Representational scepticism: the bubble puzzle

J. Robert G. Williams [*]

September 30, 2016

Credo: we have the ability think about the world we live in, from our immediate environment of shoes, ships and sealing wax, to speculations about the furthest reaches of space and time. Representational sceptics deny the credo, maintaining that we cannot form thoughts genuinely about the external world. The form of representational scepticism investigated here is that our apparently external-world thoughts (”lo, there is sealing wax on yonder envelope”) turn out to have content that is strictly less-demanding than it at first appears. For example, our representational sceptic might hold that all that is required for the truth of x’s judgement that there is sealing wax on yonder envelope is that x’s *experiences* be of a certain characteristic (sealing-waxy/envelopey) kind. She denies what we non-sceptics maintain: that the truth of such judgements requires more than that our experiences be sealing-wax-suggestive—that they require a chunk of the external world to contain wax on paper. The sceptic doesn’t doubt there is an external world beyond our experience, but denies we have the ability to form judgements about it.

There are two ways to avoid representational scepticism. The realist maneuver is to hold the line, and insist we do have judgments (and other attitudes) with standard external-world contents—that are only true if there really is *sealing wax* in front of us. The idealist maneuver is to concede that if ‘external world’ content is construed in the way indicated, as ways the world can be, independent of our experiences, then such contents are unthinkable. The idealist invites us to drop that conception of an experience-independent external world, maintaining that what we are pleased to call ‘external’ objects are nothing but constructs of our experiences—or better, that contents about those objects are constructs of contents about experience. If one can argue that the realism leads to representational scepticism, and that representational scepticism is untenable, the idealist maneuver is the natural upshot. A (historically well-trodden) route to idealist metaphysics thus has representational scepticism as a staging post.

I present an argument from an interpretationist metaphysics of representation, for representational scepticism. Berkeley presupposed a certain resemblance-based metaphysics of representation, and used this to argue that (what we are pleased to call) external objects must really be ideas—since only an idea can resemble an idea. There are few now who accept Berkeley’s metaphysics of representation, but many contemporary realist metaphysicians, under the influence of David Lewis, accept a version of the metaphysics of representation from which I start.

[*] With thanks to many people with whom I’ve discussed this result, in particular to Thomas Brouwer, Rachael Briggs, Ed Elliott, Anandi Hattiangadi, Hanti Lin, Wolfgang Schwarz, Orri Steffanson, Brian Weatherson and audiences at Leeds, Syracuse, Stockholm, York, KCL. The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 312938.

The argument I develop—the bubble puzzle—is inspired by an argument-sketch given in (Lewis, 1984) against (a version of) the interpretationist metaphysics of representation he himself advocated. It is his "official" motivation in that paper for supplementing interpretationism with extra resources—a distinction between more or less eligible contents, built out of 'reference-magnetic' properties. Absent this, Lewis argues, we are subject to this particular kind of underdetermination. The introduction of eligibility as a 'saving constraint' in the metaphysics of representation is a paradigmatic realist maneuver, in my terms. I have plenty to say about how this is best developed, but the present papers focus will be on the argument for representational scepticism, not the best (realist or idealist) response.

Related underdetermination results (Quine, 1964; Wallace, 1977; Putnam, 1980, 1981; Kripke, 1982) have generated vast literatures, but in the case of the bubble puzzle to the best of my knowledge only the original sketchy remarks in (Lewis, 1984) and a recapitulation in (Weatherson, 2013) exist. The lack of attention to the bubble puzzle is unfortunate. First, it's not obvious that it works. This paper shows that Lewis's brief sketch can indeed be developed into an underdetermination argument, but only if a particular theory of practical rationality is presupposed. If a rival theory of practical rationality is substituted (the one endorsed by Lewis himself), then the original argument lapses. Second, the bubble puzzle has a special significance compared to other more familiar underdetermination arguments. Most significantly, the bubble puzzle directly targets the metaphysics of mental content, whereas the parade cases of underdetermination arguments most often work with linguistic content, and generalize to the case of mental content only by adding strong hypotheses about the relation between mind and language. I assert—and defend elsewhere—that this has had a distorting effect on the literature.

