RADICAL MISINTERPRETATION

BY

EDWARD ELLIOTT

This paper provides an exposition and defence of Lewis' theory of radical interpretation. The first part of the paper explains what Lewis' theory was; the second part explains what it wasn't, and in so doing addresses a number of common objections that arise as a result of widespread myths and misunderstandings about how Lewis' theory is supposed to work.

1. *Introduction*

Karl is an ordinary human being, with ordinary beliefs and desires, and our task is to determine what those beliefs and desires are. The catch is that we're not allowed direct access to the facts about Karl's mental life or the meanings of his expressions, or indeed about any intentional properties whatsoever. What we are allowed to know are any and all facts about his physical constitution and environment, evolutionary ancestry, potential futures, counterfactual histories, and so on – although only inasmuch as these are expressed without invoking any 'spooky' mental or semantic properties of the sort that are often thought to raise question marks for physicalism. In short: given just the physical facts, we're to derive the facts about Karl's beliefs and desires. For the present discussion I'll be referring to this as the *problem of radical interpretation*.

In this paper, I explain and defend David Lewis' solution to this problem. I'll go into some detail, as there's significant variation in the literature on how Lewis' writings are to be interpreted and not a small amount of confusion on the matter. Some of the responsibility for this state of affairs lies with Lewis himself: the many details of his view are distributed over a dozen or so articles written over two decades, and he would frequently gloss over those

Pacific Philosophical Quarterly 104 (2023) 646–684 DOI: 10.1111/papq.12436

Pacific Philosophical Quarterly published by University of Southern California and John Wiley & Sons Ltd.

details in favour of more simplified expressions. But whatever the reason, Lewis' ideas are frequently misunderstood. So I aim to explain clearly what Lewis' theory was, and – just as important – what it *wasn't*. In so doing, I take myself to also be supplying a defence of that theory. As I see it, Lewis' ideas are far too often subject to critiques that rest heavy on mischaracterisation. The hope against hope here is that we might collectively move past the myths and the caricatures and evaluate Lewis' position on its actual merits.

There are two main parts to the paper. The first part is exposition. For this I've avoided saying very much about Lewis' 'Radical Interpretation' (1974), for that paper is the source of many a confusion. (Or so I'll argue.) The second part is then devoted to highlighting and debunking a number of common myths – including, *inter alia*, myths relating to the maximisation of rationality, source intentionality, interpretivism, and decision-theoretic representation theorems.

2. *What Lewis' theory was*

Section 2.1 provides an overview. Following that, Lewis' theory can be broken down into its reductive and non-reductive components. The non-reductive component is the *theory of content*, which is characterised by the *principles of fit* and the *principles of humanity*. Section 2.2 explains what a theory of content is, while Sections 2.3 and 2.4 say more about the principles of fit and humanity, respectively. Finally, Section 2.5 discusses the reductive component.

2.1. CONSTITUTIVE RATIONALITY

At the heart of Lewis' theory are two ideas. The first is analytic functionalism – that our concepts of belief and desire are implicitly defined by the roles they play within folk psychology (Lewis, 1974, pp. 334–335; 1979, p. 533; 1983a, p. 373; 1986, p. 39; 1994, pp. 428–430; henceforth, all otherwise incomplete citations will be to Lewis). If they are to exist at all, then our systems of beliefs and desires must be whatever they have to be in order to render folk psychology true, or near enough to true. The second is that the roles our systems of belief and desire play within folk psychology are, first and foremost, *rational* roles. Folk psychology is built around a defeasible presumption of rationality; the paradigm folk-psychological agent is one who acts in a more or less rational way given her attitudes, and who has more or less rational attitudes given the evidence of her senses. That's not the whole of what folk psychology says, but it's a big part.

Put those together and out pops a weak kind of *constitutive rationality*: folk psychology will count as near enough true only if our systems of belief and desire normally do a good job of conforming to whatever standards of rationality we find therein; hence, those standards will constitute a crucial part of what implicitly defines the attitudes. (Lewis makes this inference explicit in 1974, pp. 334–335; 1986, pp. 36–40; 1994, pp. 428–430; 2020, Letter 466 [1980].) These will include standards relating to what our systems of belief and desire should be like at any given time, how they should relate to our intentional behaviour, and how they should change in response to the evidence of our senses. For most agents most of the time, then, their systems of belief and desire must as a matter of a priori necessity be such as to mostly satisfy the folk-psychological standards of rationality in light of the facts about their experiences and behaviour.

A few notes are in order. First: despite Lewis initially characterising folk psychology as a collection of 'platitudes' (in 1972, p. 256), it is best not to think of it as a conjunction of sentences the folk themselves would be inclined to spontaneously assert or unreflectively accept (see 1994, p. 416). Our tacit understanding of grammar provides a better model. Folk psychology is a representation of the implicitly understood posits and principles that guide typical human interactions with and interpretations of one another. Not only might it involve complicated or subtle rules that are difficult to express accurately, it might also be expressed using technical machinery that the folk themselves need not easily comprehend. It would be enough, Lewis said, that a member of the folk '*could* recognise those principles as being something he believed all along, when and if someone else formulated them and explained to him what they meant' (2020, Letter 435 [1973], my emphasis; see also 1974, pp. 337–338).

Note two: folk psychology is not a precise theory. In many circumstances – especially the rare or the unusual – it will have not much determinate to say. Folk psychology is an instrument of explanation and prediction for everyday use, and there's no strong impetus for it to have something very settled to say about the many hypothetical puzzle cases that philosophers like to dream up: '[t]he advantage of being prepared is not worth the bother of solving countless problems in advance when most of them will never arise' (1997, p. 358). Lewis frequently emphasised the indeterminacy of folk psychology in edge cases, and took semantic indecision to be a natural and acceptable consequence of this. (See 1980a, p. 220; 1983b, p. 120; 1994, p. 417; 2020, Letters 436 [1973], 450 [1979], 466 [1980], 486 [1988], 506 [1994]; 629 [1992].)

Note three: I've been saying 'systems of belief and desire' for a reason. While Lewis often spoke of beliefs and desires as if these were clearly separable entities (e.g., 1983a, p. 373; 1983b, p. 119; 1988, pp. 323–325), he also often preferred to treat entire systems of belief and desire as the basic 'units' of the attitudes:

> The contentful unit is the entire system of beliefs and desires. (Maybe it divides into contentful snippets, maybe not.) (1994, p. 430)
>
> We have a disagreement about total state functionalism … I think the causal relations of the total states give us grip enough to raise the questions whether the Martian works the same way we do … I think an *a priori* functionalist had better use fairly total states, at least in the department of belief; because folk psychology hasn't anything very decisive and plausible to say about how the belief system splits into many smaller states. It splits somehow, no doubt; but folk psychology gives no good guidance on how, so we'd best look for realizations of the folk psych of total belief states. (2020, Letter 485 [1986].)

I take the implication to be that, inasmuch as folk psychology has anything reasonably determinate to say about our beliefs and desires, it will be in connection to the total systems thereof. Regardless of how a total system divides into 'snippets' – as it no doubt does – there must be a total system and it must play a certain kind of causal role. That provides a good starting point for our analysis, and one that gives us 'grip enough' for many purposes. In what follows, therefore, I'll be treating each system of beliefs and desires as though a single intentional state, with its own causal role and a single (albeit complex) content. I won't assume that these holistic states divide into 'snippets' with separable causal roles, although I won't assume they don't either.

Note four: I said 'the folk-psychological standards of rationality' for a reason. Lewis famously believed that a systematic formulation of folk psychology ought to look a lot like what we find in Bayesian models of learning and decision-making (1974, pp. 337–338; 1979, pp. 533–534; 1980b, pp. 287–288; 1994, p. 428). But do not be misled! Those models he took to be 'approximately descriptive' in many respects (2020, Letter 673 [1980]), but over-idealising in others.[1] Lewis did not believe that ordinary agents live up to the idealised standards of rationality encoded in Bayesian models, nor that folk psychology takes it to be so either:

> I think [Bayesianism] a good reconstruction of standards of rationality we would *like* to live up to … Then we can get a more realistic account by noticing that Bayesian rationality requires enormous amounts of memory and calculation, so we cut corners in a way we hope won't take us too far away. (2020, Letter 429 [1968])
>
> [It] seems unlikely that any real person could store and process anything so rich in information as the [probability and utility] functions envisaged … But it is plausible that someone who really did

[1] In *almost* every location where Lewis explicitly draws upon Bayesianism to aid in describing his theory of the attitudes, he also highlights that they over-idealise. See (1979, p. 515), (1983a, p. 375), (1986, p. 30), (1988, p. 325), (1994, p. 428), (1996, p. 303), (2020, Letters 524 [1998], 651 [1972], 673 [1980], 684 [1986], 695 [1989], 722 [1994], 737 [1999]). The sole exceptions to this trend (that I've been able to find) are (1974, pp. 337–338) and (1980b, pp. 263–288). Note that there are two types of idealisations: those imposing excessive standards of rationality, and those 'that make the topic tractable' (1996, p. 303). Lewis mentions both. For instance, he thought that 'thoroughly quantitative' (i.e., real-valued) representations of degrees of belief and desire are unrealistic (1988, p. 325 fn. 2; 1986, p. 30; 1994, p. 428).

> have these functions to guide him would not be so very different from us in his conduct, apart from his supernatural prowess at logic and mathematics and *a priori* knowledge generally. (1981, p. 7)
>
> [Decision theory] is psychologically unrealistic – sure it is … We're describing (one aspect of) what an ideally rational agent would do, and remarking that somehow we manage to approximate this, and perhaps – I'd play this down – advising people to approximate it a bit better if they can. (2020, Letter 674 [1981])

So let's distinguish between standards of *ideal rationality* – those standards we would like to satisfy but usually don't – from standards of *near-enough rationality* – those standards the satisfaction of which would make us overall more rational than irrational, while still leaving room for error. It's the latter that matter for Lewis' analytic functionalism:

> It wouldn't do to conclude that, as a matter of analytic necessity, anyone who can be said to have beliefs and desires at all must be an ideally rational *homo economicus*! Our rationality is very imperfect [and] the folk know it too … But there is no cause for alarm. Folk psychology can be taken as a theory of imperfect, near-enough rationality, yet such rationality as it does affirm can still be constitutive. (1994, p. 428)
>
> I think of folk psychology as predicting behaviour by a two-step procedure. First it introduces a (descriptive!) distinction between rational and irrational, or (better) between more and less rational; second, it predicts a certain modicum of rationality. (2020, Letter 505 [1993])
>
> In short, folk psychology says that we make sense. It credits us with a modicum of rationality in our acting, believing, and desiring. (1994, p. 428)

I'll have more to say about this below. For now, I just want to put the issue on your radar. Lewis' theory is one of constitutive rationality, but do not confuse it for something stronger than it is.

2.2. CONTENTS AND INDICES

On a common picture nowadays, our beliefs and desires are inner representations of a computational system, usually thought to be encoded in a sentence-like format, and what the belief or desire does it does partly in virtue of the content it represents. A theory of the attitudes is provided in two parts, which are usually treated independently. One is to explain how these representations come to have the contents they do – this might be via causal interactions with the environment, proper functions, inferential roles, or what have you. The other part is to explain when a representation should be categorised as one type of attitude or another – what makes it the case that this sentence belongs in the 'belief box' as opposed to the 'desire box'. The type and the content of the representation then fixes the role it plays within your cognitive economy.

