

My description of Typikarl will be based on a simplified 'Jeffreyan' conception of decision-making (see Jeffrey 1965). Some of the arguments that follow—including the underdetermination argument—will make use of the 'Jeffreyan' aspects, and I'll say more about that when the time comes. I emphasise, though, that my goal here is to describe a *model*. We are not aiming for the ultimate systematisation of folk psychology, but a simplified formal system that's 'close enough' for drawing specific conclusions about underdetermination. I discuss some of the simplifications in the Appendix.

Beliefs at a time

We'll start by describing Typikarl's beliefs and desires at a time. Any vaguely Bayesian theory will require at minimum that beliefs and desires are in some sense *graded*, and furthermore that they're each coherent enough to permit the comparison of expected values (Lewis 1974: 337).

I will assume that propositions are subsets of some set of possible worlds, Ω. For simplicity, I'll also assume that Ω is *finite*; consequently there are only finitely many propositions with each belonging to $\wp(\Omega)$, the powerset of Ω. Also for simplicity, I will assume that Typikarl has beliefs regarding every proposition. Given that, a nice and familiar way to ensure a minimally coherent system of beliefs is to assume:

> $\mathcal{B}$-Eligibility: Typikarl's beliefs at a time can be represented by a function $\mathcal{B} : \wp(\Omega) \mapsto \mathbb{R}$, which obeys the laws of probability.[4]

$\mathcal{B}$-Eligibility sets a minimum bound on our eligibility constraints: only those systems of belief that are representable by a probability function are eligible for assignment. Note also that we're taking the probability function to capture the *content* of a system of beliefs as a whole. We are thus understanding 'content' more broadly than it's sometimes used—the 'content' of a system of beliefs is not a proposition, but a space of propositions plus a measure of the strengths with which each is believed.

Desires at a time

With regards to Typikarl's desires, we'll want to start with his *intrinsic values*. We can think of his intrinsic values as that aspect of his overall conative state that's independent of his beliefs, to be contrasted with that aspect of his conative state which is influenced by his beliefs (e.g., by something like means-ends reasoning).

Within the Jeffreyan framework, intrinsic values will be most easily represented by a distribution of real numbers over the worlds in Ω:

[4] That is: $\mathcal{B}(\Omega) = 1$, and for all p, q, $\mathcal{B}(p) \geq 0$ and if $p \cap q = \varnothing$ then $\mathcal{B}(p \cup q) = \mathcal{B}(p) + \mathcal{B}(q)$.

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

You might usefully read $\mathcal{V}$ as representing the relative strength with which TypiKarl desires that the actual world is ω, for each world ω in Ω.[5]

We can now talk about Typikarl's desires more generally, which will be a function of his beliefs and his intrinsic values. Say first that '$\mathcal{B}^p$' designates $\mathcal{B}$ *conditionalised on* p; that is, for any 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. We can then say that the strength with which Typikarl desires a given proposition p is just the $\mathcal{B}^p$-weighted average $\mathcal{V}$-value of the worlds that constitute p:

$$\sum_{\omega \in p} \mathcal{B}^p(\{\omega\})\mathcal{V}(\omega).$$

However, for our purposes all that really matters are Typikarl's relative desires—his *preferences*. So, if from now on we use '$p \succsim q$' to mean that Typikarl desires p at least as strongly as he desires q, then we have:

$\succsim$-Coherence: If Typikarl has beliefs $\mathcal{B}$ and intrinsic values $\mathcal{V}$, then $p \succsim q$ just in case

$$\sum_{\omega \in p} \mathcal{B}^p(\{\omega\})\mathcal{V}(\omega) \;\geq\; \sum_{\omega \in q} \mathcal{B}^q(\{\omega\})\mathcal{V}(\omega).$$

Changes over time

Now for how Typikarl changes over time. We'll assume, as Lewis usually did, that Typikarl updates by conditionalising on his sensory evidence (cf. Lewis 1980b: 288; 1983a: 374; 1994: 428–9). To make this precise, let me introduce some more notation and a couple of background assumptions.

First, we'll let

$$\boldsymbol{\tau} \;=< \tau_1, \ldots, \tau_n >$$

be an ordered set of times from the beginning of Typikarl's life onwards. Then, we let

$$\mathbf{E} \;=\; \{e_1, \ldots, e_n\}$$

designate the set containing all and only those propositions which characterise—in all relevant detail—all possible 'streams' of sensory evidence from any time to any later time; for example, *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.

Purely for technical convenience, we'll assume that $\mathbf{E}$ includes Ω, the 'trivial' evidence. (This merely helps to simplify some definitions below.) If we then say that $\mathcal{B}$ is *consistent with* e just in case $\mathcal{B}(e) > 0$, then we will assume that Typikarl's *initial* system of beliefs is consistent with every $e \in \mathbf{E}$. That is, Typikarl doesn't rule out any

[5] As standard, I'll assume that strengths of desire are measurable on nothing stronger than an interval scale. Consequently, if $\mathcal{V}$ and $\mathcal{V}'$ are related by an interval-preserving transformation (i.e., if there's an $\mathrm{x} > 0$ and a y such that for all ω, $\mathcal{V}(\omega) = \mathrm{x}\mathcal{V}'(\omega) + \mathrm{y}$), then I will presume that $\mathcal{V}$ and $\mathcal{V}'$ represent the very same intrinsic values, and I won't be fussed about distinguishing between them.

possible life history of sensory evidence prior to having had any experiences whatsoever. (This just makes the underdetermination argument go through a bit smoother.) With all that in place, the main driving force for psychological change is:

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

We also have to describe how Typikarl's desires change over time. Since Typikarl's beliefs will generally change as a result of his evidence, so too will his preferences tend to shift and change about as time goes on. On the other hand, I will also assume:

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

This assumption is implicit in a number of Lewis' works (e.g., 1974: 336–7; 1980b: 288; 1983a: 374–5), and it is implicitly relied upon for his underdetermination argument. In his (1983a: 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 certainly not that they will *often*. In any case, we stick with Static Values for now. (I discuss the implications of denying it in the Appendix.)

Choice dispositions

Since Typikarl's beliefs $\mathcal{B}$ and values $\mathcal{V}$ at any time jointly determine his total system of desires at that time, from now on we'll represent Typikarl's *total belief-desire state at a time* using pairs of the form $(\mathcal{B}, \mathcal{V})$. The only thing left is to describe how Typikarl's total belief-desire state at a time relates to his choices at that time.

It's clear enough that if Typikarl were able to make any proposition whatsoever true, and he knew this, then he would simply make it so that the actual world is whatever world he values most. But Typikarl does not have magical powers, and in fact he probably has no direct influence over the truth or falsity of the vast majority of propositions that he's able to contemplate. So, before we can say anything very specific about Typikarl's choice dispositions, we first need to fully characterise his *options*. And that's something I'm *not* going to do, 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, we will follow Lewis (1974: 337; 1994: 416–7) in treating Typikarl's options at any time as a partition of propositions specifying how he behaves at that time.[6] Given this, an initial thought for how we might model Typikarl's options at a time τ would be to use the partition

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

that captures in maximal detail each the specific ways b that Typikarl behaves at τ. But this won't work, since Typikarl needn't have preferences over all of the b in $\mathbf{B}^\tau$. In particular, if $\mathcal{B}(b) = 0$, then b has no location within Typikarl's preference ranking—and if we didn't have the simplifying assumption that Typikarl has beliefs regarding *all* propositions, then it could well turn out that *none* of the b in $\mathbf{B}^\tau$ find a place in his preference ranking.