Section 1 introduces interpretationism as a metaphysics of mental content. Section 2 gives an informal sketch of the sort of underdetermination that threatens. Section 3 develops the formal results in the context of one theory of practical rationality. Section 4 brings the bubble puzzle together. Section 5 outlines the two main strategies of reacting to the puzzle. Sections 6 and 7 consider some objections based on the complaints about the presupposed formalism—the model of learning and the model of decision, respectively. In each case, I show conditions under which the complaints are valid, and ways of generalizing the argument to avoid them.

1 Interpetationism

Suppose you're given information about what choices Karl takes and what perceptual evidence he receives. The information you receive isn't just about choice and evidence in the here and now, but in close possibilities. In sum: you know Karl's dispositions to act in the light of evidence.[1]

So you've got a lot of richly described information about the inputs and outputs of Karl's psychology. Question: if you try to 'fill in' the rest of his psychology, the beliefs and desires he must have, in order to be disposed to produce *those* choices given *that* input, can you do it?

In order to assist, we give you one last constraint. Karl is *rational*. Let's be clear what kind of constraint this is—it means that the *internal patterning* of his psychology and its relation to the inputs and outputs is not messed up. He is not believing contradictory things; his beliefs and

[1] I'll be assuming that the information comes in intentional descriptions, so that we know that Karl is disposed to choose to p over q, r, s, where the $p \ldots s$ are propositions.

desires rationalize his choices; and he updates on input evidence in the right kind of way. Later we'll nail down a specific model of these rational constraints.

I've asked whether given sufficient information about inputs and outputs, plus the constraint that the psychology be rational, you could identify Karl's beliefs and desires. But it's not really about *you*. The question concerns the following property: *being a psychology which rationalizes dispositions to choose C under patterns of evidence E*. Does this property uniquely characterize Karl's actual belief-desire psychology, when C and E are filled with Karl's choices and evidence? In the words of Lewis (1974), we're not so much interested in how we determined the facts about Karl's beliefs and desires, but about how the facts (about choices, evidence and rationality) determine the facts (about what he believes and desires).

Questions of this kind are central to one of the leading accounts of the metaphysics of representation of the last fifty years—'interpretationism'. The account starts from the following biconditional:

> Karl has belief-desire psychology BD iff correct belief-desire interpretation of Karl ascribes to him psychology BD

Everyone can agree to this—the obvious thought is that it is right because it is necessary and sufficient for a belief-desire interpretation of Karl to be correct, that it ascribe him the beliefs and desires that he actually has. The obvious way of thinking of this, that is, just thinks of it as an instance of the general schema that if p is the case, then the correct answer to the question *whether p?* is p.

Interpretationists certainly needn't dispute this reading of the biconditional. But they think there is another way of reading it, on which the right hand side has metaphysical priority. What makes it the case that the left hand side obtains—that Karl has the belief-desire psychology he does—is that the right hand side obtains. This account will loop around in circles if it relied on an understanding 'correct' interpretation which pointed back to the left hand side. So interpretationists offer substantive accounts about what the correctness of an interpretation consists in. *Rationalizing the agents' dispositions to choose in the light of their evidence* is a candidate substantive analysis of 'correctness' of the kind the interpretationist strategy demand. Hence the interest among metaphysicians of representation in the question of whether this property *really does* characterize Karl's psychology uniquely—whether it's fit to play the explanatory role we've just identified.