Lewis was ambivalent towards this picture. He thought the hypothesis that our propositional attitudes correspond to inner representations a piece of ‘plausible, *not* unfounded, neurophysiological speculation’ (2020, Letter 441 [1977]), but also something that goes beyond folk psychology and therefore something that has no proper place in an *a priori* functional analysis (1994, pp. 421–423; 2020, Letters 441 [1977], 485 [1986]). He did not deny that the brain somehow manages to represent things, but he did deny that we *must* conceive of our attitudes as corresponding to any specific representations in a modern theory of cognition. More importantly, it’s no part of the Lewisean picture that our intentional states have their causal roles by virtue of any content they may or may not represent. That gets the order of explanation precisely backwards. Better to say instead that *we* represent our intentional states by associating them with contents in a manner that systematically corresponds to their distinctive folk-psychological roles. Contents, on Lewis’ view, are indices we assign to keep track of and reason about nodes in a network of causal relations; what, if anything, the nodes themselves might represent is entirely besides the point.

Compare the use of numbers to index some quantity – say, temperature. (Lewis first draws the analogy between indexing physical quantities with numbers and indexing mental states with contents in a letter to Stalnaker; see 2020, Letter 441 [1977].) Each specific temperature constitutes a node in a non-numerical relational network: some temperatures are *hotter* than others, *colder* than others, *between* this temperature and that, and so on. Such relations are analogous to relations between numbers, and thus, it’s possible to assign to each temperature a unique (up to choice of unit and zero point) numerical value such that the relations between them are mirrored in the relations between the numbers so-assigned. Indexing temperatures with numbers affords mathematical generalisations and reasoning about what is an essentially non-mathematical domain of inquiry. The temperatures *themselves* do not represent those numbers in any further interesting sense; instead it is *we* who represent temperatures using numbers.

As with temperatures and numbers, so too with intentional states and their contents. Our intentional states constitute nodes in a causal network posited by folk psychology, and we can index nodes in that network by assigning to each some abstract content that systematically corresponds to its location therein. Indeed, Lewis held that the folk employ multiple indexing schemes:

> I don’t like the question ‘whether belief is fine-grained or coarse-grained’. […] There are belief states; these are not, in their nature, fine- or coarse-grained. They do not consist of external relations to fine- or coarse-grained propositions in abstract heaven! Rather, they probably consist of patterns of synaptic interconnection. We characterise these states by indexing them with content, much as we characterise states of molecular motion by indexing them with numbers. In either case, the detour through the scheme of indexing facilitates generalisation about the causes and effects of the states. In either case, we have more than one fruitful way of doing the indexing. (2020, Letter 478 [1983])

The first and more fundamental scheme has us assign coarse-grained contents – the kinds of contents that can be captured within a standard possible worlds framework – to intentional states in a manner designed to encode those state's functional roles. I'll have more to say about how that all works in due course, but by way of initial example suppose that the content of a belief-desire system can be approximated by a distribution of probabilities and utilities over the space of possible worlds. (Not *a* space; *the* space.) Then, very roughly, if one system of beliefs and desires normally causes more coffee-directed behaviour than another system does, say, then all else equal we might suppose that the former should be indexed by some probability-utility pair that assigns greater utility to worlds where the subject possesses coffee than whatever probability-utility pair we use to index the latter system.[2]

The *theory of content* provides a more systematic account of how this first indexing scheme is supposed to work, of how contents are associated with intentional states so as to reflect their respective functional roles. That will be the focus of my discussion below, but it's worth saying a bit more here about the *other* indexing scheme and how it relates to the first. This secondary scheme (or perhaps jumble of schemes) is employed to help make sense of ordinary language belief- and desire-attribution sentences, which otherwise have very little relevance to understanding the theory of content (2020, Letter 478 [1983]). Whereas the first scheme will have us attribute coarse-grained contents, the second scheme makes use of a more fine-grained conception of content:

> One system of content-indexing assigns coarse-grained propositions (or better, properties); and does so in a narrowly psychological way, on the basis of the functional roles of the belief states being characterised. […] Another system (or several systems jumbled together) assigns fine-grained propositions, and does so in such a way that the fine-grainedness makes a difference; and does so in a broadly psychological way, so that narrow-psychological duplicates get assigned different fine-grained content. It is this second system (mixture of systems?) that is encoded in the belief sentences of ordinary language. (2020, Letter 478 [1983])

The second scheme has us assign contents by relating the contents as assigned by the first scheme to external properties or things in the actual world via relations of acquaintance, and it does so in a variety of ways. There is no single coherent principle to neatly characterise how this second scheme works – no 'unified formula to cover all cases' (1986, p. 33) – but one example might help. Suppose that Karl is watching someone sneak

[2] As Schwarz (2015) points out, where the content of a system of attitudes is given by a distribution of probabilities and utilities over the space of possible worlds, functional differences between systems of belief and desire will correspond not so much to differences in the *propositions* towards which the agent does or does not have attitudes, but rather in the varying *strengths* with which they have those attitudes.

through the shadows. The sneak just so happens to be Bernard, although Karl himself cannot make out who it is. Under the first scheme, we might assign to Karl a system of beliefs according to which the content *the person in front of me is sneaking through the shadows* is attached to a high probability, but not one where *Bernard is sneaking through the shadows* is likewise highly probable. That is the assignment of content that will best represent the causal role of Karl's beliefs. On the other hand, under the second indexing scheme, we might appeal to the perceptual acquaintance relation that just so happens to hold between Karl and Bernard to identify Bernard as 'the person in front of me' specified in the content assigned by the first scheme, and thus assert that Karl believes Bernard is sneaking through the shadows. (See 1986, pp. 32–34; 1979, 536ff, for more discussion and examples.)

The lesson going forward is that we shouldn't seek to understand the theory of content via an analysis of ordinary language, nor should we seek to show how the truth or assertability conditions of attribution sentences can be derived directly from the contents assigned by the first scheme. That way madness lies, for the contents assigned by the second scheme, or schemes, and embedded in propositional attitude attributions, 'are a far cry from the contents that best serve to index belief [and desire] states in a way that codifies their functional roles' (2020, Letter 478 [1983]). There is no straightforward connection between the two. For the same reason, it helps to keep in mind that our formulation of folk psychology – of the principles that tacitly guide our interpersonal interactions and interpretations of one another – need not be cashed out in the same language the folk themselves would use. When we are spelling out the theory of content, we are not *ipso facto* providing a theory of attitude attributions, and apparent conflicts between the contents assigned by the first indexing scheme with how we might pre-theoretically talk about our beliefs and desires need not spell doom for Lewis' theory. Such conflicts need to be resolved, but their resolution need not be a part of the theory of content itself (cf. 1986, pp. 34–36, on coarse-grained contents and hyperintensionality).

Do keep in mind also that the theory of content is *not* supposed to provide us with a reductive analysis of intentional states in terms the physicalists would find acceptable. That will come later. To provide a theory of content is to detail what *folk psychology* says about the typical causal roles of our intentional states, and thus comes prior to any functional analysis. (One cannot define beliefs and desires as *whatever comes closest to satisfying the belief-and-desire roles in folk psychology* until one has specified what folk psychology says about those roles.) I will therefore make free use of intentional notions – 'intentions', 'experiences', 'evidence' – when describing Lewis' theory of content; my doing so shouldn't be considered problematic.

2.3. THE PRINCIPLES OF FIT

A theory of content, Lewis says, can be characterised by its constraining principles. There are two types of principle: *fit* and *humanity*. The principles of fit are the more important:

> Given the functional roles of the states, the problem is to assign them content. States indexed by content can be identified as a belief that this, a desire for that, a perceptual experience of seeming to confront so-and-so, an intention to do such-and-such. […] The problem of assigning content to functionally-characterised states is to be solved by means of constraining principles. Foremost among these are principles of fit. (1983a, pp. 373–374)

The principles of fit tell us how contents relate to causal roles. There are principles of fit not only for systems of belief and desire, but also for intentions and sensory experiences. Such principles will include, for instance:

> If a state is to be interpreted as an intention to raise one's hand, it had better typically cause the hand to go up [1]. If a state (or complex of states) is to be interpreted as a system of beliefs and desires … according to which raising one's hand would be a good means to one's ends, and if another state is to be interpreted as an intention to raise one's hand, then the former had better typically cause the latter [2]. Likewise on the input side. A state typically caused by round things before the eyes is a good candidate for interpretation as the visual experience of confronting something round [3]; and its typical impact on the states interpreted as systems of belief ought to be interpreted as the exogenous addition of a belief that one is confronting something round, with whatever adjustment that addition calls for [4]. (1983a, p. 374)

There are four causal relationships mentioned here, marked [1] to [4], which are represented schematically in Figure 1. But we can be more systematic still, if make use of the 'approximately descriptive' Bayesian models. Thus, Lewis asks us to consider …

> … an oversimplified picture of interpretation as follows … P is a probability distribution over the worlds, regarded as encapsulating the subject's dispositions to form beliefs under the impact of sensory evidence: if a stream of evidence specified by proposition e would put the subject into a total state S – for short, *e yields S* – we interpret S to consist in part of the belief system given by the probability distribution $P(\cdot|e)$ that comes from P by conditionalisation on e. U is a function from worlds to numerical desirability scores, regarded as encapsulating the subject's basic values: if e yields S, we interpret S to consist in part of the system of desires given by the $P(\cdot|e)$-expectations of U. Say that P and U *rationalise* behaviour b after evidence e iff the system of desires given by the $P(\cdot|e)$-expectations of U ranks b at least as high as any alternative behaviour. Say that P and U *fit* iff, for any evidence-specifying e, e yields a state that would cause behaviour rationalised by P and U after e. That is our only constraining principle of fit. (Where did the others go? – We built them into the definitions whereby P and U encapsulate an assignment of content to states.) (1983a, p. 374; symbols altered for consistency)

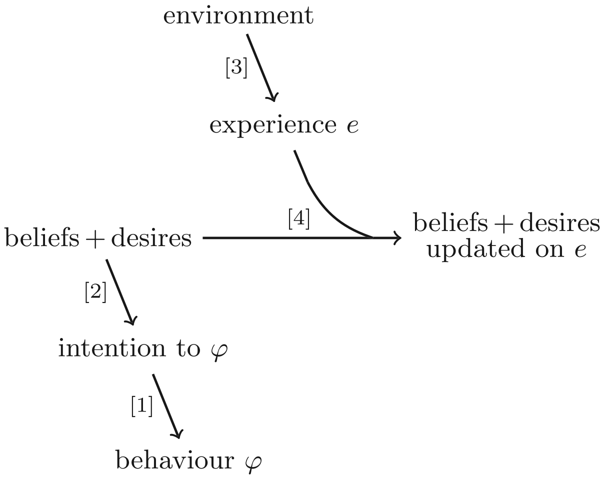

Figure 1 Causal role of belief-desire systems.

There's a lot going on in that quote, so it'll help to start by breaking down the model to which Lewis appeals. It has three parts:

Synchronic Coherence. At any time, an agent's beliefs can be represented by a probability distribution P over the space of possible worlds; her basic desires can be represented by a utility distribution U over that same space of worlds; and the strength of her desire for any proposition p is given by the $P(\,\cdot\,|p)$-weighted average U-value of the worlds where p is true.

Diachronic Coherence. If an agent at time t has a system of attitudes represented by $\langle P, U\rangle$, and her sensory evidence from t up to some later time t' is given by the proposition e, then her system of attitudes at t' will be represented by $\langle P(\,\cdot\,|e), U\rangle$.

Expected Utility Maximisation. Given a 'suitable' partition of propositions specifying how the agent behaves, the agent will behave so as make true the one (or one of the ones) she desires most.[3]

Aficionados will recognise this as involving an evidential decision theory, cashed out in a 'Jeffreyan' as opposed to 'Savagean' framework (cf. Savage, 1954; Jeffrey, 1965). I'll be referring to it again below. Inasmuch as the decision rule diverges from Lewis' own preferred causal decision theory (cf. 1981), those differences shouldn't much matter to the discussion.