[6] Arguably, agents decide between *actions*, not *behaviours*. If so, then this would present some potential problems for the Lewisean approach to radical interpretation (cf. Williams 2019). But, with the appropriate minor adjustments, Lewis' underdetermination argument goes through either way, and the points I raise in §3 and §4 are essentially unchanged if we replace all talk of 'behaviours' with 'actions'. So let's set this aside as a matter to be dealt with elsewhere.

More generally, we've yet to consider the 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: 170)

Something similar has been advocated by Sobel (e.g., 1983)—roughly, p counts as an option only if the decision-maker is certain she can make p true by an act of will. Related conditions are discussed in (Hedden 2012) and (Schwarz forthcoming), and important connections between options and beliefs are discussed in (Hausman 2000; 2012: 27ff).

I do not know whether we should want anything as strong as Sobel's constraint or Jeffrey's 'actuality' constraint, but here's not the place to decide the issue. What's important for our discussion is that there are plausibly *some* conditions on what counts as an option and that these depend at least in part on what decision-makers *believes* they can do. A minimal constraint, I take it, is that Typikarl won't be disposed to choose something that's impossible by his lights. So we should say that p counts as an option only if $\mathcal{B}(p) > 0$. Furthermore—and this is the really important point—any proposition about his behaviour that *does* count as an option at τ will be equivalent to a disjunction of the more specific behaviour-propositions in $\mathbf{B}^\tau$. (This will be relevant for the underdetermination argument in §2, and for some of the arguments noted in §3.)

Given this, we can say at a minimum that if Typikarl has beliefs $\mathcal{B}$ at τ, then his options at τ will be set of a mutually exclusive and jointly exhaustive propositions about his behaviour at τ,

$$\mathbf{O}^\tau_\mathcal{B} = \{o_\varnothing, o_1, ..., o_n\}$$

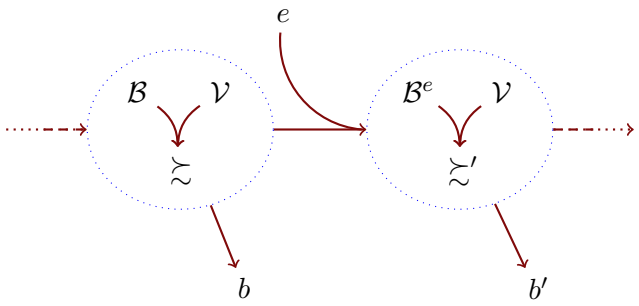
where what we'll call the *null option*, $o_\varnothing$, is the largest union of the b in $\mathbf{B}^\tau$ such that $\mathcal{B}(o_\varnothing) = 0$. It won't be located anywhere in Typikarl's preference ranking. The remaining non-null options o will each be equivalent to some union of b in $\mathbf{B}^\tau$, and will satisfy at least the minimal condition $\mathcal{B}(o) > 0$.[7] We now have everything we need:

> $\succsim$-Maximisation: If Typikarl has beliefs and desires $(\mathcal{B}, \mathcal{V})$ at τ, then he will make true some b in $\mathbf{B}^\tau$ that entails one of the options in $\mathbf{O}^\tau_\mathcal{B}$ that he desires most.

1.3 Fitness

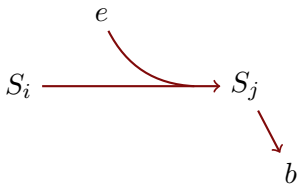
To summarise: Typikarl's beliefs and intrinsic values determines his desires, and his total system of beliefs and desires at a time determines his choices and hence how he's disposed to behave at that time. Each system of beliefs and desires is also poised to give rise to new systems of belief and desire by conditionalising on sensory input, and each such system might itself come about from some earlier systems conditionalised on the appropriate (possibly trivial) evidence.

[7] I will assume henceforth that there's always a unique such $\mathbf{O}^\tau_\mathcal{B}$ that characterises Typikarl's options at τ, given that his beliefs at τ are $\mathcal{B}$. What I've said so far in no way licenses this assumption, and it may well be false. But it isn't used in the underdetermination argument, and it only plays a simplifying role in the arguments of §3 and §4—so I'm not worried about taking it for granted here.



The only thing left is to say what a scheme of interpretation is, and what it is for a scheme to maximise *fit*. It's evident from how Lewis defines his 'only constraining principle of fit' in 'New Work' (1983a: 374) that he was understanding schemes of interpretation as functions from *total* momentary physical states to *total* systems of belief and desire $(\mathcal{B}, \mathcal{V})$, and we'll follow him in this regard. (The Appendix discusses the implications of alternative ways of characterising schemes of interpretation.)

'Fitness' will need to be defined recursively, so start by saying that 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 brings about b, where b entails some option o that maximises expected value with respect to $(\mathcal{B}^e, \mathcal{V})$. In visual form, the total momentary physical state S_i matches $(\mathcal{B}, \mathcal{V})$ whenever, for all relevant evidence-specifying e,



where b entails o and o maximises expected value 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 behaviours that maximise expected value 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, since we've said nothing to guarantee that the state S_j that S_i yields given e will itself match $(\mathcal{B}^e, \mathcal{V})$. We need also that the interpretations are *aligned* in the appropriate way in light of the causal relationships between the states to which they're assigned; thus,

FIT. A scheme of interpretation $\mathcal{I}$ *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 $\mathcal{I}(S_j) = (\mathcal{B}^e, \mathcal{V})$; and
(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 $\mathcal{B}_i^e = \mathcal{B}$.

FIT gives us an ideal. Imperfect fit can be cashed out in terms of how close a scheme comes to satisfying this ideal. As far as the underdetermination argument is concerned, we assume that typical agents do indeed satisfy the Bayesian theory I've just outlined, so at least one perfectly fitting scheme is guaranteed to exist.

2. The Underdetermination Argument

My cards are on the table: with the exception of Static Values, I think that the kind of functionalism that I've been describing is *basically* right. Some tweaks are 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 thoroughgoing physicalist terms, we'll also need to say a lot more about how we can get access to the facts about evidence (cf. Pautz 2013; Williams 2019), and how options ought to be characterised in relation to beliefs. In any case, though, I doubt that the main outlines are going to change much. Whatever we end up with after crossing all the t's and dotting all the i's should end up looking a lot like the kind of theory characterised by $\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, $\succsim$-Coherence, $\succsim$-Maximisation, and Conditionalisation.

So that brings us to the underdetermination argument. According to Lewis, if one scheme of interpretation maximises fit then there will always be others, and some of these competing alternatives will differ radically from one another—that is, *unless* we impose stronger eligibility constraints on the contents of beliefs and desires. Lewis only *very* briefly sketches the argument for this in 'New Work', though most of the neglected details have more recently been filled in by Schwarz (2012), Weatherson (2012: 5), and Williams (2016). Here's how it goes.