One reason to concentrate on this sort of property is that we have available some beautiful proofs-of-principle of how this kind of thing can work. If you know what Karl is disposed to choose X out of the options X, Y, you know he prefers X to Y.[2] If you know Karl's rational preferences over all options, you can work out what degree of belief he must assign to each proposition, and what utility (desirability) he must assign to every outcome, in order for those beliefs and desires to have those preferences as their rational upshot. The 'representation theorems' that underpin (and add nuance) to these claims from the interpretationists' point of view are well labelled—they can play a key role in defending the presuppositions of the interpretationists' substantive analysis of 'correct interpretation'.[3]

[2] Though see (Maher, 1993) for evaluation of this transition.

[3] The results that I'm referring to need careful study. In the version due to Savage, there are limitations due to the combination of (i) the restriction of attitudes of belief, desire and action to three distinctive kinds of entities—states, outcomes and acts respectively (for discussion see Elliott, forthcoming); (ii) richness assumptions about the space of available acts (Joyce, 1999; Meachem & Weisberg, 2011); (iii) the status of the conditions on the

It is important to emphasize that the core interpretationist thesis here is a metaphysics of belief and desire. It is *not* a general scheme to reduce the intentional to the non-intentional. We can see this clearly if we look at what the explanatory base for Karl's beliefs and desires are, for my interpretationist—Karl's rationality, choices and evidence. Setting aside the status of the rational constraints one needs to appeal to, what Karl's *evidence is* in a particular situation, and what contrastive *choice* he has made, seem to me to be paradigmatically intentional facts. Relevant to the first are that Karl sees that there is a cube three metres to the left of him; that he hears that the trumpets are sounding, etc. Relevant to the second is that Karl *throws the cube at the trumpets* rather than *sitting on the cube and enduring the noise*. One could elaborate the core interpretationist scheme by adding in some further account of the intentionality of (perceptual?) evidence and of contrastive choice, but that's strictly beyond anything we're committed to here.

Interpretationism is not in the first instance a reductive metaphysics of intentionality, then (though it could be used in the service of such a project, if you are so minded). It's an account of how one form of intentionality grounds another—how the 'source intentionality' of perceptual and action grounds the 'secondary intentionality' of belief and desire. If you feel the pull of the inchoate thought that our ability to represent the world needs to be explained in terms of our immediate encounters with that world—whether as patient or agent, then interpretationism offers you a way to turn that urge into a systematic theory. The source intentionality itself could be theorized in all sorts of ways, compatibly with this. Source intentionality could be seen as metaphysically primitive (Pautz, 2013); 'grounded' somehow in phenomenology; given a naturalizing treatment, perhaps inspired by Dretske (1981); Millikan (1984); or somehow folded into a generalized interpretationism that takes bare behaviour (rather than contrastive choice) and sensory stimulation (rather than intentional evidence) as basic (which seems to be what (Lewis, 1974) had in mind).

With the pep talk done, let's turn to evaluation. Can *rationalizing choices in the light of evidence* really play the role that interpretationists want it to?

preference ordering as *rationally* required (extensively discussed, most famously by Allais, 1953; Ellsberg, 1961); (iv) that rational beliefs and desires take the form of probability and utility functions in the standard way. It is really the richness assumptions which are being targetted below (Zynda, 2000). In versions due to Jeffrey and Bolker, beliefs, desires and choice all attach to propositions; but the representation theorems are restricted in a precisely-specifiable way—they are unique only up to certain transformations. Joyce argues that the non-uniqueness is damaging—inequivalent mental states are represented by the belief-desire descriptions related by this transformation. In reaction, Joyce urges us to start not with a preference ordering potentially elicited from choice behaviour, but also a credence ordering. Credence and preference ordering together then determines belief and desire description much more precisely. Now, the spirit of the limitations to be discussed below is that the materials available to the radical interpreter are more restricted than is required to run these theorems. And so we get additional dimensions of non-uniqueness (or indeterminacy). The restriction to "act propositions" is discussed in detail below. In relation to Joyce's comparative belief primitive, one might think that from information about the evidence available to an agent, one learns not only total probability, but also can directly generate rational restrictions on the relative likelihood of various propositions—presumably propositions having something to do with the visible or audible scene one is perceiving. If we include the propositions involved in these comparative constraints in the set of 'evidence propositions' below, then it'll follow that (since the probabilities ultimately attached to evidence propositions are the same in any of the models of belief and desire that we discuss below) that the comparative constraints are met by every belief-desire pair. So helping ourselves to *certain* comparative belief constraints as well as (something close to) comparative preference constraints is fully consistent with the argument to follow. In each cases, the key thing for the purposes of this appear is the restricted range of propositions available to the radical interpreter in this setting. That won't worry those (like Joyce) whose interested is more in the relation between ordinal and cardinal characterizations of belief-desire psychology; but it matters to those interested in these results in their familiar role as getting a grip on the metaphysics of attitudes. Thanks to Rachael Briggs and Orri Steffanson for discussion here.