[3] Regarding 'suitability': in (1981, pp. 7–8), (1986, p. 37), and (2020, Letter 714 [1993]), Lewis mentions that an agent's option partition ought to be such that she can make each option true at will, and there shouldn't be any more specific ways of making them true for which she likewise has that capacity. See also (1994, p. 417) for discussion on how the behaviour-specifying propositions that make up an agent's options might be understood.

Given that, there are two main parts to the principle of fit that Lewis describes. The first relates the content of a belief-desire system to what I will be calling the agent's *evidence-counterfactual dispositions* – how the agent would behave given different sequences of evidence. (This concept will be important later on.) It might be reconstructed like so:

> A state S consists in part of the system of beliefs and desires represented by $\langle P, U\rangle$ only if, for any evidence-specifying e, if those in S who receive evidence e go into state S', then those in S' will behave so as to maximise expected utility relative to $\langle P(\cdot|e), U\rangle$

That is *not* the complete principle that Lewis puts forward. Content codifies functional role, and the functional role of a belief-desire system is not exhausted by the evidence-counterfactual dispositions with which it's associated. For example, another important part of the role of any system of beliefs and desires is that it's linked to other such systems before and after via updating on evidence. The partial principle does not capture these inter-system connections. Consider: if S is to be interpreted using $\langle P, U\rangle$, and S plus e leads to S', then S' should (all else equal) be interpreted using $\langle P(\cdot|e), U\rangle$. However, the partial principle by itself does not entail this. For suppose that the latter state S' is associated with evidence-counterfactual dispositions that 'fit' the $\langle P(\cdot|e), U\rangle$ interpretation; still, there may be some alternative interpretation, $\langle P', U'\rangle$, that *also* fits, and the partial principle says nothing about whether we ought to prefer $\langle P(\cdot|e), U\rangle$ over $\langle P', U'\rangle$.

For the more complete statement of the principle, we need to include 'the others' that Lewis mentions are built into 'the definitions whereby P and U encapsulate an assignment of content to states'. Specifically: if the agent is in some state S' that consists in part of the system of attitudes given by $\langle P, U\rangle$, and S' when fed evidence e leads to S, then we are supposed to

> … interpret S to consist in part of the belief system given by the probability distribution $P(\cdot|e)$ [and] in part of the system of desires given by the $P(\cdot|e)$-expectations of U.

These 'definitions' are intended to capture the connections between systems of attitudes that go via evidence. Explicitly writing them into the principle gives:

> A state S consists in part of the system of beliefs and desires represented by $\langle P, U\rangle$ only if, for any evidence-specifying e, if those in S who receive evidence e go into state S', then S' will consist in part of the system of beliefs and desires represented by $\langle P(\cdot|e), U\rangle$, *and* those in S' will behave so as to maximise expected utility relative to $\langle P(\cdot|e), U\rangle$

Consequence: if S consists in part of the system of beliefs and desires represented by $\langle P, U\rangle$, and S follows from S' when that prior state S' is fed evidence e, then S' should itself consist in part of a system represented by some

$\langle P', U \rangle$ such that P' conditionalised on e is P; and hence, also, S should cause behaviour that maximises expected utility with respect to $\langle P, U \rangle$. The full principle thereby systematically links the content of a system of beliefs and desires given by $\langle P, U \rangle$ more thoroughly to its total functional role. On the input side: it connects $\langle P, U \rangle$ to the belief-desire systems that precede it, if such there are, as well as to the sensory experiences that provide the evidence on which those prior systems are updated. And on the output side: it connects $\langle P, U \rangle$ directly to behaviour, and to later belief-desire systems that follow from $\langle P, U \rangle$ after updating in light of new experiences.

I reiterate that what I've described here is only an approximation to a single principle of fit among many. It is designed to capture the gist of the functional role of a system of beliefs and desires in relation to its immediate causal inputs (sensory experiences and prior belief-desire systems) and causal outputs (behaviours and later belief-desire systems). There is still much more to be said, both in making the principle of fit for beliefs and desires more realistic, and in explicitly spelling out the principles for other kinds of intentional states. Missing from Lewis' 'oversimplified picture' is any representation of intentions, which sit as intermediaries between belief-desire systems and behaviour as per relations [2] and [1] respectively in Figure 1. Nor is there any accounting for change in basic desires over time. (Lewis mentions both of these among several other 'dire' oversimplifications of his toy model.) Likewise, we still need to spell out the principle of fit for sensory experiences, which are linked to external properties on the input side and systems of belief and desire on the output side, as per relations [3] and [4], respectively. But one can readily imagine how those principles might go, at least in outline, and Lewis *does* go into some detail regarding fit for experiences in (1997) and (2020, Letter 501 [1991]) – the rough idea being that a state should be interpreted as an experience of some secondary quality F being present only if (a) in normal circumstances, something's being F is part of what brings that state about, and (b) the state in turn is part of what gives rise to a system of beliefs and desires updated on the content that something F is present.

2.4. THE PRINCIPLES OF HUMANITY

That's enough for the principles of fit. They make up the main part of the theory of content, but alone they won't suffice to pin down our indexing scheme[4]:

[4] Lewis variably refers to the 'second part' of his theory of content in terms of the principles (plural) of charity, the principles of humanity, and in terms of reasonableness, intelligibility, and/or eligibility. The inconsistent terminology does not make for easy exegesis. I've chosen 'principles of humanity' to help distinguish them from the Charity principle described in 'Radical Interpretation'. They are related, to be sure, but they are not identical, and being clear about the difference is useful for avoiding some of the misunderstandings that I'll discuss later.

> [A theory of content], I said, should have two parts. One part says what it is for an assignment of content to states to fit the functional roles of the states … But principles of fit can be expected to underdetermine the assignment of content very badly … Therefore a theory of content needs a second part: as well as principles of fit, we need 'principles of humanity', which create a presumption in favour of some sorts of content and against others. (1986, p. 107)

In 'New Work' (1983a, pp. 374–375), Lewis provides a condensed argument for this underdetermination thesis, formulated using the simplified Jeffreyan model he employs for that discussion. He provides another informal version of the argument in *Plurality* (1986, pp. 37–38). The details of these arguments needn't concern us, and they've been discussed quite thoroughly already (see especially Williams, 2016). What's important is just that we can have two or more rather different indexing schemes that diverge in how they assign contents, and yet do equally well according to the principles of fit.

We therefore require additional constraints on our indexing scheme. These will be constraints directly on contents, independent of functional role:

> The saving constraint concerns the content – not the thinker, and not any channels between the two … Believing this or desiring that consists in part in the functional roles of the states whereby we believe or desire, but in part it consists in the eligibility of the content. (1983a, p. 375)

Compare again the measurement of physical quantities. We assign numbers to quantities such that relations between them are usefully mirrored in the relations between the numbers assigned. That's what's essential for the numerical indexing scheme to perform its function, but more than one indexing scheme will do the trick. The choice of unit and zero point provides us with further constraints in the case of temperature – 'fixed points' with which we can pin down the remainder of the scheme. Likewise for the attitudes. Contents are indices assigned to intentional states in the first instance to codify each state's location in a causal network. That's what's essential for the scheme to perform its function, and that's what's covered by the principles of fit. The problem is that more than one scheme will do the trick; the solution is further constraints.

The extra 'fixed points' come in the form of restrictions on the *reasonableness* or *intelligibility* of contents, which primarily serve to break ties between schemes that are otherwise equally good:

> We need further constraints … 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 she has lived. […] These principles will select among conflicting interpretations that equally well conform to the principles of fit. They impose a priori – albeit defeasible – presumptions about what sorts of things are apt to be believed and desired; or rather, about what dispositions to develop beliefs and desires, what inductive biases and basic values, someone may be rightly interpreted to have. (1983a, p. 375)

Lewis refers to these all as rationality constraints, but admits also that he employs an unusually broad sense of 'rationality' (see 1986, pp. 38–39; 1994, p. 428). Whatever we call them, though, the important part is that the additional constraints belong to folk psychology. There are three types that Lewis explicitly distinguished, regarding (i) what kind of basic desires might be assigned, (ii) what kinds of 'inductive biases' we might have, and (iii) what kinds of contents we might believe and desire. As he summed them up in 'Reduction of Mind':

> Folk psychology sets presumptive limits to what basic desires we can have or lack: *de gustibus non disputandum*, but still a bedrock craving for a saucer of mud would be unintelligible. Likewise it sets limits to our sense of plausibility: which hypotheses we find credible prior to evidence, hence which hypotheses are easily confirmed when their predictions come true. And it sets limits on what our contents of belief and desire can be … Especially gruesome gerrymanders are *prima facie* ineligible to be contents of belief and desire. (1994, pp. 416–417)

The third 'rationality' constraint relates to *naturalness* – agents should not be interpreted such that they expect unexamined emeralds to be grue, nor as having a basic desire for a-long-life-unless-one-was-born-on-Monday-and-in-that-case-life-for-an-even-number-of-weeks, for example (1983a, p. 375; 1986, p. 107). Later commentators have focused much on this aspect of Lewis' view, but naturalness was only ever one factor among several that go into the principles of humanity (see 1983a, p. 375; 1986, pp. 38, 107; 1994, pp. 416–417; 1996, p. 306; 2020, Letter 499 [1991]; see also Schwarz, 2014). I don't have a lot to say about it, because it doesn't strike me as a particularly essential part of the Lewisean view and serves mostly as a distraction.

2.5. ANALYTIC FUNCTIONALISM

The principles of fit and humanity together characterise the theory of content, which should systematically capture what folk psychology says about how our many intentional states relate to one another, to sensory inputs, and to behavioural outputs. Lewis never claimed to have spelled the theory out to completion. He provided only sketches of the principles of fit, and said even less about the principles of humanity – what you can see in the quotes above is more or less the extent of it.

Perhaps Lewis didn't think it was his place to fully spell the principles out. As he said in 'Reduction of Mind': 'I offer not analyses, but a recipe for analyses' (1994, p. 416). In any case, that recipe will be familiar to most readers so I'll be brief. (For more detailed discussion, a good starting place is Lewis' own 1970 and 1972.) Once we have our theory of content, and thus, a specification of the causal roles that our intentional states are supposed to enter into, we can use the standard Ramsey–Carnap–Lewis method to extract explicit definitions:

> [Folk psychology] associates with each mental state a typical causal role. Now we have our recipe for analyses. Suppose we've managed to elicit all the tacitly known general principles of folk psychology. Whenever M is a folk-psychological name for a mental state, folk psychology will say that the state M typically occupies a certain causal role: call this the M-role. Then we analyse M as meaning 'the state that typically occupies the M-role'. (1994, p. 416)

Or more accurately,

> Suppose, for instance, that folk psychology had only three names for mental states: L, M, N. We associate with this triplet of names a complex causal role for a triplet of states, including causal relations within the triplet: call this the LMN-role. Folk psychology says that the states L, M, N jointly occupy the LMN-role. That implies that M occupies the derivative role: coming second in a triplet of states that jointly occupy the LMN-role. Taking this as our M-role, we proceed as before. Say that the names L, M, N are *interdefined*. The defining of all three via the LMN-role is a package deal. (1994, p. 416)

Or more accurately still: if no states jointly occupy the LMN-role, the world may still supply us with so-called imperfect deservers; thus, we say, finally, the folk-psychological names will refer to whatever states jointly come closest to jointly occupying the LMN-role, provided they come close enough.[5]

It's important to note that folk psychology itself admits two kinds of exceptions to the causal roles it posits. There are *individual-level exceptions*: there might be some occasions in which an individual's mental states do not have the kinds of causes and effects they're normally supposed to have, even if for that individual they usually do. And then there are *population-level exceptions*: there might be some few individuals within a population whose intentional states rarely if ever have the kinds of causes and effects they're normally supposed to have (1980a; 1983b; 2020, Letter 436 [1974]). The latter are important. Folk psychology posits inner mental states that can and do recur in many individuals throughout a given kind (e.g., humankind), and which have 'normal' causal properties defined relative to that kind. The physical states with which they are identified must therefore also be the kinds of states that can recur across many individuals in a kind, and which have normal causal properties defined relative to that kind. The anti-individualism that results is not an optional extra to Lewis' view; it is a direct implication of his analytic functionalism.