First, 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 of the times τ) plus any e in $\mathbf{E}$ will be a member of the *algebra of propositions generated by* $\mathbf{G}$, designated $\mathcal{A}(\mathbf{G})$, and defined as the set of all propositions equal to the union of members of $\mathbf{G}$:

$$\mathcal{A}(\mathbf{G}) \;=\; \{p \;\mid\; \exists g_i, ..., g_j \in \mathbf{G} : p = g_i \cup \cdots \cup g_j\}.$$

The upshot is that any disjunction of behaviour-specifying propositions—and hence any option-specifying proposition—as well as any evidence-specifying proposition will belong to $\mathcal{A}(\mathbf{G})$. Given this, the key assumption needed for the underdetermination result is that there are propositions regarding which Typikarl can have non-trivial strengths of belief and/or desire regarding propositions that do not belong to $\mathcal{A}(\mathbf{G})$. Since we've assumed that Typikarl has beliefs regarding every proposition, this amounts to just saying that some members of $\mathbf{G}$ contain more than one possible world.

An example will help to show where the argument goes from here. Suppose that for some g in $\mathbf{G}$, $g = \{\omega_1, \omega_2, \omega_3\}$. Now take some 'decent, reasonable' system of initial beliefs and intrinsic values, $(\mathcal{B}_1, \mathcal{V}_1)$. The specific numbers don't really matter, so suppose:

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

We now 'twist' $\mathcal{B}_1$ into a new probability distribution that assigns the same values to every proposition in $\mathbf{G}$; and we twist $\mathcal{V}_1$ 'in a countervailing way', so as to end up with the same expected values for every proposition in $\mathbf{G}$. For example,

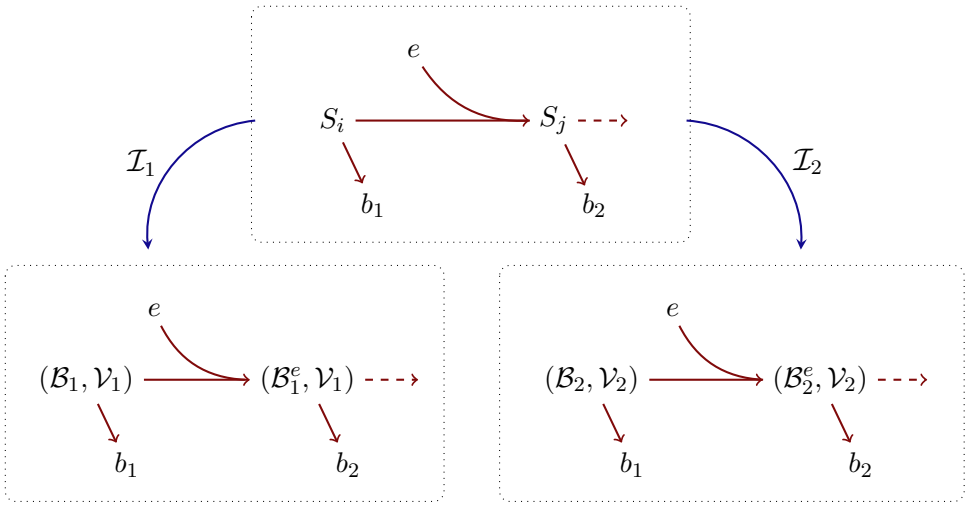
$$\mathcal{B}_2(\{\omega_i\}) = \begin{cases} 0.3, & \text{if } i = 1 \\ 0.3, & \text{if } i = 2 \\ 0, & \text{if } i = 3 \\ \mathcal{B}_1(\{\omega_i\}) & \text{otherwise} \end{cases} \qquad \mathcal{V}_2(\omega_i) = \begin{cases} 7, & \text{if } i = 1 \\ 7, & \text{if } i = 2 \\ 5, & \text{if } i = 3 \\ \mathcal{V}_1(\omega_i) & \text{otherwise} \end{cases}$$

Which gives the same expected value for g:

$$\sum_{\omega \in g} \mathcal{B}_2^g(\{\omega\})\mathcal{V}_2(\omega) \;=\; \sum_{\omega \in g} \mathcal{B}_1^g(\{\omega\})\mathcal{V}_1(\omega) \;=\; 7.$$

More generally, from $\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, and $\succsim$-Coherence, $(\mathcal{B}_2, \mathcal{V}_2)$ will determine the same strengths of belief and desire for all propositions in $\mathcal{A}(\mathbf{G})$. Several things follow from this. First, since $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ must therefore determine the same expected values for any disjunction of behaviour propositions at any time, they will also determine the same preferences over those propositions. So, by $\succsim$-Maximisation, they must determine the same choice dispositions over Typikarl's options (whatever those may be). Furthermore, the same holds even after conditionalising $\mathcal{B}_1$ and $\mathcal{B}_2$ on any evidence-specifying proposition in $\mathbf{E}$—that is: for any such e, $(\mathcal{B}_1^e, \mathcal{V}_1)$ and $(\mathcal{B}_2^e, \mathcal{V}_2)$ will *also* determine the same choice dispositions over Typikarl's options (whatever those may be). So by Conditionalisation plus Static Values, we get the result 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.

Given the definition of Fit, then, and assuming that there are no further constraints on eligibility beyond those given by $\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, it follows that there must be multiple competing schemes of interpretation each assigning perfectly eligible interpretations with equally perfect fit:



This establishes *underdetermination* with respect to any propositions not in $\mathcal{A}(\mathbf{G})$. We get *radical* underdetermination by noting that Typikarl will have beliefs and/or desires regarding numerous propositions not in $\mathcal{A}(\mathbf{G})$—i.e., propositions that are not equivalent to disjunctions of propositions specifying how he behaves and what evidence he receives in what order—and that schemes which agree with respect to $\mathcal{A}(\mathbf{G})$ can vary wildly with respect to propositions not in $\mathcal{A}(\mathbf{G})$.

(Do not misunderstand what Lewis' argument teaches us. It does *not* tell us that the evidence-and-choice facts determine the facts about Typikarl's beliefs and desires up to the propositions in $\mathcal{A}(\mathbf{G})$, but no further. If that were the case, then there would be no further concerns about underdetermination if only $\mathbf{G}$ were maximally fine-grained—for example, if we were to say that any proposition whatsoever could serve as the content of Typikarl's evidence, or if Typikarl really could choose to make any proposition true. But, as we'll see, making $\mathbf{G}$ more fine-grained won't solve the problem. Lewis' argument

establishes that the evidence-and-choice facts *at most* determine Typikarl's beliefs and desires up to those propositions in the algebra generated by **G**. For all we've said so far, though, the evidence-and-choice facts need not determine Typikarl's beliefs and desires, or even his preferences, for those propositions either.)

So how can we fix this? Lewis' proposal was to suggest stronger eligibility constraints than those that are entailed by $\mathcal{B}$-Eligibility and $\mathcal{V}$-Eligibility alone—constraints based on substantive considerations of *reasonableness*:

> 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. (1983a: 375)

Continuing the example, suppose that $(\mathcal{B}_1, \mathcal{V}_1)$ is a more reasonable than $(\mathcal{B}_2, \mathcal{V}_2)$. S matches the interpretation $(\mathcal{B}_1, \mathcal{V}_1)$ just when it matches $(\mathcal{B}_2, \mathcal{V}_2)$. A scheme that assigns $(\mathcal{B}_1, \mathcal{V}_1)$ to all such states will therefore be considered better overall, *ceteris paribus*, than any scheme which assigns $(\mathcal{B}_2, \mathcal{V}_2)$ to these states. The upshot is that the latter can never be assigned to *any* physical state by the correct scheme (or schemes) of interpretation. Some pairs of probability and value functions represent no genuinely possible belief-desire state—they 'can never be correctly assigned because, whenever [they] fit the functional roles of the thinker's states, some more favoured content also fits' (1986: 108).