2 The bubble puzzle: first pass

Lewis is the paradigmatic interpretationist (his ambitions stretch beyond the core interpretationist project just characterized, but include it as a part). But he constructed an argument that *rationalizing choices in the light of experience* underdetermines Karl's psychology. Lewis's own argument is presented rather quickly, in a mix of the technical apparatus of decision theory and high-level intuitive glosses. Here we'll separate these out: so this section gives the intuitive gloss of the sort of challenge Lewis thinks is available; and the next section will exhibit the formal detail.

Karl sees a pile of banknotes in a box beyond his reach. They are mounted on one end of a seesaw. Karl realizes that if he hits the end near him the banknotes will fly up into the air and he can catch them. He can hit the seesaw or not (so by his lights there are just two courses of action open to him). He in fact hits it, a manifestation of a counterfactually robust disposition to hit it in evidentially and practically similar cases. This is a description of Karl's relevant evidence and disposition to choose (NB: *over*describing his evidence and choice is not a worry here, since we'll be looking to establish under-determination results. The reader should rather be on the alert for any *under*description).

First candidate psychological rationalization of Karl: he prefers an outcome where he lives in a world much like ours (a 'normal', non-counterinductive world) where he has riches, to a 'normal' world where he's in poverty. He believes himself to be in a normal world, and further to be such that if he hits the lever, he'll bring about a world where he has riches, whereas if he fails to hit it, he'll be stuck with the status quo: a normal world where he's in poverty. If he considers non-normal worlds at all, he assigns them vanishingly small confidence, and so any desires that he might have that discriminate between non-normal worlds won't be a rationally relevant factor in this choice. We can suppose that he's indifferent between these non-normal scenarios.

Second candidate psychological rationalization of Karl: he desires an outcome where he has riches and lives in a world very different from ours (a non-normal, counterinductive world, where things behave in the usual ways only within Karl's immediate 'bubble', the region of space-time within a few hundred miles of his location), to the same sort of non-normal world where he's in poverty. He believes himself to be in a non-normal world, and further to be such that if he hits the lever, he'll bring about the first scenario; if he fails to he'll bring about the second. If he considers normal worlds at all, he assigns them vanishingly small confidence, and so the desires that he might have that discriminate between normal worlds won't be a rationally relevant factor in this choice. We can suppose that he's indifferent between these normal scenarios.

Observation: either rationalization can explain the action Karl takes on this occasion. The postulate about desire over non-normal worlds can play the role of desire over normal worlds, so long as we also make a switch to ascribe a belief that Karl is in the non-normal worlds rather than the normal ones.

The strategy for coming up with the second candidate rationalization looks generalizable. One starts from the sensible rationalization in terms of beliefs and desires primarily over normal worlds. One finds a 'copy' of the relevant scenarios in worlds that are non-normal. The bubble-believing hypothesis switches around the attitudes to normal and non-normal worlds one finds in the the standard rationalization. Where we found discriminating desires among the normal worlds, we now find them among the non-normal bubble worlds. Where we found high belief that the world is normal (and fine gradations of degrees of belief among different normal scenarios) we now ascribe high belief that the world is non-normal (and fine gradations of degree of belief among the different ways the local bubble can be laid out).