[5] Lewis attempted no precise account of what makes a potential role-occupant 'good enough', nor how we might go about deciding between competing imperfect deservers. He would have seen doing so as a fruitless endeavour, for several reasons. For one, it was Lewis' general presumption that folk psychology will prove more or less accurate in the large majority of actual causes – thus, the choice between imperfect deservers is unlikely to make much of a difference to anything except in rare cases or strange hypothetical scenarios. Moreover, there's no point in seeking precision where there's no precision to be had.

The upshot is that when we're trying to determine which of Karl's physical states deserve to be identified with what intentional states, we need to consider in the first instance the causal properties of those states not only as they happen to be in Karl but also as they happen to be in others of Karl's kind, including those at nearby possibilities 'sufficiently similar in the anatomy of their inhabitants and in the relevant laws of nature' (1986, p. 39; see also 1980a; 1981, p. 14; 1983a, pp. 373–375; 1983b, pp. 119–121; 1994, pp. 416–418, 427–430; 2020, Letters 478 [1983], 486 [1988], 501 [1991], 503 [1993]). To solve the problem of radical interpretation for Karl, we must first solve the problem for Karl's kind.

To summarise: Lewis proposed a solution to the problem of radical interpretation that proceeds in four main steps:

Step 1. Specify the principles of fit and humanity; these will characterise the normal causal relations that hold between our intentional states, sensory inputs, behavioural outputs, and so on, according to folk psychology.

Step 2. Using the Ramsey–Carnap–Lewis method, extract a joint definition of all our intentional states in a vocabulary acceptable to the physicalist.

Step 3. Locate physical states that can serve as (good enough) occupants for the relevant folk-psychological roles relative to Karl's kind.

Step 4. Determine what beliefs and desires Karl has at each time by considering what physical states he is in at those times.

There are still many details I've left out – for example: constraints on what counts as an appropriate population for characterising kind-relative typical causal roles (1980a; 1983b); *de se* content and the proposal that properties are what really serve as the objects of belief and desire (1979); extending folk psychology to include psychophysical connections between experiences and secondary qualities (1997; 2020, Letter 501 [1991]); the relation between behaviour described under an intentional description and under a physical description (1994, pp. 416–417); doublethink and fragmentation of attitudes (1986, pp. 28–30; 1994, pp. 425–427) … the list goes on. I cannot do full justice to all of it in the course of a single paper. But I've said enough that we might now clearly recognise some important missteps in the recent literature.

3. *What Lewis' theory wasn't*

In the remainder, I want to highlight and address a number of common myths and misconceptions about Lewis' theory. Most of these, I think, can be traced back ultimately to the way that Lewis framed his ideas in 'Radical Interpretation' (1974). For that reason, I will start with a summary of the relevant parts of that paper.[6]

3.1. 'RADICAL INTERPRETATION'

The theory in 'Radical Interpretation' is simple enough. An 'interpretation' is here understood to be an assignment of (graded) beliefs and desires to an agent at a time. The correct interpretation of Karl at a time is defined as the one, if there is just one, that does best overall with respect to the two principles of interpretation, *Rationalisation* and *Charity*, and if there's more than one tied for equal best then the truth is indeterminate between them.

Rationalisation is the simpler of the two principles. In Lewis' words,

> … the beliefs and desires ascribed to Karl should be such as to provide good reasons for his behaviour, as given in physical terms by the physical facts. Thus, if … Karl's arm goes up at a certain time, [we] should ascribe beliefs and desires according to which it is a good thing for his arm to go up then. (1974, p. 337)

Lewis immediately goes on to spell the principle out in decision-theoretic terms:

> 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. A precondition: those ascribed beliefs and desires should be coherent enough to permit the comparison of expected utilities of alternative ways of behaving. (1974, p. 337)

Note the 'coherent enough'. Note also that there's no mention of rationalising Karl's behavioural dispositions – the principle as stated tells us to assign attitudes to Karl as would provide good reasons for his *behaviour*, nothing more. It's unclear whether this terminology was deliberate, but it was consistent.

Next is Charity. At its most abstract level, the principle says that 'Karl should be represented as believing what he ought to believe, and desiring what he ought to desire' (1974, p. 336). But what is it that Karl ought to believe and desire? Lewis doesn't commit to anything very specific, but he does offer suggestions:

[6] The claim here isn't that 'Radical Interpretation' presents a picture that's vastly different from the one described in Section 2. Of course it doesn't! But it *is* different, or (inclusive 'or') can be *reasonably interpreted* as different. The point is that even small or subtle differences in presentation can snowball into more significant misunderstandings, and it's my hypothesis that this is what's happened with at least most of the myths discussed.

> Perhaps an improved Principle of Charity would require Karl's beliefs and ours to be related as follows: there must exist some common inductive method *M* which would lead to approximately our present systems of belief if given our life histories of evidence, and which would likewise lead to approximately the present system of beliefs ascribed to Karl if given Karl's life history of evidence according to [the physical facts]. As for desires: there must exist some common underlying system of basic intrinsic values *U* which would yield approximately our systems of desires if given our systems of beliefs, and which would likewise yield approximately the system of desires ascribed to Karl if given the system of beliefs ascribed to Karl … (1974, p. 336)

There's several parts to Charity as characterised here. One part concerns Karl's basic desires. The idea appears to be that we tend to ultimately care about the same sorts of things, and so we should avoid attributing any basic desires that would deviate too far from our own.[7] Another part concerns the relation between Karl's beliefs and his sensory evidence. Agents should be interpreted as adhering to a 'common inductive method', such that if they have the same life history of evidence then they should end up with approximately the same beliefs. Lewis doesn't say anything further about the 'common inductive method', but given his writings elsewhere it's clear enough he was imagining it would look something approximately but not exactly like conditionalisation (e.g., 1980b, p. 288; 1983a, p. 374; 1994, p. 428; 2020, Letters 503 [1993], 737 [1999]). For any such method, where we end up after updating on our evidence depends in part on where we started – the agent with high initial confidence they're living in a world filled with grue emeralds, say, or in a counter-inductive world, will have rather different beliefs than our own even given the very same evidence. Thus, Charity also constrains the kinds of beliefs we ought to ascribe to Karl before he receives any evidence at all (cf. 1980b, p. 288; 1983a, p. 374; 1994, p. 417).

(Given this, one might roughly think of Charity as a mixture of the principles of humanity – at least those dealing with basic desires and inductive biases – and a part of the principle of fit from 'New Work' – the part dealing with the connections between belief-desire systems via updating on evidence. But only roughly: Charity in 'Radical Interpretation' is a constraint on how one should interpret an individual given only facts about their evidence, whereas the principles of fit and humanity are constraints on how folk psychology indexes mental states with contents given their respective locations in a typical causal network. There's more than one difference there, and those differences are important.)

And that's it. In the final pages of 'Radical Interpretation', Lewis writes as though Charity and Rationalisation are supposed to constitute the full and complete account of how one might go about solving the problem of radical interpretation:

[7] The 'we' refers to humans – Karl is not just any agent, but a *human* agent (see 1974, p. 335). The 'common underlying system of basic intrinsic values' is presumably common to humans; doubtless Lewis would have allowed that Martians might have basic desires that differ from ours.

> … using the physical facts both as a source of information on Karl's behaviour and as a source of information on his life history of evidence, fill in his attitudes completely by means of the Rationalisation Principle and the Principle of Charity. (1974, p. 341)

There are two possibilities. If an interpretation perfectly satisfies both principles, Lewis conjectures, then it will do so uniquely (1974, p. 343). He offers no proof, although it does come with the caveat that should it be false then that falsity will merely show that we've yet to specify all the relevant constraints. If no perfect interpretations exist, then the correct interpretations – there will likely be more than one – will be those that reach the best compromise between Charity and Rationalisation.

3.2. THE MYTH OF INTERPRETIVISM

Onward now with the myths. I start with the most straightforward and easiest to debunk. It goes like this. Charity and Rationalisation tell us how to attribute systems of belief and desire *directly* to Karl, given the facts about *his* evidence and *his* behaviour, respectively. Because the correct interpretation of Karl is then defined simply as that which does best overall with respect to Charity and Rationalisation, 'Radical Interpretation' implies that the facts about Karl's beliefs and desires are grounded wholly in the facts about his evidence and behaviour: *what it is* for Karl to have the beliefs and desires he has *just is* for him to have certain patterns of behaviour in response to certain sequences of evidence.

This, I take it, is a paradigmatically *interpretivist* position, and one that's frequently attributed to Lewis. (See, e.g., Fodor and Lepore, 1992; Eriksson and Hájek, 2007, pp. 199–202; Hattiangadi, 2019; Hattiangadi and Stefánsson, 2021, pp. 6478–6479). But Lewis explicitly rejected precisely this kind of picture on more than one occasion (see 1981, p. 14; 1983b, pp. 119–121; 1986, pp. 39–40; 1994, pp. 428–429; 2020, Letter 503 [1993]). In 'Reduction of Mind', he described it as a species of behaviourism:

> A behaviourist analysis might say, roughly, that a subject's beliefs and desires are those beliefs and desires, attribution of which would best make sense of how the subject is disposed to behave, and of how his changing behavioural dispositions depend on the changing perceptible features of his surroundings. But [my opponent] is a robust realist about beliefs and desires. He takes them to be genuine inner states, and causes of behaviour. He won't like an analysis that dispenses with efficacious inner states in favour of mere patterns of dispositions […] I applaud these misgivings. I too am a robust realist about beliefs and desires. (1994, p. 428)

It's for this reason that Lewis wrote that 'Radical Interpretation' was both 'too behaviouristic' (2020, Letter 499 [1991]) and 'unduly individualistic' (1983b, p. 119). These are two sides of the same coin. Rather than assigning beliefs and desires directly to agents, Lewis thought it better to identify states of belief and desire with recurrent inner physical states on the basis of their kind-relative typical causal roles (1986, p. 39). Consequently, the correct interpretation of Karl needn't be that which best rationalises his behaviour relative to his evidence:

> Karl might believe himself a fool, and might desire fame, even though the best interpretation of Karl considered in isolation might not assign those attitudes to him. For the best interpretation of Karl's kind generally might be one that interprets two states respectively as a belief that one is a fool and a desire for fame, and Karl might be in those two states. (1983b, p. 119)

The non-individualism was part of Lewis' stance from the beginning. For example, a kind-relative notion of 'typicality' is necessary for how Lewis proposes to handle the case of the total paralytic in his very first publication (1966, p. 22): the causal role that the firing of C-fibres typically plays *in the paralytic* is very different from the role it typically plays *in others of the paralytic's kind*, and it's the latter we need if we're going to say the paralytic is in pain by virtue of being in a state that typically occupies the pain-role. For the same reason, in a 1974 letter to Sydney Shoemaker, Lewis highlights that it's important for functionalists to use of a kind-relative notion of 'typicality', because otherwise they'll be 'no better off than a behaviourist in providing for the amputated brain whose experiences do not occupy at all their proper causal roles' (2020, Letter 436 [1974]).