Lewis thought it likely that, with strengthened eligibility constraints in place, we would be able to sufficiently narrow down the admissible schemes of interpretation. He didn't provide much of an argument for this; nor for that matter did he say very much about what the eligibility requirements were supposed to be. But how exactly the solution is supposed to work and whether it's successful is a topic for another paper. Let's turn now to representation theorems, and whether these have any interesting implications for Lewisean functionalism.

3. Representation Theorems and Lewisean Functionalism

Representation theorems for expected utility theory are often taken to supply us with conditions under which we can derive the facts about an agent's beliefs and desires once 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 four decades since.

I couldn't hope to cover every relevant theorem in any detail here, so here's what I'm going to do. I'll start by describing in a very generalised way the stereotypical representation theorem (§3.1). Then I'll consider the suggestion that Lewisean underdetermination is consistent with these theorems so long as we deny some of the 'act-richness' axioms they rely on (§3.2). Finally, I'll discuss the deeper reason why there's no genuine conflict between the underdetermination argument and the representation theorems (§3.3).

3.1 The stereotypical representation theorem

Any precisification of expected utility theory will do two things. First, it will 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. Second, the theory will say that an agent's preferences $\succsim$ will be determined by her beliefs and intrinsic values according to (some version of) the expected utility rule.

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, and
$\boldsymbol{\pi} = \{\succsim, \succsim', ...\}$ = the possible systems of preference.

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

> Representation Theorem. If a system of preferences $\succsim$ satisfies axioms a_1, ..., a_n, then $\mathcal{T}_{\text{EU}}(\mathcal{B}_i, \mathcal{V}_i) = \mathcal{T}_{\text{EU}}(\mathcal{B}_j, \mathcal{V}_j) = \succsim$ if and only if $(\mathcal{B}_i, \mathcal{V}_i) = (\mathcal{B}_j, \mathcal{V}_j)$.

The axioms $a_1, ..., a_n$ will include, for example, the requirement that $\succsim$ is transitive and complete. I'll say a bit more about some of the axioms in §3.2, but for now we can safely ignore the specifics of what the $a_1, ..., a_n$ actually say.

Ramsey's theorem takes the above form, and Savage's does 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 agents choose whichever of their options they consider best. If we know what the relevant options are, then observing the agent's actual choices at most tells us what she considers best from amongst those. That's a long way off from full information about her preferences, so more is needed.

Well, we know how to solve this one, right? According to a long-standing tradition, counterfactual choices reveal preferences. Specifically,

> Revealed Preference. If Typikarl has beliefs and desires $(\mathcal{B}, \mathcal{V})$, then for any possible set of options $\mathbf{O}$, if his options were $\mathbf{O}$ then he would be disposed to make true any one of the options in $\mathbf{O}$ that he desires most.

In other words, Typikarl's preferences correspond directly to his choice dispositions under counterfactual hypotheses about his available options. Indeed, suppose we define an Typikarl's *counterfactual choice ranking*, $\succsim^c$, as follows: $o_1 \succsim^c o_2$ just in case, if o_1 and o_2 were Typikarl's only options, then o_1 is one of the options he would be disposed to make true. If Revealed Preference is true, then (at least as far as the possible options are concerned) Typikarl's counterfactual choice ranking $\succsim^c$ *just is* his preference ranking $\succsim$, more or less.[8]

3.2 Underdetermination and act-richness

So here's where we're at: if there's an appropriately tight connection between choice dispositions and preferences, of the kind given by Revealed Preference, then it's plausible enough to say that a result like Representation Theorem describes sufficient conditions for when the facts about an agent's choice dispositions can uniquely

[8] Revealed Preference tells us that (for possible options o_1, o_2), $o_1 \succsim o_2$ implies $o_1 \succsim^c o_2$. The other direction is trickier, since $\succsim$ need not be complete. However, at least in Typikarl's case, $\succsim$ is complete over those o such that $\mathcal{B}(o) > 0$; for any remaining possible options, Revealed Preference entails they'll be minimal in $\succsim^c$. Hence, $o_1 \succsim^c o_2$ implies $o_1 \succsim o_2$ if neither o_1 nor o_2 is bottom ranked in $\succsim^c$; otherwise o_1 and o_2 aren't $\succsim$-related. In sum: $\succsim^c$ is $\succsim$, more or less.

determine the facts about her beliefs and desires. And this is exactly the kind of thing that you'd *think* a Bayesian functionalist ought to be very interested in, especially when facing off against a problem of radical underdetermination. There can be no reasonable doubt that Lewis would have been aware of Ramsey's and Savage's theorems by the time he was writing 'New Work'. So why did he never mention them?

Let me with two answers that I don't think get to the heart of the issue. (Readers not interested in the details can skip to §3.3 without loss of comprehension.)

First, as Schwarz (2012) and Williams (2016) have both pointed out, the underdetermination argument in 'New Work' presupposes Jeffrey's (1965) decision theory. And, famously, on Jeffrey's theory preferences do not suffice to determine unique beliefs and desires—instead, Jeffrey's theorem only establishes axioms under which $\mathcal{T}_{\text{EU}}(\mathcal{B}_i, \mathcal{V}_i) = \mathcal{T}_{\text{EU}}(\mathcal{B}_j, \mathcal{V}_j) = \succsim$ if and only if $(\mathcal{B}_i, \mathcal{V}_i)$ and $(\mathcal{B}_j, \mathcal{V}_j)$ are related by a fractional linear transformation. So there's *already* underdetermination of belief and desire by preference in this version of expected utility theory. Lewis' conclusion is nothing new to report.[9]

Now, there's a good sense in which this is all correct: Lewis' conclusion is that there is underdetermination; his argument for this presupposes Jeffrey's theory, and the representation theorem for Jeffrey's theory admits underdetermination; hence, Lewis' conclusion is consistent with the relevant theorem—indeed, it's 'old news'. With that said, and as Schwarz (2012) rightly points out, it clearly wasn't Lewis' intent for his argument to rest on the specifics of Jeffrey's theory—and more importantly, we have good reasons to think that it doesn't. Williams (2016) shows how to get the same *kind* of underdetermination result for Lewis' (1981) causal decision theory, while (Elliott 2017a) contains much the same style of argument applied to a wide class of 'Savagean' decision theories. In both cases, the upshot is the same: an agent's choices over options might *at most* determine her preferences up to a limited subset of the propositions regarding which she has beliefs and desires, but for the remaining propositions we have underdetermination. So Lewis' conclusion seems to be quite robust against variations in the decision theory being applied. The conclusion *is* perfectly consistent with Jeffrey's representation theorem, but *that's* not yet getting at the heart of the matter.

The second suggestion is trickier to deal with. In his (2014: 21–2), Schwarz states that 'Lewis did not trust these [representation theorem] results'. Schwarz pins that distrust on the fact that amongst the axioms $a_1, \ldots, a_n$ that Savage, Ramsey, *et al.*, use to prove their uniqueness results there will usually be some implausibly strong requirements relating to the 'richness' of space of options over which the agent's choice ranking is defined. (See also Schwarz 2012; 2015: 513; Williams 2016: 430 suggests something similar.) If this is right, then you can see one easy way to reconcile Lewis' conclusion with Savage's and Ramsey's results: if the axioms $a_1, \ldots, a_n$ needed to ensure a unique representation are never jointly satisfied, then Representation Theorem doesn't provide us with reasons to believe that choices can uniquely determine beliefs and desires.