Is there *any* disposition to action that Karl could have that couldn't be rationalized by a bubble psychology? The formal argument in the subsequent section will give reason to think that *everything* can be captured in this way (to anticipate: crucial to this will be identifying the right boundaries of the bubble. It has to large enough that normal and non-normal world pairs will be identical as regards the perceptual stream of basic evidence that Karl receives, and the basic actions that he takes in response. Different accounts of action and perception set these boundaries differently).

If this threat is realized, then two very different psychologies could rationalize the totality of Karl's dispositions to choose in the light of evidence. That causes trouble for the interpretationist metaphysics of representation. It proposed that *the facts* about Karl's psychology were grounded in *the correct belief-desire interpretation* of him, which was in turn to be given a substantive analysis in terms of *the belief-desire interpretation which rationalizes his choices in the light of evidence*. But the the uniqueness presupposition in the last step now seems to fail.

A failure of the uniqueness presupposition shouldn't by itself lead one to scrap the project. There are various ways to refine the proposal to allow for some non-uniqueness, and we have motivation to do this because it's quite plausible there there is some indeterminacy in the beliefs and desires we have—both on whether we are really *believing* some specific content, and on what the exact content of the beliefs we announce is. I'll assume that when P is the proposed analysis of 'correct interpretation', and there are multiple belief-desire psychologies satisfying P, then the interpretationist will be committed to saying that it is indeterminate (among this range) which is the correct interpretation; and thus indeterminate what beliefs and desires Karl has.

But accepting the indeterminacy in belief-desire content that bubble-rationalizations of action would give you would be astonishingly radical. It would be to say that there's no fact of the matter about whether you really are thinking about stars and the early history of the universe, as opposed to thinking about worlds in which the local bubble around you is *merely as if* there were distant stars and a particular early history of the universe. This is representational scepticism, pure and simple—a denial of our capacities to represent announced in the Credo.

For the metaphysician of representation, there are three reactions to this. The first is to embrace and advocate this as a feature, not a bug, of the story about representation. I am not tempted by this and neither was Lewis, but I think we can find figures in the history of philosophy who do build on this thought. The second reaction is that the interpretationist emphasis on rationalizing action and evidence was a wrong step, and some other entirely different starting point should be sought. The third reaction simply takes such results as evidence that not all constraints on correct interpretation have yet been found—underdetermination results like these then have a positive role in showing that there are further 'saving' constraints on correct interpretation, beyond rationalization, that we must seek to identify. I will discuss these later.

3 The bubble argument: second pass

3.1 Preview of the formal results

The informal sketch of the argument leaves quite a lot of details suppressed. Exactly how should we represent evidence and choice? What are the rational patternings of belief/desire/update/action that constrain interpretation? These details matter, if we're going to explore how robust the argument is under varying assumptions.

The three components that Lewis worked with include (i) a Bayesian account of rational partial belief, with probabilistic constraints ruling out interpretations on which e.g. inconsistent claims are both believed; (ii) a Bayesian account of update, on which a learning experience is associated with a proposition *p*—the total incremental information learned with certainty in that experience—to which the rational response is to update one's prior beliefs by conditionalizing on *p*; (iii) a decision-theoretic account of the rational patterning between belief (credence), desire (utility) and contrastive choice. Out of a set of options, the one to choose is that which has highest utility, where utility assignments to options for action (hit the seesaw!) are constrained by utility assignments for outcomes (obtain lots of money) and conditional credences (high in: obtaining lots of money conditionally on hitting the seesaw).

Lewis's own version of the argument, in a nutshell, is to consider a total rational credence and utility assignment over a space of worlds. Among this vast space will be a normal world, and its counterinductive, non-normal bubble twin. He claimed that if you permute the credence and utility assigned to such a pair you'll get a deviant psychology that rationalizes all the same actions in the light of the same evidence. We'll see however that the general result turns crucially on aspects of the model of rationality in play that are disputable (and, ad hominem, Lewis himself denied).