So first question: why did Lewis write 'Radical Interpretation' the way he did? At least part of the answer is that he was framing it in response to Davidson:

> But see ['Radical Interpretation'] with caution: it began as a conversation with Donald Davidson, and I went rather too far in granting undisputed common ground … I ignored the possibility that deviant Karl might believe something in virtue of the causal role of his inner state not in Karl himself but in others who are more typical members of Karl's kind. (1994, p. 429)

For Davidson, principles of interpretation telling us how to interpret agents on the basis of their behaviour really do seem to have definitional status: *what it is* for the agent to have the beliefs and desires she does *just is* for her to be interpreted as such by an application of those principles. For Lewis, the 'principles of interpretation' are just a framing device. They're a stylised means of expressing complementary parts of what folk psychology says about the normal causal roles of our attitudes in relation to evidence, behaviour, and each other, as well as what folk psychology says about 'rational' restrictions on basic desires and inductive biases (1974, p. 334). They map *approximately* on to the principles of fit and humanity, which supply the theory of content from which a functional analysis can be extracted, but they do not directly define Karl's attitudes.

Followup question: was Lewis an interpretivist? Well, that depends – ask five philosophers what interpretivism is supposed to be, exactly, and you'll get ten different answers. If 'interpretivism' just picks out any theory according to which standards of rational belief, desire, and action are somehow constitutive, then Lewis counts. But that's just another way to say that he was an analytic functionalist, and he thought the central roles of beliefs and desires within folk psychology are mostly characterised by rational norms. If *that's* interpretivism, then it's rather different from what we find in Davidson, and it doesn't seem there's much value in muddying the water with vague labels.

3.3. THE MYTH OF INDEPENDENCE

Next is the *Myth of Independence*. This one is closely related to the *Myth of Interpretivism*, but more subtle. The rough idea is that Charity and Rationalisation are often taken to impose *independent* constraints on interpretation – the former a constraint on attitudes in relation to evidence, the latter a constraint on attitudes in relation to choices or behaviour.

It'll help if I use a specific example, so consider Hattiangadi's (2019) 'voting' formulation of Lewis' account. It goes like this. We have a set of *candidate interpretations* of Karl (i.e., alternative assignments of beliefs and desires), and we have three *voters*: Ms Charity, Ms Rationalisation, and Ms Naturalness. Each has their own *preferences* (i.e., their own opinions regarding how candidate interpretations ought to be ranked from best to worst), based on what they individually think is important. For Ms Charity, what matters is how well the interpretation satisfies an evidential constraint – she prefers an interpretation to the extent that it renders Karl epistemically rational given the facts about his evidence. For Ms Rationalisation, what matters is how well the interpretation satisfies a behavioural constraint – she prefers an interpretation to the extent that it renders Karl pragmatically rationality given the facts about his behaviour. Finally, Ms Naturalness doesn't care about Karl's evidence or behaviour, only the relative naturalness of the contents of his attitudes. The correct interpretation is then one that does the best job of balancing the preferences of each voter off against the others.

There can be no doubt that 'Radical Interpretation' suggests something along these lines. Charity imposes a constraint relating to evidence; it does other things, too, but what matters is that Charity says nothing about behaviour. Rationalisation imposes a constraint relating to behaviour, and says nothing about evidence. And, as Lewis says, if no interpretation fits all the constraints perfectly then we will need to 'strike a balance' between their competing recommendations (see 1974, p. 343). This lends itself naturally to being formulated within the framework of social choice theory, in *more or less* the way Hattiangadi suggests.

Nevertheless, it's not the right way to understand Lewis' theory. The thing to note in particular is the manner in which the evidential and behavioural constraints are taken to contribute *independently* to the final result. If the correct interpretation is some optimal compromise between preferences of Ms Charity and Ms Rationalisation (plus any further voters), then correct interpretation can depend only on what those voters' preferences depend on in turn – anything deemed irrelevant according to each voter *independently* cannot be a factor in what determines correct interpretation. And that's a problem, for it entails we ought to ignore most facts about evidence-counterfactual dispositions.

Imagine a simplified scenario, as depicted in Figure 2. At time t_0, Karl begins in state S_0, whereupon he might receive either evidence e_1 (which would put him in state S_1 at t_1) or e_2 (which would put him in state S_2 at t_1). As a matter of fact, he receives and then updates on e_1, and so goes into S_1. Then at t_1, he might go on to receive either evidence e_3 (putting him in S_3 at t_2) or e_4 (putting him in S_4 at t_2). Each state is causally associated with a behaviour – for example, S_0 causes b_0, S_1 causes b_1, and so on. Given that, suppose now we want to interpret Karl at t_1, while he's in S_1. Ms Charity would rank interpretations by how well they make sense relative to the evidence e_1 received between t_0 and t_1. We may also factor in reasonableness constraints on initial beliefs and basic desires – this matters not for the present point. What does matter is that Ms Charity has no interest in Karl's actual or counterfactual behaviour, she cares only for his actual history of evidence up to t_1. Ms Rationalisation would instead rank candidate interpretations by how well they rationalise the behaviour b_1. Ms Rationalisation may or may not also factor in Karl's momentary choice dispositions – for example, how he would

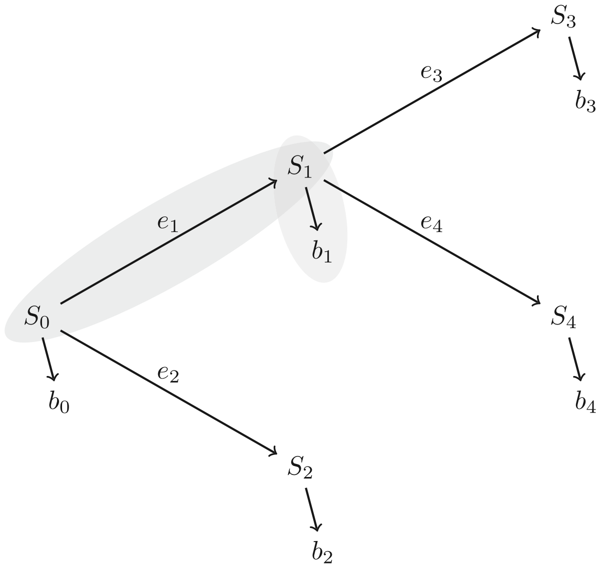

Figure 2 What matters to Ms Charity and Ms Rationalisation.

have chosen among his remaining options at t_1 should his first choice have not been available. I'll talk about that more below, but again it matters not for the present point. What does matter is that Ms Rationalisation has no interest in Karl's evidence or his evidence-counterfactual dispositions. Nothing in Lewis' description of the principle tells us to consider the behaviours Karl would have at t_1 if he had received some alternative sequence of evidence up to that point, nor the various ways he would go on to behave at t_2 given this or that future sequence evidence. The facts about Karl's evidence-counterfactual dispositions are irrelevant to the two 'voters' individually, and consequently irrelevant to which interpretation affords the best compromise between their recommendations.

But this directly conflicts with what's implied by the principle of fit Lewis gives in 'New Work'. According to that principle, the interpretation of the state S_1 should depend in part on the interpretation of the states actually and counterfactually downstream of S_1, and hence derivatively on the behaviours associated with those states. So S_1 should be assigned some probability-utility pair $\langle P, U\rangle$ only if b_3 is rationalised by $\langle P(\cdot|e_3), U\rangle$ and b_4 is rationalised by $\langle P(\cdot|e_4), U\rangle$. But that's not all: the principle also implies that S_1 should be assigned $\langle P, U\rangle$ only if S_0 is assigned some $\langle P', U\rangle$ such that $P'(\cdot|e_1) = P$. Consequently, S_1 should be assigned $\langle P, U\rangle$ only if the behaviour b_2 caused by S_2 – the state Karl would have been in had he received e_2 after t_0 – is rationalised by $\langle P'(\cdot|e_2), U\rangle$. In short: the causal role of the attitudes in relation to behaviour and their causal role in relation to evidence *interact* to generate a network of possible states and behaviours linked by actual and potential causal relations, with the interpretation of any state in that network constraining the interpretation of every other.

Between the two of them, Ms Charity and Ms Rationalisation are ignoring a great deal of relevant information – all that matters for them is Karl's actual life history of evidence and his actual sequence of behaviours. But evidence-counterfactual dispositions are important for any sensible functionalist theory of the attitudes. Consider: Karl might fear clowns by being in a state that disposes him to run away from the sight of red noses and silly shoes. Such dispositions need never manifest if the circus never comes to town, but still they're an important part of *what it is* for Karl to have that fear. Likewise, part of what makes a system of attitudes *what it is* is that it is poised to bring about any number of different later systems of attitudes, and hence potentially many differing behaviours, conditional on what evidence it happens to be updated upon; and furthermore, that it will itself have, in most cases, be the result of some earlier system of attitudes that was updated on some evidence. So we need consider evidence-counterfactual dispositions. And Lewis knew it:

> Roughly, what makes it so that a certain credence function is yours is that you are disposed to act in more or less the way that it rationalises. (Better: what makes it 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 rationalised by the pair of them together, taking into account the modification of credence by conditionalising 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; and further, no other such pair would fit your dispositions more closely.) (1980b, pp. 287–288; see also 1975, p. 548; 1983a, p. 374; 1986, p. 37)

Let me be clear: the issue is not with the use of social choice theory *per se*. Rather, the problem is in failing to properly carve out the independent factors that go into the assignment of content. Those are not the *evidential* versus *behavioural* constraints (as encoded by Charity and Rationalisation), but the *causal* versus *non-causal* constraints (as encoded by the principles of fit and humanity). It would be not be inaccurate to represent the combined role of fit and humanity within the theory of content by means of a voting metaphor – so long as we keep squarely in mind that fit is foremost, and humanity serves primarily to filter between those schemes with equal fit. But it *is* an error to represent the evidential and behavioural constraints as separable, each to be considered independent of the other. They are two ineliminable parts of a single complex causal role, and for a functionalist theory of the attitudes their interaction matters.

3.4. THE MYTH OF SOURCE INTENTIONALITY

Next is the *Myth of Source Intentionality*. The idea here is that certain ‘lower-level’ intentional facts – namely, those about Karl’s sensory experiences and how he decides between which options – need to be fixed in place *before* Charity and Rationalisation can be applied to determine the facts about Karl’s beliefs and desires. (The expression ‘source intentionality’ is borrowed from Pautz, 2013).

If we think of Charity and Rationalisation as taking the facts about evidence and choice as ‘inputs’ respectively and then spitting out recommendations for which interpretations are better than others, then those principles can jointly *define* what it is for an agent to have the beliefs and desires she does only if the appropriate evidence-facts and choice-facts can themselves be specified prior to or otherwise independently of the facts about beliefs and desires – else circularity threatens. We thus find this Myth underlying the oft-expressed concern that Lewis’ solution to the problem of radical interpretation requires an independent reductive theory of sensory evidence and intentional choice:

> … Karl's history of experiences and hence evidence play a crucial role in constitutively determining the contents of his beliefs. Given this, what experiences and evidence he has cannot in turn be pinned down by his beliefs. That would be circular. To avoid circularity, his experiences and their contents must be determined in a belief-independent way, for instance, by causal connections to the world. (Pautz, 2013, p. 231 n. 31)

> … [Lewis' interpretationism] does not give us a complete characterisation of intentional states of Karl in terms of a purely physicalistic description of Karl's behaviour and environment. Rather, it takes as given certain 'low-level' intentional states of his (his evidence and choices, or at least a certain range of them) and, using the device of interpretation, fills out the rest of Karl's contentful states. If we want a theory that fully characterises intentionality in terms of the non-intentional, we'd need add an independent, prior theory that tells us how to attribute a basic set of evidence propositions and choices to an agent. (Brouwer *et al.*, 2021, p. 3381)

> [The 'Lewisean' theory], even if it succeeds, only reduces the belief and desire facts to evidence and action facts. If the facts about Karl's evidence and actions are themselves representational facts, or if they are grounded in representational facts, then the proposal does not give us a completely reductive theory of the representational to the non-representational. (Buchanon and Dogramaci, Forthcoming, p. 5)

'Radical Interpretation' clearly *suggests* a two-stage approach. (I wouldn't go so far as to say the paper *unambiguously* suggests a two-stage approach, but there's a reason this interpretation of Lewis is so common!) But you can guess what I'll say next: it is not the right way to understand Lewis' theory. I'll start by discussing 'source intentionality' in connection to Charity.