But I don't think this gets us to the heart of the matter either. Lewis certainly *might* have rejected these domain-richness axioms—a great many others have! But 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. This is

[9] Of course, the $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ in the example of §2 aren't related by a fractional linear transformation either. But it's easy to see that this, too, is entirely consistent with Jeffrey's representation result: since $\mathcal{A}(\mathbf{G})$ is a proper sub-algebra of $\wp(\Omega)$, if the evidence-and-choice facts can be taken to determine no more than $\succsim$ over $\mathcal{A}(\mathbf{G})$, then Jeffrey's theorem at most establishes that Typikarl's beliefs and desires over $\mathcal{A}(\mathbf{G})$ can be determined up to a fractional linear transformation, and might be much more radically underdetermined for the remainder.

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.

We'll focus on Savage. Savage's formalisation of expected utility theory starts off with two partitions: a set of *states*, and a set of *consequences* that specify ways the world might be in as much detail as makes a difference to what we care about. Think of the consequences as sets of equally-desirable worlds. From there, Savage defines his preference relation over total functions from states to consequences, each intended to represent a distinct *act*. If an act-function takes us from state s_1 to consequence c_1, s_2 to c_2, and so on, then that function represents the act such that if it's chosen and s_1, then c_1 results; if it's chosen and s_2, c_2 results, and so on. So far so good. But one of the axioms that Savage uses to prove his uniqueness result requires that *every* act-function represents a distinct act. This implies, amongst other things, the existence of so-called *constant acts*—acts with the same consequence regardless of what state happens to be true—which has long been a focal point of criticism for Savage's theorem. (See, for example, Fishburn 1981: 162; Maher 1993: 182–5; Joyce 1999: 107–8.) Would Lewis have rejected Savage's assumption too?

Before we answer that, we should get clear on just what the constant acts problem *is*. First, if Savage was assuming that every act-function represents an *available* act—something the decision-maker could choose to perform in their current circumstances—then it's highly unlikely that there's an act available to me that will bring about any level of value I so care to choose. But it's also unlikely that this is what Savage had in mind. If every constant act were an option then any minimally rational agent would just choose the constant act that guarantees the best consequence come what may (Joyce 1999: 67). Savage's theorem would be relevant only to omnipotent gods can directly choose any consequence at will, and that's clearly not how Savage understood the import of his work. More importantly, the Revealed Preference methodology doesn't *require* us to consider only the available acts. After all, if we're going to be talking about what Typikarl would choose given counterfactual option sets, then why would we restrict ourselves to his available acts? Typikarl still has preferences over possible-but-not-available acts, and dispositions to choose between them if they were to become available.

Ok—but isn't the problem that some constant acts won't even be *possible*? Well, that depends on what the space of consequences looks like. On the one hand, imagine that Typikarl has extremely opinionated 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 consequences, corresponding to *some* ways an agent's intrinsic values might be, Savage's act-richness axiom does entail the existence of impossible acts. On the other hand, suppose that Typikarl doesn't care about very much at all—say, there are only two or three relevantly distinct consequences 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 consequences there's at least one *possible* act that guarantees that consequence regardless of which state happens to be true.

As a general rule of thumb, the more fine-grained the consequence-partition, the less likely it will be that there exists a set of states such that every function from states to consequences corresponds to a possible act. Because most of us usually do care about quite a lot of things, Savage's assumption is problematic. *That's* the problem of constant acts. But the flip-side of this is that the more coarse-grained the consequences are, the

more plausible Savage's act-richness assumption becomes. And, crucially, the axioms $a_1, \ldots, a_n$ of Savage's theorem do not require the consequence-partition to be extremely fine-grained—and neither does the underdetermination argument require any special assumptions about how opinionated Typikarl's intrinsic values can be.

The upshot here is that there are cases where Lewis' argument seems to apply but for which we've no special reason to reject Savage's act-richness axiom. So whatever it is that's going on between Lewis' argument and Savage's theorem, it's not—or is not *only*—a disagreement over the richness of the act-space. We're going to need something more general than this if we're going to reconcile the two.[10]

3.3 The source of underdetermination

To apply a result like Representation Theorem to Typikarl's case we need an appropriate theoretical link between his behaviour and his preferences. If Typikarl's preferences can be simply read off of his counterfactual choice ranking, then the application is straightforward. But at no point is this how Lewis describes things.

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

> 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's 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 [the physical facts] 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: 337, emphasis added; see also 1983a: 374; 1986: 36ff)

That is, the physical facts supply us not a *ranking* over a space of propositions, but a single proposition about his behaviour that entails (one of) the option(s) that Karl considers best given his beliefs and desires. In a later work, Lewis adds more detail:

> … 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*… (1980b: 287–8, emphasis added)

So in Typikarl's case, for each time τ we know:

i) which evidence-specifying e characterises Typikarl's sensory evidence up to τ,
ii) which behaviour-specifying b Typikarl makes true at τ, and (more generally),
iii) which b' he *would* have made true if he *were* to have had some other history of sensory evidence e' instead.

[10] 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 weakened. More recent theorems which closely follow the structure of Savage's own manage to establish conditions sufficient for unique expected utility representations without assuming anything quite as strong (e.g., Abdellaoui and Wakker 2005; Gaifman and Liu 2018).

Note that I've been saying 'behaviours' for a reason—not 'options'. Whatever behaviour he performs at a time must entail an option o that is amongst those options he desires most at that time. But since what counts as Typikarl's options depends in part on his beliefs, we cannot in general presume to know which options Typikarl is choosing between given just the physical facts.

So let's make a distinction. On the one hand there's the matter of what *options* an agent would choose if her *options* were thus-and-so. We'll call these *o-counterfactual choices*. These are the kinds of facts about choices that get 'encoded' in a counterfactual choice ranking if Revealed Preference is true. On the other hand, there's how an agent would *behave* if her *evidence* were thus-and-so. We'll call these *e-counterfactual choices*. The facts about e-counterfactual choices are the kinds of facts that Lewis thought would be available for the purposes of radical interpretation. So the question for us now is: if we knew enough about Typikarl's e-counterfactual choices, as well as his actual evidence and behaviour, then would this be enough to determine his preferences? (Spoiler alert: they won't be.)

Lewis argued that any given pattern of e-counterfactual choices can be consistent with multiple hypotheses about initial beliefs and desires. That argument made use of some idiosyncratic aspects of Jeffrey's decision theory, as well as some assumptions about the 'grain' of the sets $\bigcup(\mathbf{B}^\tau)_{\tau\in\mathfrak{T}}$ and $\mathbf{E}$, but there are some more general reasons to think that e-counterfactual choices will underdetermine preferences. Let's get three of the obvious ones out of the way first:

1. What behaviours Typikarl performs will depend on what his options he is choosing between, which in turn depends at least in part on his beliefs. Consequently, it might not be possible to know what Typikarl's options are until we first know his beliefs, and if we cannot know his options then we cannot determine his preferences over those options merely from the facts about his behaviour.
2. Even if we knew what Typikarl's options are, his performing a behaviour b that entails the option o might indicate that he uniquely prefers o above the alternatives, or it might only indicate that he considers o one of the best.
3. Even if we assume there are never any ties for equal best, observing Typikarl's choices at a time will let us determine which options he considers best at that time, but won't help us draw comparisons between options across different times.

The first of these relates to a long-standing problem for using behavioural information to determine preferences in the absence of assumptions about belief, and is usually discussed in connection with revealed preference theory (e.g., Hausman 2000; 2012: 27ff; Thoma forthcoming). The second concerns another long-standing problem with behaviourally distinguishing preference from indifference (cf. Savage 1954: 17; Maher 1993: 12–4). The third is a unique problem that arises from when we appeal to e-counterfactual choices, as opposed to o-counterfactual choices, to gather information about preferences.

All three are important potential sources of underdetermination that Lewis would likely have been aware of, but they are *not* the ones I want to focus on. Consequently, let's assume for the sake of argument (*per impossibile*) that

1. we can know what Typikarl's options are despite not knowing his beliefs,
2. Typikarl is never indifferent between any options, and
3. any option is available for Typikarl to choose at any given time.

Even given these recklessly implausible assumptions, *still* Typikarl's e-counterfactual choices won't determine his preferences.

Suppose that Typikarl's options are $\mathbf{O} = \{o_1, o_2, o_3, o_4\}$, and that he ranks these $o_1 \succ o_2 \succ o_3 \succ o_4$. In order to 'extract' these preferences from Typikarl's e-counterfactual choices, we will need the appropriate propositions about Typikarl's evidence to exist in $\mathbf{E}$—specifically, we need e that line up nicely with the appropriate restrictions on his options $\mathbf{O}$. Here's what I mean by that. Typikarl will choose o_1 from $\mathbf{O}$, but we still need to know how he ranks the sub-maximal options, so we need to give him evidence that o_1 is unavailable. It would suffice to consider either:

i) Typikarl's choices after conditionalising on $o_2 \cup o_3 \cup o_4$ (to determine $o_2 \succ o_3$ and $o_2 \succ o_4$), and after conditionalising on $o_3 \cup o_4$ (to determine $o_3 \succ o_4$); or

ii) Typikarl's choices after conditionalising on $o_2 \cup o_3$ (to determine $o_2 \succ o_3$), and after conditionalising on $o_3 \cup o_4$ (to determine $o_3 \succ o_4$).

If we have Typikarl conditionalise on anything else, then we run the risk of distorting the results. (For example, if e entails the negation of o_1 but cross-cuts o_2, o_3, and o_4, then Typikarl's choices after updating on e won't necessarily tell us what he prefers between o_2, o_3, and o_4, only what he prefers between $e \cap o_2$, $e \cap o_3$, and $e \cap o_4$.) And that's a problem, because we're unlikely to find *any* of the propositions $o_2 \cup o_3 \cup o_4$, or $o_2 \cup o_3$, or $o_3 \cup o_4$, in $\mathbf{E}$. After all, the evidence-specifying propositions are supposed to characterise *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 tastes and so on Typikarl 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. No proposition with that content is going to be equivalent to a disjunction of propositions about how Typikarl behaves at some time.

Note that this problem arises even if we assume that the space of options is *arbitrarily* rich—indeed, the richer the space of options available, the more evidence-specifying propositions we'll need in $\mathbf{E}$ if we're to use Typikarl's e-counterfactual choices to determine his preferences over those options. And there's a further problem still if we assume that *every* proposition can count as an option or as evidence. For if Typikarl can choose to make any proposition true, then he will always just choose to make true whichever world ω has maximal value. After conditionalising on e, then if e contains ω he'll choose ω; otherwise he'll choose whichever world ω' is most valuable in e. In this case, with a rich enough $\mathbf{E}$ we could use Typikarl's e-counterfactual choices to determine his preferences over (singleton sets of) *worlds*, but this will radically underdetermine his preferences over the full suite of propositions regarding which he has beliefs and desires.

Contrast all this with the Revealed Preference method for extracting preferences from Typikarl's o-counterfactual choices. If we allow ourselves to stipulate arbitrary hypotheses about Typikarl's options, then we can say, for example, that his options are exactly $\{p, \neg p\}$, and thus determine Typikarl's preferences between p and $\neg p$. By contrast, if we start off saying that Typikarl's possible options can include any propositions whatsoever, then for any e the options that Typikarl would take to be available after conditionalising on e will include all subsets of e. This means that there *is* no e we can have him update on such that he's left believing his options are exactly $\{p, \neg p\}$—for if $e = p \cup \neg p$ then his beliefs won't change at all; whereas if $e \subset p \cup \neg p$ then his options won't be p and $\neg p$ but rather $p \cap e$ and $\neg p \cap e$ plus every other subset of e. So making the space of options richer *won't* solve the problem of underdetermination.

The above are all different ways of getting at a very general reason why Typikarl's e-counterfactual choices underdetermine his preferences. If you want to extract preferences over some space of options $\mathbf{O}$ from the facts about choices, then you need to consider choices between specific subsets of those options. That's key to how the Revealed

Preference methodology is supposed to work. But we cannot apply the same method with e-counterfactual choices, because we won't find the right evidence-specifying propositions, corresponding to appropriate restrictions in the option set, for Typikarl to conditionalise on. The problem isn't—or isn't *only*—the 'richness' of the space of options, nor is it just about what kind of propositions can serve as evidence. The problem is that the facts about e-counterfactual choices just don't contain the right kind of information from which sufficient information about preferences can be extracted.

So here, I think, is the deeper reason why there's no conflict between Lewisean underdetermination and a result like Representation Theorem. The latter entails that under the right conditions $a_1, ..., a_n$, Typikarl's preferences $\succsim$ will be uniquely determined relative to some expected utility theory $\mathcal{T}_{\text{EU}}$ by some system of beliefs and desires $(\mathcal{B}, \mathcal{V})$. If we combine this with Revealed Preference, then there will exist some $\succsim^c$ such that

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

But the content of folk psychology *isn't* given by $\mathcal{T}_{\text{EU}}$, and it *doesn't* include Revealed Preference (as I'll argue soon). Instead, the folk psychological theory tells us that if a typical agent were to start off with beliefs and desires $(\mathcal{B}, \mathcal{V})$, then if she were to receive evidence e she would be disposed to behave in some way b that entails some option o that has maximal expected value relative to $(\mathcal{B}^e, \mathcal{V})$. The content of the theory is not given by $\mathcal{T}_{\text{EU}}$, but better represented by

$$\mathcal{T}_{\text{FP}} : (\boldsymbol{\beta} \times \boldsymbol{\delta} \times \mathfrak{T}) \mapsto \Big\{ \mathcal{F} : \mathbf{E} \twoheadrightarrow \bigcup_{\tau \in \mathfrak{T}} \big(\wp(\mathbf{B}^\tau)\big) \Big\}.$$

Specifically, $\mathcal{T}_{\text{FP}}(\mathcal{B}, \mathcal{V}, \tau)$ picks out a function $\mathcal{F}$ that represents a specific pattern of actual and e-counterfactual choices, taking us from the evidence e compatible with $\mathcal{B}$ to the set of behaviours available at the later time $\tau + e$ that entail some option in $\mathbf{O}^{\tau+e}_{\mathcal{B}^e}$ that maximises expected value relative to $(\mathcal{B}^e, \mathcal{V})$, where '$\tau+e$' is τ plus the whatever the duration of e happens to be. And for all that a result like Representation Theorem might tell us, there need not be any $\mathcal{F}$ such that

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

Moreover, we've just seen several reasons to think that such an $\mathcal{F}$ *won't* exist. There is no conflict with Representation Theorem. There *is* underdetermination.