3.4.1. Evidence and experience

Here's one reason why the reliance on some prior unreduced 'source intentionality' might be considered problematic in the case of Charity: it is plausible that the contents of our experiences might be influenced by our prior expectations. We sometimes see or hear what we want or expect to see or hear. Thus, if Lewis requires us to *first* pin down the facts about Karl's experiences *before* we appeal to Charity to help determine his beliefs, but the facts about Karl's beliefs can influence the facts about his sensory experiences, then the theory is doomed to failure. As Hattiangadi and Stefánsson (2021) put the worry, the Lewisean account will be either viciously circular or it will presuppose access to intentional facts regarding the content of experiences that aren't directly accessible just the physical facts; either way it's in trouble.

But while Lewis never said *much* about how sensory experiences are causally related to the external world, he did say enough to debunk this Myth. On Lewis' preferred picture, sensory experiences can be understood *in part* by their causal relations to the secondary qualities that (according to folk psychophysics) usually cause them under normal conditions. As he put it in ‘New Work’, a ‘state typically caused by round things before the eyes is a good candidate for interpretation as the visual experience of confronting something round’ (1983b, p. 374). According to Pautz,

> This suggests a *simple causal principle*: very roughly, if, in the relevant population, state *S* would be caused by something’s being *F*, in so-and-so range of actual and counterfactual cases, then *S* is an experience with the content *that something F is present*. (Pautz, 2013, pp. 222–223)

That’s the kind of principle one would expect given a two-stage solution to the problem of radical interpretation – the simple causal principle tells us how to determine the content of an experience in a wholly belief-independent way, and so provides us ‘source intentionality’. But that’s not what Lewis had in mind, for to characterise the contents of sensory experiences purely in terms of the properties that normally cause them is to focus only on the ‘backwards-looking’ part of their total functional role. Sensory experiences need to be analysed also in part by their ‘forwards-looking’ roles in connection to other intentional states (see 1972, p. 257 n. 15; 1994, pp. 416; 1997; 2020, Letter 501 [1991]). One key part of that role involves the changes they cause in belief:

> Part of the definitive role for a colour, and part of the definitive role for the corresponding colour-experience, is that the former typically causes the latter (at least on a cloudy day in Scotland). Of course there’s more to the roles than that … Part of the role of colour-experiences is to give rise to beliefs about identity over time of the things before one’s eyes. (2020, Letter 501 [1991]; see also 1983b, p. 374; 1994, p. 416)

A physical state that’s typically caused by round things before the eyes might be a good candidate for interpretation as the visual experience of confronting something round, all else equal – but if that state does not *also* cause the kinds of changes in belief and subsequent behaviours we’d normally expect of one who has seen round things before them, then it is *not* so good a candidate for interpretation as the experience of confronting something round after all.

The facts about external sensory inputs will be a *part* of what constrains the facts about sensory experiences and their contents, which in turn constrain the facts about beliefs and desires. But that was only ever taken to be one variable in a larger equation. The facts about behaviour will *also* constrain the facts about experiences, because they constrain the facts about beliefs which in turn constrain the facts about experiences. Beliefs, desires, experiences, intentions – all are implicitly interdefined by reference to their respective locations in a mess of causal interrelations tied down at both ends by sensory inputs and behavioural outputs. How we interpret any part of the causal network constrains how we might interpret the rest.

I'm not trying to argue that Lewis provided us with a full and complete theory of how the facts about an agent's sensory experiences reduce to the purely physical facts. He did not. (Whereupon the sceptic will complain it's still incomplete. Well, *all* philosophical theories about the reduction of sensory experience to physical facts are incomplete – how is this one worse off?) The point is that Lewis' theory of how we might reduce the facts about sensory experiences to the physical facts, and his theory of how we might reduce the facts about beliefs and desires to the physical facts, are *the same theory*. There are no necessary hierarchies of definitional priority among intentional mental states on Lewis' picture.

3.4.2. Choice and options

A similar circularity worry is sometimes raised in connection with Rationalisation. It starts with idea that Lewis' theory requires us to first fix the facts about how Karl chooses (or is disposed to choose) between which of his behavioural options, only after which we can appeal to Rationalisation to help decide which interpretations are better than others. But, the worry goes, we cannot understand what Karl's choices are, nor how he's disposed to choose between which options, unless we already know how he conceives of those options. Karl might lack confidence in his capacity to do something that he can in fact do at will, for example, and thus fail to treat it as an option when making his decisions. Consequently, in order to apply the principle of Rationalisation, we first need to know the facts about Karl's choices and what he takes himself to be choosing between. But these are intentional facts seemingly dependent on the very beliefs that we're trying to determine. (This concern is raised by Hattiangadi and Stefánsson, 2021, pp. 6480–6481, 6488–6489; see also Williams, 2019, pp. 169–174, for closely related discussion).

As it's described in 'Radical Interpretation', Rationalisation characterises a relationship between Karl's attitudes at a time, his behaviour at that time, and a 'suitable' partition of behaviour-specifying propositions – roughly, it tells us to assign attitudes such that, however Karl behaves at a time, that behaviour is the behaviour of one who maximises expected utility relative to a 'suitable' set of options. Nothing that Lewis says in that paper or elsewhere entails that we must determine what Karl's options are *before* we fix the facts about Karl's beliefs and desires. Lewis' statement of the principle is consistent with letting the facts about Karl's options be another variable to be determined alongside the others in the process of interpretation. That is, if it turns out that 'suitability' depends in part on the facts about Karl's beliefs, then we can suppose that Karl's options are determined alongside his attitudes under the constraints that, *inter alia*, (a) the options are 'suitable' relative to the attitudes assigned, and (b) his behaviour makes true the option (or one of the options) that has maximal expected utility relative to those beliefs and desires.

That '*inter alia*' is important, for the reasons I keep emphasising. If Karl's options depend in part on his beliefs, then they depend in part on what those beliefs depend on in turn. We should therefore not expect to be able to determine what Karl's options are independently of the facts about his evidence and his experiences. Consider, for example, Williams' (2019, p. 172) hypothetical 'option-fatalist', who is always certain there's exactly one thing she can do, which just so happens to be what she does. The option-fatalist never really faces a choice: her option partition consists of just one proposition that describes exactly what she does; everything else she deems impossible. The principle of Rationalisation, taken by itself in the absence of all other considerations, is *always* consistent with an option-fatalist interpretation. If that principle were all we had to go on, we could never rule such an interpretation out. But it's not all we have to go on. The option-fatalist's beliefs are not the kinds of beliefs we would expect of one who had a reasonable system of initial beliefs and subsequently updated in a more or less rational way on a normal life history of sensory evidence. Constraints on reasonableness plus fit with evidence will help to rule out option-fatalist interpretations in normal cases.

(We won't be in a position to know the details of Karl's evidence and prior beliefs *before* we know what his options are. But we don't need to; what's required for the point to go through is that the combined constraints on belief update plus reasonableness constraints on prior beliefs will likely rule out any option-fatalist interpretation given a normal life history of evidence. That I can confidently assert without knowing the details of Karl's evidence.)

Note that this strategy for dealing with the option-fatalist interpretation does not appear to be available on Williams' own explicitly two-stage theory of radical interpretation, according to which we first pin down the 'source intentionality' facts, and only then apply Charity and Rationalisation to solve for the best assignment of beliefs and desires (Williams, 2019, pp. 9–11, 167–172). Included among the 'source intentionality' facts are those about Karl's options – they need to be pinned down *before* anything else. On such a picture, then, there doesn't seem to be any scope for letting the facts about Karl's options depend in part on the facts about his beliefs, which in turn depend in part on the facts about his evidence and constraints on the reasonableness of his initial beliefs. What's needed – and what Williams aims to achieve – is an account of how an agent's options can be determined prior to and independently of their attitudes. That strikes me as hopeful, but more importantly unnecessary. Why not embrace the interrelation of beliefs and options, and let them be interdefined?

3.5. THE MYTH OF RATIONALITY MAXIMISATION

Often in conjunction with many of the foregoing myths is the *Myth of Rationality Maximisation* – that Lewis believed an agent's beliefs and desires are always (or even typically) those the attribution of which would make her *as rational as possible* given her evidence and behaviour.

This is probably one of the more common myths in the literature today. For examples of it in the wild, see (Buchak, 2016, pp. 801–802), (Thompson, 2016, p. 387), (Williams, 2018, pp. 47–48; 2019, p. 16), (Brouwer *et al.,* 2021, p. 3389), and for a rare example of authors recognising it *qua* myth, see Smithies *et al.* (2022). The myth also plays a crucial role in the critical arguments of Hattiangadi (2015; 2019, see esp. pp. 289–290) and Eriksson and Hájek (2007, esp. pp. 200–201), and is the primary target of Buchanon and Dogramaci (Forthcoming), according to whom Lewis endorsed:

> An assignment of beliefs and desires to Karl is correct in virtue of the fact that it best jointly maximises (1) his epistemic rationality given the evidence of his senses, and (2) his practical rationality given his dispositions to act.

Try as you might, though, you won't find Lewis anywhere saying that the process of interpretation ought to go via the maximisation of rationality. What he did say, sometimes, was that an agent's attitudes can – *in most cases*, *roughly* – be thought of as those that would *best rationalise* her behaviour given her evidence (e.g., 1980b, p. 288; 1981, p. 14; 1983b, p. 374; 1986, pp. 36–38).

Of course, if one equates 'best rationalises' with 'makes as rational as possible', then you can see how something like this myth gets going (cf. Williams, 2018, p. 47; 2019, p. 16). The problem is that we shouldn't be equating the two! For one thing, if the best rationalisation were the one that *maximises* the subject's rationality, then there could in principle be no such thing as *over*-rationalising – and Lewis made it clear that we should avoid over-rationalising interpretations (e.g., 1974, p. 337; 1994, p. 428). Moreover, if the best interpretation is to be understood as the one that maximises the subject's rationality, then this is the sort of thing that would need to make sense within the context of Lewis' analytic functionalism – and that would seem to be so only if the constitutive folk-psychological standards of rationality were themselves ideal standards of rationality. For it will only be in that case that the rationality-maximising interpretation will be systematically identical to the interpretation that most closely adheres to what folk psychology has to say. And we've seen already that this is very much not how Lewis conceived of folk psychology.

Still further evidence is provided in 'Reduction of Mind', where Lewis explicitly distinguishes two routes by which an agent might happen to be less-than-ideally rational (1994, p. 428). One route is when the agent's attitudes relate to her evidence and behaviour in a deviant or abnormal way, contrary to what folk psychology predicts. The other route is when the agent's attitudes relate to her evidence and behaviour in just the way they're supposed to according to folk psychology, which only requires conformity to near enough standards of rationality. (Thus, it's a prediction of folk psychology that even perfectly typical humans will be less-than-ideally rational.) If the constitutive standards of rationality were ideal standards, then the only route by which anyone might count as less-than-ideally rational would be the former – to be irrational at all would require deviation from the folk-psychological expectations.