Lewis' conclusion was that the evidence-and-choice facts won't determine Typikarl's beliefs and desires for any propositions outside of $\mathcal{A}(\mathbf{G})$. What I've been arguing now is that the evidence-and-choice facts won't determine Typikarl's beliefs and desires over the propositions *within* $\mathcal{A}(\mathbf{G})$ either. The underdetermination is worse than we thought. Even if we make the partition $\mathbf{G}$ maximally fine-grained—whether by letting any proposition count as evidence, or by enriching Typikarl's space of options beyond all sense of plausibility, or both—still we can expect plenty of underdetermination of the belief-and-desire facts by the evidence-and-choice facts on the Lewisean functionalist theory. At best, Typikarl's e-counterfactual choices will help us to determine bits and pieces of his preferences over $\mathcal{A}(\mathbf{G})$, but the majority of his preference ranking will not be 'revealed' no matter how much we know about his evidence and his e-counterfactual choices.

4. Representation Theorems and Radical Interpretation

Finally, we should talk about the elephant in the room: even if the *Lewisean* theory of radical interpretation makes no reference to o-counterfactual choices, an alternative theory might do so. Nothing I've said so far entails that we *can't* or *shouldn't* use the facts about an agent's o-counterfactual choices to pin down the facts about her preferences and, on that basis and given the appropriate representation theorem, determine 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. Extracting information about preferences from information about o-counterfactual choices requires something like Revealed Preference to mediate the inference, so first we need to consider whether that thesis is something we ought to accept. For if not, then representation theorems don't tell us how to determine beliefs and desires given choices on *any* plausible theory of radical interpretation.

We should reject Revealed Preference. As in the previous section, in arguing for this I will set aside some common concerns that arise when we try to determine preferences between options on the basis of facts about behaviour. These are well-known epistemic problems with the standard methodology of revealed preference theory, but they aren't reasons to reject the Revealed Preference thesis itself. Consequently, we'll assume again that we can know what Typikarl's options are even if we don't know his beliefs and how he conceives of his options, and we'll assume that he's never indifferent between any options. The *real* problem with Revealed Preference isn't methodological, it's that the thesis is false: the idea that o-counterfactual choices 'reveal' actual preferences just doesn't fit well with any plausible theory of belief update.[11]

In any realistic decision situation, there's going to be a *very* wide range of available options an agent might choose between. We ignore most of these when drawing up a decision table, but instead of just *going out for Thai food* versus *going out for Italian*, 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 between are given by $\mathbf{O} = \{o_\varnothing, o_1, ..., o_{100}\}$, with

$$o_1 \succ o_2 \succ \ldots \succ o_{99} \succ o_{100}$$

$\succsim$-Maximisation and Revealed Preference each predict that o_1 will be chosen.

But now let's go to the counterfactual scenario where his options are given by, let's say, $\mathbf{O}^\star = \{o_{83}, o_{84}\}$. $\succsim$-Maximisation makes no predictions about what Typikarl would do in this scenario—that principle tells us how Typikarl will choose amongst his options given his preferences; it does not tell us how Typikarl would choose if his options were thus and so given that his preferences actually thus and so. After all, it's entirely consistent with $\succsim$-Maximisation that Typikarl's preferences in the counterfactual scenario are vastly different than his actual preferences. Revealed Preference, on the other hand, predicts that Typikarl will choose o_{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 o_1 on the basis of his beliefs and

[11] An argument similar to the one I'm about to make here is also made by Hausman (2012: 31–3). For some reason, it seems to have been neglected in responses to that work. To my mind it points to a much deeper problem with revealed preference theory than the observation (also in Hausman 2012: 27ff) that we cannot determine preferences over options from observations of behaviour in the absence of hypotheses about belief—which is what defences of revealed preference theory have focused on so far.

desires; in the counterfactual he purportedly chooses o_{83}, but somehow his beliefs don't change? But what else could plausibly explain the difference in what's chosen aside from a change in beliefs about what choices are available? Surely we don't want to posit that Typikarl has some magical means of direct access to what options are available to choose between, which informs his behaviour *without* in any way affecting his beliefs.

No: if there's a difference in Typikarl's behaviour across the different scenarios, then the most plausible explanation is that his beliefs have changed—presumably, by learning about the restrictions to his option set. If that's the case, though, then it matters a great deal exactly *how* Typikarl's beliefs change in the new 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 a significant change in your beliefs, since the nearest possible world where anything like that could be true would be quite far off indeed. So Typikarl's beliefs in such strange circumstances are not likely to be much like his beliefs in the actual world. And if they're not, then why think his o-counterfactual choices have any strong connection to what's going on inside his head in the actual world?

Here's a more precise way to put that point. Assume that Typikarl's intrinsic values never change. We don't know what his beliefs and desires are, except that they're given by some $(\mathcal{B}, \mathcal{V})$ and that $o_{83} \succ o_{84}$. In this case, we can be certain that Typikarl will choose o_{83} in the scenario where his options are restricted to $\{o_{83}, o_{84}\}$ only if we can be certain that his beliefs in that scenario will be given by $\mathcal{B}$ conditionalised on $o_{83} \cup o_{84}$. But what Typikarl knows in the counterfactual is not *I will choose* o_{83} *or* o_{84}, but the much stronger proposition o_{83} *and* o_{84} *are the only options available*. And there's a very big difference between coming to believe $o_{83} \cup o_{84}$ under the supposition that he could have chosen any of o_1 through to o_{100}, versus coming to believe that o_{83} and o_{84} exhaust the extent of his choices. The latter requires a major rethink of the causal structure of the world, so his counterfactual beliefs aren't likely to be very similar to $\mathcal{B}$ conditionalised on $o_{83} \cup o_{84}$.

So the truth of Revealed Preference seems to require either that decision-makers' choices to magically reflect changes in the option set despite no changes in their beliefs, or implausible assumptions about how we would update our beliefs upon learning that our options have been severely restricted. The thesis is plausible enough if we consider only 'minor' restrictions to the option set, and cases where we can reasonably assume that the decision-maker's beliefs won't change *too* much if we take away some of their options. But that's not going to be enough for the purposes of radical interpretation, where we will want to know the decision-maker's preferences over a rich space of options. For the general case, then, Revealed Preference isn't remotely plausible, and it seems that o-counterfactual choices don't determine preferences either.

5. Conclusion

I promised to answer two questions:

(i) How do representation theorems like Savage's and Ramsey's relate to Lewis' underdetermination argument?
(ii) Are there *any* plausible theories of radical interpretation under which these theorems show that beliefs and desires are (sometimes) determined by choice dispositions?

For the first, the answer is 'they don't really'. The representation theorems tell us that it might be possible to determine an agent's beliefs and desires given enough knowledge

about her preferences—but we're never going to *have* enough information about the agent's preferences given just the facts about choices, since preferences are underdetermined by *e*-counterfactual choices. For the second question, the answer is 'probably not'. An agent's *e*-counterfactual choices underdetermine her preferences, and her *o*-counterfactual choices determine her preferences only under 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 what representation theorems tell us about the relationship between our attitudes and our choices, and what we can learn about the former given only information about the latter. 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 theory.

Appendix

I've said that I think Lewis was right about underdetermination, but there were several important simplifications that went into the underdetermination argument presented in §2. In this short appendix, I'll discuss several of these simplifications and why I don't think they matter too much to truth of the conclusion.

First: obviously, $\mathcal{B}$-Eligibility, $\mathcal{V}$-Eligibility, $\succsim$-Coherence, Conditionalisation, and $\succsim$-Maximisation are all highly idealised, even in relation to typical agents. But we can reasonably expect that the underdetermination argument will be robust against various ways of 'de-idealising' Bayesianism. For example, the argument doesn't hinge on $\mathcal{B}$ and $\mathcal{V}$ assigning *precise* numerical values, 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 that can considered possible—e.g., by allowing for the possibility of imprecise and probabilistically incoherent systems of belief—will only make the degree of underdetermination worse. Likewise, it's a consequence of $\succsim$-Coherence, $\succsim$-Maximisation, Conditionalisation, and Static Values that Typikarl's choices over time are determined by his initial beliefs, intrinsic values, and evidence. That's implausible. But weakening these constraints isn't likely to help—the fewer constraints there are on transitions between states, the worse we can expect the underdetermination to be.

Second: Static Values isn't very plausible, but it does play an important role in the underdetermination argument. Interestingly, *how* we go about denying Static Values can make a difference to the conclusion. 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 we'd expect to see a greater degree of underdetermination. On the other hand, suppose that 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 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 of what I mean here, 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; and 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 perfectly 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 confidence he has that p is true. The 'reasonable' and 'deviant' interpretations will then diverge in the choices they predict at later times due to the systematic change in values.)

But I doubt this will significantly affect the conclusion. 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. So I suspect that even if Static Values were replaced with something more plausible, we could still expect to see significant underdetermination.

The third and final limitation to Lewis' argument that I'll discuss concerns how *schemes of interpretation* are understood. As Lewis (1983b: 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 total systems of belief-desire $(\mathcal{B}, \mathcal{V})$ directly to total physical states depending on how well those states fit the functional role of $(\mathcal{B}, \mathcal{V})$ considered as a whole. This is the kind of scheme that Lewis makes use of for his argument in 'New Work'. 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 $\mathcal{V}$.

This distinction matters because coarse-grained schemes of interpretation ignore causal and counterfactual relationships between sub-total mental states. Here's an example. Let $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ be as they were described in §2. The functional role of the *total* belief-desire state $(\mathcal{B}_1, \mathcal{V}_1)$ is isomorphic to the functional role of $(\mathcal{B}_2, \mathcal{V}_2)$. That's what Lewis' underdetermination 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 $(\mathcal{B}_2, \mathcal{V}_2)$. *However*, for all we've said so far, the functional role of $\mathcal{B}_1$ need *not* be isomorphic to the functional role of $\mathcal{B}_2$; nor need it be the case that the functional role of $\mathcal{V}_1$ is isomorphic to the functional role of $\mathcal{V}_2$. For example, we know that

$$\sum_{\omega \in g} \mathcal{B}_1^g(\{\omega\})\mathcal{V}_1(\omega) \;=\; \sum_{\omega \in g} \mathcal{B}_2^g(\{\omega\})\mathcal{V}_2(\omega) \;=\; 7,$$

and thus $(\mathcal{B}_1, \mathcal{V}_1)$ generates the same expected values over options as $(\mathcal{B}_3, \mathcal{V}_3)$. But these generate different expected values than $(\mathcal{B}_1, \mathcal{V}_2)$ and $(\mathcal{B}_2, \mathcal{V}_1)$:

$$\sum_{\omega \in g} \mathcal{B}_1^g(\{\omega\})\mathcal{V}_2(\omega) \;=\; 6, \qquad \sum_{\omega \in g} \mathcal{B}_2^g(\{\omega\})\mathcal{V}_1(\omega) \;=\; 4.5.$$

So under recombinations with the values $\mathcal{V}_1$ and $\mathcal{V}_2$, $\mathcal{B}_1$ and $\mathcal{B}_2$ generate distinct expected values for at least some propositions in $\mathcal{A}(\mathbf{G})$, 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 (singleton sets of) worlds in the $g \in \mathbf{G}$. Define permutations for value functions similarly. Now let the *recombination function* $\mathcal{R}^{\mathcal{B}}_{\mathcal{V}}$ select for each p in $\mathcal{A}(\mathbf{G})$ 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 space of worlds it turns out that if $\mathcal{B}_i$ is not a permutation of $\mathcal{B}_j$ *or* $\mathcal{V}_i$ is not a permutation of $\mathcal{V}_j$, then

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

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 some p in $\mathcal{A}(\mathbf{G})$ when combined with different systems of value (beliefs). And that's interesting, because—as Williams (2016) points out—*most* of the underdetermination that's established by the argument in §2 *isn't* between permutations.[12]

But I don't think these facts will do much to allay the worries about underdetermination. First, while $\mathcal{B}_1$ and $\mathcal{B}_2$ generate different patterns of *expected values* for some of the propositions in $\mathcal{A}(\mathbf{G})$ under recombination, this doesn't yet entail they will generate different *preferences over options*; still less does it entail that they will generate different (*e*-counterfactual) *choices*. More importantly, we cannot use recombinations to distinguish between permutations. For example, even if we use fine-grained schemes of interpretation we still won't be able to distinguish between $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_3, \mathcal{V}_3)$:

$$\mathcal{B}_3(\{\omega_i\}) = \begin{cases} 0.3, & \text{if } i = 1 \\ 0.2, & \text{if } i = 2 \\ 0.1, & \text{if } i = 3 \\ \mathcal{B}_1(\{\omega_i\}) & \text{otherwise} \end{cases} \qquad \mathcal{V}_3(\omega_i) = \begin{cases} 9, & \text{if } i = 1 \\ 6, & \text{if } i = 2 \\ 3, & \text{if } i = 3 \\ \mathcal{V}_1(\omega_i) & \text{otherwise} \end{cases}$$

The functional role of $\mathcal{B}_1$ is isomorphic to the functional role of $\mathcal{B}_3$, and likewise for $\mathcal{V}_1$ and $\mathcal{V}_3$. So while we *might* be able to use recombinations to tell $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_2, \mathcal{V}_2)$ apart, this won't help us to distinguish between $(\mathcal{B}_1, \mathcal{V}_1)$ and $(\mathcal{B}_3, \mathcal{V}_3)$. The underdetermination result holds, though it might not be *quite* as radical as the argument of §2 makes it out to be.

References

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

Bermúdez, J. (2009). *Decision Theory and Rationality*. Oxford University Press.

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

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

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

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

[12] Lewis briefly notes a preference for fine-grained schemes in (1983b: 119). Given this, and since fine-grained schemes of interpretation cannot help us to distinguish between permutations, it's entirely possible that coarse-grained schemes were used in 'New Work' merely to simplify the discussion.

Elliott, E. (2017b). A Representation Theorem for Frequently Irrational Agents. *Journal of Philosophical Logic 46*(5), 467–506.

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.

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

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

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. (1965). *The Logic of Decision*. University of Chicago Press.

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

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

Joyce, J. (1999). *The foundations of causal decision theory*. 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. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, pp. 263–293. 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.

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

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

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

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

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

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. 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 University Press.

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

Williams, J. R. G. (2019). *The Metaphysics of Representation*. New York: Oxford University Press.