To err is human, as the folk are well aware – they know better than to over-rationalise. The best rationalisation is not the one that maximises the subject's rationality. It's the one that best conforms to the imperfect, near-enough standards of rationality implicit in folk psychology.

3.6. RATIONALISATION BY REPRESENTATION THEOREM

In a letter to Michael McDermott, Lewis explains that his ideas on constitutive rationality were inspired by earlier work on the axiomatic foundations of decision theory – starting with Ramsey (1931), and followed later by, *inter alia*, Savage (1954) and then Jeffrey (1965) with the aid of Bolker (1967):

> What I mostly had in mind under the heading of constitutive rationality [was] the fit between actions and belief and desire. Yes, this does sound like Davidson – although without his irrealism and without his emphasis on the interpreter. But still more it should sound like decision theory, and the project of recovering subjective probabilities and utilities from the agent's dispositions to choose between gambles. This was a common source for Davidson and for me. (2020, Letter 503 [1993])

Lewis here alludes to what I'll call the *representational project*, which revolved around decision-theoretic representation theorems. These theorems are commonly taken to imply that if an agent's preferences satisfy certain constraints (or 'axioms'), then there will exist a unique probability-utility pair $\langle P, U\rangle$ that *represents* those preferences – specifically in the sense that those probabilities and utilities determine exactly those preferences via the theory's decision rule (e.g., expected utility maximisation).[8] If we then posit a very tight relationship between an agent's preferences at a time and her momentary choice dispositions, such that the latter suffice to determine the former and vice versa, then those axioms on preferences can be

[8] Being pedantic: Ramsey and Savage posit axioms on a binary preference relation that suffice to determine an expected utility representation $\langle P, U\rangle$, where the P is unique and the U is unique up to a positive linear transformation. Because utilities are usually taken to be measurable on nothing stronger than an interval scale, this means U is exactly as unique as it needs to be – hence I just say 'unique'. For Jeffrey's theorem, the base axioms won't suffice for a unique representation, however they will if supplemented appropriately (see Jeffrey, 1965, 119ff).

translated directly into axioms on the momentary choice dispositions that supposedly 'reveal' them. Given this, it was common for those engaged in the representational project to remark that subjective probabilities and utilities are 'really' just a way of representing facts about subjects' choice dispositions.

Hattiangadi & Stefánsson (2021, henceforth H&S) have recently objected to Lewis, arguing more or less that the success of the Lewisean theory depends crucially on the success of the representational project – so much so, in fact, that they might as well be the very same thing. Here's what H&S have to say. First, they assert that Lewis needs an a priori 'guarantee that the constraints on interpretation will be uniquely satisfied' (Hattiangadi and Stefánsson, 2021, p. 6478). This requirement – or, *purported* requirement – presents an obvious problem, because Rationalisation all on its ownsome is *prima facie* unlikely to pinpoint a single uniquely best interpretation. But luckily for Lewis, representation theorems promise a solution[9]:

> Lewis [was] struck by the power of [the Bolker–Jeffrey theorem], which *proves* that if an agent's preferences satisfy certain minimal constraints – the Bolker–Jeffrey axioms – it is possible to deduce probability and utility functions that can be understood as representing the agent's degrees of belief and desire. By appeal to this representation theorem, [Lewis] motivates the claim that there is an a priori entailment from the truths about Karl's preferences to truths about his beliefs and desires … (Hattiangadi and Stefánsson, 2021, pp. 6478–6479)

Because the Bolker–Jeffrey theorem 'proves' that we can deduce the facts about Karl's attitudes from the facts about his preferences, we need to know first what his preferences are. These we derive directly from the physical facts about behaviour[10]:

> Lewis has little to say about this aspect of the enterprise, suggesting merely that an agent's preferences might be knowable on the basis of 'raw behaviour'. (Hattiangadi and Stefánsson, 2021, p. 6479)

Specifically,

> … the Rationalisation Principle tells the radical interpreter to assign preferences to Karl on the basis of information about his physically-described choice behaviour in such a way as to satisfy the Bolker–Jeffrey axioms, and to assign a credence and a utility function to him that make him out to maximise expected utility. (Hattiangadi and Stefánsson, 2021, p. 6479)

[9] I've altered the following quotes from their originals, changing 'Karla' to 'Karl', 'her' to 'him', and so on, to maintain consistency with the rest of the paper.

[10] H&S cite 'Radical Interpretation' page 338 for this attribution, which appears to be a misreading. In that passage, Lewis says that instead of formulating Rationalisation as a constraint on attitudes in relation to behaviour described physically, we might instead formulate it as a constraint relating attitudes to behaviour under an intentional description. Lewis never suggested that an agent's preferences are derivable from the facts about their behaviour alone.

And so, finally, Lewis is said to argue:

> … if the radical interpreter knows that Karl's preferences satisfy the Bolker–Jeffrey axioms, [the interpreter] can rely on the Bolker–Jeffrey representation theorem to deduce Karl's degrees of belief and desire; if *what it is* to have those beliefs and desires *just is* to be representable as such by the lights of decision theory, the radical interpreter simply cannot fail. (Hattiangadi and Stefánsson, 2021, p. 6479)

In sum: the radical interpreter first determines Karl's preferences from his choice behaviour, and then uses them to determine his beliefs and desires by appeal to a representation theorem; and this is possible because beliefs and desires *just are* a means of representing preferences and the choices that 'reveal' them. Does that sound familiar?

$$\begin{tikzcd}
\text{choice dispositions} \arrow[d, "\textit{directly reveal}"] \\
\text{preferences} \arrow[d, "\textit{if axioms satisfied}"] \\
\text{beliefs \& desires}
\end{tikzcd}$$

What H&S describe isn't Lewis' theory; I've said enough in previous sections to establish that much. But I don't want to dwell on that. Connections between Rationalisation and representation theorems are frequently drawn, so H&S are certainly not alone in thinking that they're related. (See, e.g., Cozic and Hill, 2015, pp. 8–9; Buchak, 2016, pp. 800–801; Williams, 2016, p. 422; Elliott, 2017, pp. 388–389; Brouwer *et al.*, 2021, p. 3381 n. 22). So for this final section I want to focus on some general lessons regarding the relationship between the representational project and Lewis' theory – far too often one can find reasonable scepticism towards the former bleeding over into misplaced scepticism towards the latter. They are similar in many ways, but importantly dissimilar in others, and we should be careful not to conflate them. Each of the two main steps of the representational project – from choices to preferences, and then from preferences to beliefs and desires – sits very ill-at-ease with the Lewisean approach (read: analytic functionalism). I tackle those steps in reverse order.

3.6.1. From preferences to beliefs and desires

I should start by making a distinction between the constraints imposed on preferences by a *theory* and those imposed by a *theorem*. For example, there is a significant difference between what it takes for a system of preferences to conform to Jeffrey's distinctive variety of decision theory, and what it takes for a system of preferences to satisfy the axioms of Jeffrey's representation theorem. These are sometimes conflated – hence, for instance, one will occasionally see complaints that Jeffrey's decision theory requires preferences over an uncountably infinite algebra of propositions (e.g., Hattiangadi and Stefánsson, 2021, p. 6486), when in fact this is only a condition of his theorem. Jeffrey himself was clear about the distinction, noting that several of his theorem's axioms are considerably stronger than what's required by his theory (1990, pp. 147–149). (For preference axioms that are necessary and sufficient for consistency with Jeffrey's theory restricted to finite algebras, see Domotor, 1978).

This distinction is important, so it's worth being systematic about it. Suppose we arrange sets of axioms on preference into classes ordered by the implied relative strength of the conjunction of their members. Where 'representability' is taken to be relative to a specific decision theory – Jeffrey's theory, for instance – we can do it as in Figure 3.

Sets of type A_1 will always be strictly stronger than those of type B or C, in the following sense: the members of the A_1-set will jointly entail any and all axioms in any B-set or C-set, but not vice versa. In the same sense, A_2 is always strictly stronger than C, while B is always at least as strong as C and strictly stronger just in case the decision theory in question is consistent with systems of preference admitting of more than one representation. Finally, note that an agent's preferences will be consistent with a given decision theory just in case they satisfy all the axioms in an appropriate set of type C – in other words, exactly when those preferences are *representable* as arising from at least one coherent system of attitudes according to the relevant decision rule.

With that in mind, a question to ponder – what reason would an analytic functionalist have to suppose that Karl's preferences must satisfy all the axioms of Jeffrey's representation theorem? In short: none. Those axioms constitute a set of type A_2, or A_1 if supplemented in the right way. So even if we were to indulge the fantasy that we're to interpret Karl as an ideally rational *homo economicus* – the perfect evidential decision theorist – still we would require only that his preferences satisfy the axioms in a weaker set of type C. The same points apply in the case of Ramsey's and Savage's theorems, both of which posit axioms belonging to sets of type

A_1. Sufficient but not necessary for unique representability

A_2. Sufficient but not necessary for representability

B. Necessary and sufficient for unique representability

C. Necessary and sufficient for representability

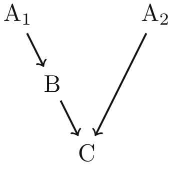

Figure 3 classes of axiom sets ordered by implied strength.

A_1.[11] There is no reason at all to imagine that Karl's preferences must satisfy any further constraints, such as those in a set of type B – or, *gasp*, A_1 – unless we want to insist on the thesis that his full suite of beliefs and desires can in all cases be fully determined from the facts about his preferences alone in the absence of any further considerations.

You won't find Lewis *anywhere* ascribing to that thesis, nor even hinting at it. And you shouldn't expect to either, unless there's a plausible case to be made that it's a clear part of folk psychology. But that would be very surprising indeed! The idea that preferences determine beliefs and desires has its origins with mid-Twentieth Century behaviourists overly-impressed by the strong uniqueness results that came attached to representation theorems like Ramsey's and Savage's, combined with the purported reduction of preferences to choice dispositions afforded by the recent advent of revealed preference theory (more on that shortly). If one is already inclined to think that mental states are really just fancy behavioural states, and one imagines they can reliably reduce preferences to behaviour, then one is strongly motivated to try to reduce beliefs and desires to preferences. It's a part of contemporary philosophical folklore that something like this might be possible, but it's not a part of folk psychology.

In any case, the real lesson of these theorems is the exact opposite: unique representability is typically achieved only under idealised and unrealistic conditions; take those away and we generally find that there are systems of preference that are multiply representable. A deeper appreciation of what the theorems are telling us is that the *homo economicus*' preferences do not, in general, determine her subjective probabilities and utilities. Sometimes they do, as the theorems show, but we probably shouldn't presuppose the kinds of idealisations required to make those theorems work. And the fact that Lewis frequently emphasised constraints relating to evidence and reasonableness and intelligibility is a clear indication that he was well aware of this. The idea that Lewis' theory requires Karl's preferences to satisfy all the axioms of some contemporary representation theorem is just false.

3.6.2. From choices to preferences

So it is not and never was a requirement of the Lewisean theory that Karl's preferences should determine his beliefs and desires due to satisfying the

[11] It's worth flagging that Savage's theorem will sometimes be presented as though his axioms belong to a set of type B. This way of presenting the theorem is possible if certain strong 'structural' assumptions are built into the definition of the expected utility rule itself. But it is better to separate such assumptions out of the decision rule, as Joyce (1999, p. 83) for example does with Savage's 'constant acts' assumption. The consequence of not doing so is a muddying of the useful distinction between the necessary conditions for *representability*, and the specific structural assumptions employed to establish *unique* representability.

axioms of some representation theorem. Still less did Lewis presume that we ought to be able to read Karl's preferences off of his 'raw behaviour'. The first step of the representational project is fundamentally at odds with Lewis' approach, given the role our attitudes are supposed to play in connection with evidence.

Let me say a bit more about how momentary choice dispositions are supposedly connected to preferences. The rough idea is that our preferences are revealed by our choices relative to different option sets. This can be precisified in a few ways (see Sen, 1971), but the most common version goes like this:

> Karl (actually) prefers *p* over *q* iff, if he were able to choose between making *p* true and making *q* true, with no other options available, then he would choose (or be disposed to choose) *p* rather than *q*

Call this the *revelation hypothesis*. It posits a direct connection between an agent's actual preferences and her option-counterfactual dispositions – that is, how she would be disposed to choose if her options were thus and so.

The first thing to note is that the revelation hypothesis is not a consequence of the expected utility rule. The expected utility rule relates an agent's choices in a given decision-situation to her preferences *in that situation*. It says that if an agent is presented with a space of options to choose from, she will choose the one or one of the ones that maximises her expected utility. It entails precisely nothing about the relationship between the agent's actual preferences and her option-counterfactual dispositions, because it imposes no restrictions on what her beliefs and desires (and hence her preferences) will be like in the relevant counterfactual scenarios. This means that the revelation hypothesis is consistent with the expected utility rule if and only if, for all *p* and *q*, if an agent's actual attitudes are such that she prefers *p* to *q*, then her attitudes in the specified counterfactual scenario will likewise be such as to determine a higher expected utility for *p* than for *q*. Once we recognise this, an obvious problem for the revelation hypothesis arises.

Consider a specific case. Karl is cycling down a two-lane path, and sees another cyclist coming towards him on the wrong lane. If they continue as-is, they'll collide. Karl has three options, which he strictly prefers in this order: he can CHANGE lanes, or CONTINUE and let the other go around him, or STOP. These are Karl's options in at least the minimal sense that for each, he can make it true and is confident he can make it true. Call this the *actual scenario*. In the actual scenario, Karl will choose CHANGE. What will he choose in the counterfactual scenario, where CONTINUE or STOP are his only options?

We assume again that Karl is an ideally rational *homo economicus*; we can presume also that Karl's basic desires are the same across the actual and counterfactual scenarios. There are then two possibilities: either Karl's beliefs are likewise the same, or they're not. If Karl's beliefs are the same in the actual and counterfactual scenarios, then CONTINUE will have higher expected utility as required. But if Karl's beliefs are the same, then whence the variance in behaviour? In the actual scenario, Karl chooses CHANGE, and does so *because of* his beliefs and desires. In the counterfactual, his attitudes are no different, and yet he chooses CONTINUE – are we to imagine that some exogenous change in what Karl can do magically influences his behaviour in the required way, *without* influencing the attitudes that normally cause such behaviour? No: if there's any difference in behaviour across the two scenarios, it cannot *just* be because one of his options has been stripped away without his knowing; there must also be some difference in Karl's beliefs. Perhaps, for example, Karl first tries to change lanes and finds he cannot, and so strikes that option off the list and reconsiders.

The only sensible way to make sense of the revelation hypothesis is to imagine not only that Karl's available options have been restricted in the counterfactual scenario, but also that Karl somehow *learns* of these restrictions and makes his decision on the basis of this information. So let whatever it is that Karl learns be specified by e. First: if what Karl learns e is *anything* other than the proposition CONTINUE $\vee$ STOP, then there's guaranteed to be at least one system of beliefs and desires $\langle P, U\rangle$ that's consistent with my description of Karl in the actual scenario, but where $\langle P(\cdot\,|e), U\rangle$ fixes a preference for STOP over CONTINUE. Second: what Karl learns isn't CONTINUE $\vee$ STOP – that is, 'I *will* continue or stop' – but instead something stronger: he *lacks the option* to change lanes. If he didn't learn this, after all, then he would try to choose his preferred option of CHANGE.[12] So, $e \neq$ CONTINUE $\vee$ STOP, and the revelation hypothesis is inconsistent with the expected utility rule on the assumption that Karl updates by conditionalisation. (For the simpler version of the same point: one might imagine that, if Karl were to learn to his amazement that he suddenly *cannot* change lanes, then he might suppose the other cyclist cannot either, and so would choose STOP – contrary to the revelation hypothesis.)

No doubt there's things fans of the revelation hypothesis could say in response to this kind of problem. Or perhaps they'll wave their hands a bit and say that the connection between preference and choice is 'tricky'. But do not mistake the point of the foregoing discussion. The goal is not to establish that there's no version of the revelation hypothesis that might work, and that advocates of revealed preference theory should be hanging their heads

[12] We can guarantee that Karl will try to choose CHANGE inasmuch as e doesn't rule it out, if we assume that every CHANGE-world has a utility higher than any $\neg$CHANGE-world; this can be built into the stipulation of the case without changing the point being made.

in shame. The point instead is that *if* there's any version of the revelation hypothesis that's going to make sense within the context of Lewis' analytic functionalism, then it had better be one that fits with what that theory says about the functional role of beliefs in relation to evidence and choice. To the extent that option-counterfactual dispositions matter on the Lewisean approach, then it's only inasmuch as they're a special instance of evidence-counterfactual dispositions – how the subject would choose if they were to *learn* that their options have been altered thus-and-so. That is what the argument above is teaching us. And the only way we're going to extract useful information about Karl's actual preferences from his evidence-counterfactual dispositions is if we pay attention to his evidence. Depending on what Karl learns in the various option-counterfactual scenarios he finds himself in, he may or may not choose in a manner that mirrors his actual preferences.

What sets the representational project most squarely at odds with Lewisean functionalism is that, according to the former, we are supposed to read Karl's preferences off of his 'raw behaviour', or his moment-to-moment choice dispositions. We're to imagine just that we know how Karl would choose given this or that hypothetical restriction to his option set, and somehow work backwards from there to the actual facts about his preferences. The Lewisean approach, on the other hand, considers not Karl's option-counterfactual dispositions, but his evidence-counterfactual dispositions; and it considers not just how Karl would choose under the assumption that he learns of this or that change to his options, but how he would choose (or would have chosen) given any range of sensory experiences he might have (or have had). With respect to the *kinds* of dispositions that matter, Lewis' approach and the representational project are very different beasts.

4. *Conclusion*

The foregoing myths often show up as a package deal. The unfortunately common interpretation of Lewis is that Karl's attitudes are determined by the best compromise between the independent recommendations of two principles of interpretation, perhaps alongside some further 'naturalness' factor. Charity tells us to maximise Karl's epistemic rationality, given the facts about his evidence. Rationalisation tells us to maximise Karl's pragmatic rationality, given the facts about his choices. Both are to be cashed out in Bayesian terms, with Rationalisation formulated specifically via appeal to some decision-theoretic representation theorem telling us when we can infer from option-counterfactual choices to preferences to beliefs and desires. Because certain kinds of 'source intentionality' facts serve as the inputs to those principles, they require an independent – and still missing – reduction to the physical facts.

That's not Lewis. It's Lew*ish*, maybe, but from the perspective of analytic functionalism it takes a few crucial missteps. No wonder, then, that it has been such an easy target for all these years.

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

REFERENCES

Bolker, E. (1967). 'A Simultaneous Axiomatization of Utility and Subjective Probability,' *Philosophy of Science* 34(4), pp. 333–340.

Brouwer, T., Ferrario, R. and Porello, D. (2021). 'Hybrid Collective Intentionality,' *Synthese* 199, pp. 3367–3403.

Buchak, L. (2016). 'Decision Theory,' in L. Buchak (ed.) *Alan Hájek and Christopher Hitchcock*. Oxford University Press, pp. 789–814.

Buchanon, R. and Dogramaci, S. (Forthcoming). 'Representation and Rationality,' *Philosophy and Phenomenological Research*.

Cozic, M. and Hill, B. (2015). 'Representation Theorems and the Semantics of Decision-Theoretic Concepts,' *Journal of Economic Methodology* 22(3), pp. 292–311.

Domotor, Z. (1978). 'Axiomatizaton of Jeffrey Utilities,' *Synthese* 39, pp. 165–210.

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

Eriksson, L. and Hájek, A. (2007). 'What Are Degrees of Belief?' *Studia Logica* 86, pp. 183–213.

Fodor, J. and Lepore, E. (1992). '*Holism: A Shopper's Guide*,' Blackwell: Cambridge, MA.

Hattiangadi, A. (2015). 'Metasemantics Out of Economics?,' in I. Hirose and A. Reisner (eds) *Weighing and Reasoning: Themes From the Philosophy of John Broome*. Oxford University Press: New York, pp. 52–60.

Hattiangadi, A. (2019). 'Radical Interpretation and the Aggregation Problem,' *Philosophy and Phenomenological Research* 101(2), pp. 283–303.

Hattiangadi, A. and Stefánsson, H.O. (2021). 'Radical Interpretation and Decision Theory,' *Synthese* 199, pp. 6473–6494.

Jeffrey, R. (1965). '*The Logic of Decision*,' 1. McGraw-Hill Book Company.

Jeffrey, R. (1990). '*The Logic of Decision*,' 2. University of Chicago Press.

Joyce, J. (1999). '*The Foundations of Causal Decision Theory*,' Cambridge University Press.

Lewis, D. (1966). 'An Argument for the Identity Theory,' *The Journal of Philosophy* 63(1), pp. 17–25.

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

Lewis, D. (1972). 'Psychophysical and Theoretical Identifications,' *Australasian Journal of Philosophy* 50(3), pp. 249–258.

Lewis, D. (1974). 'Radical Interpretation,' *Synthese* 27(3), pp. 331–344.

Lewis, D. (1975). 'Languages and Language,' *Minnesota Studies in the Philosophy of Science* 7, pp. 3–35.

Lewis, D. (1979). 'Attitudes de dicto and de se,' *The Philosophical Review* 88(4), pp. 513–543.

Lewis, D. (1980a). 'Mad Pain and Martian Pain,' *Philosophical Papers*. Vol. 1. Oxford University Press: New York, pp. 122–130.

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

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

Lewis, D. (1983a). 'New Work for a Theory of Universals,' *Australasian Journal of Philosophy* 61(4), pp. 343–377.

Lewis, D. (1983b). 'Postscripts to 'Radical Interpretation',' *Philosophical Papers*. Vol. 1. Oxford University Press: New York, pp. 119–121.

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

Lewis, D. (1988). 'Desire as Belief,' *Mind* 97(387), pp. 323–332.

Lewis, D. (1994). 'Reduction of Mind,' in D. Lewis (ed.) *Samuel Guttenplan*. Blackwell, pp. 412–431.

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

Lewis, D. (1997). 'Naming the Colours,' *Australasian Journal of Philosophy* 75, pp. 325–342.

Lewis, D. (2020). '*Philosophical Letters of David K. Lewis, Volume 2: Mind, Language, Epistemology*,' in H. Beebee and A. R. J. Fisher (eds) Oxford University Press: Oxford.

Pautz, A. (2013). 'Does Phenomenology Ground Mental Content?' in A. Pautz (ed.) *Uriah Kriegel*. Oxford University Press, pp. 194–234.

Ramsey, F. (1931). 'Truth and Probability,' in R. B. Braithwaite (ed.) *The Foundations of Mathematics and Other Logical Essays*. Routledge: London, pp. 156–198.

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

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

Schwarz, W. (2015). 'Analytic Functionalism,' in W. Schwarz (ed.) *Barry Loewer and Jonathan Schaffer*. John Wiley & Sons, pp. 504–518.

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

Smithies, D., Lennon, P. and Samuels, R. (2022). 'Delusions and Madmen,' *Synthese* 200(3), pp. 1–30.

Thompson, N. (2016). 'Is Naturalness Natural?,'*American Philosophical Quarterly* 53(4), pp. 381–395.

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

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

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