3.2 Definitions

For simplicity, the following results are established against a background with finitely many possible worlds.

An **evidence proposition** (for *x*) is a proposition which describes a possible piece of *evidence* for *x*.

Comment: Evidence here should be construed restrictively, as *total immediate evidence*. Pre-theoretically, my evidence might include that I'm sitting in my house. But perhaps this is only evidence for me now *due to* a combination of my background beliefs (the study in my house looks like *this*) and the immediate evidence of my senses (I'm seeing *this*). On the Bayesian conception we are assuming, mediate evidence is the result of conditionalizing priors that include the background belief, on the immediate evidence. There's of course lots of room for arguing about what's included in immediate evidence (does it include just the facing surfaces of objects perceived, information concerning the whole object, or even a distribution of sense-data?). A strength of the argument is that it doesn't require us to take a stance on this at the beginning—instead, different stances on this point will adjust how we read the eventual conclusion.

Let $\varepsilon = \{E_1, E_2, \ldots\}$ be a partition of possible worlds, such that for every evidence proposition E, its intersection with any E_i is either null or E_i itself. Thus any such E is the direct union of cells of ε.

An **action proposition** (for x) is a proposition which describes an *option for action* that x may choose in a possible choice situation.

Comment: Possible actions here should be construed restrictively, as *total immediate actions*. Pre-theoretically, when I walk to the kitchen and cook a big bowl of pasta, that could be counted as the action of obtaining nourishment for tomorrow's long run. This isn't at all odd as an intentional description of what I do—it does indeed figure in my long term plans. But perhaps the right way to describe this is that I choose to *cook* the pasta, because I desired to obtain nourishment for the long term run, and had the background belief that if I cook the pasta I'll obtain the nourishment. Just as with evidence, there's lots of room for arguing about what's included in immediate actions. (Does it include very local and short term acts, or should the information include long-term plans? Or internal strivings whose known causal upshot are the bodily movements that implement the plans?) Again, a strength of the argument is that it doesn't require a stance on this, and different views will adjust how we read the conclusion.[4]

Let $\alpha = \{A_1, A_2, \ldots\}$ be a partition of possible worlds, such that for every evidence proposition A, its intersection with any A_i is either null or A_i itself. Thus any such A is the direct union of cells of α.

Let γ be a partition of possible worlds which is a common refinement of both α and ε. So every cell in either of the latter two partitions is the direct union of cells of γ. It follows that every action proposition and every evidence proposition is the direct union of cells of γ; and every *conjunction* of some evidence proposition with some action proposition is likewise a direct union of cells of γ.

3.3 Probabilities

Now consider a probability distribution over the algebra Γ generated by the cells of γ. Note that every action proposition and every evidence proposition is an element of this algebra. Furthermore, if X is an action or evidence proposition or conjunction thereof, and it is a direct union of cells $c_1, \ldots, c_n$, then the probability of X is the sum of the probabilities of these c_i.

Consider a probability distribution over the larger algebra A generated by the possible worlds. By the above, any two probability distributions which *agree* on their assignments to the cells of γ, will agree on their assignments to any evidence or action proposition or conjunction thereof.

Corollary: they will also agree on the *conditional probability* assignments $p(E|A)$, where A is an action proposition and E is a evidence proposition (for the conditional probability, where defined, is a ratio of the probability of $A \wedge E$ to the probability of A, on which ex hypothesi the probabilities agree.[5]

[4] One way of attacking the argument is by making the case that any proposition could in principle be the content of an immediate act. Perhaps God could offer you a choice between P and $\neg P$, for almost any P. Or perhaps (if one individuates options for actions in terms of what the agent takes themselves to be able to do) possible choice situations in which one was under the delusion that one was God could lead to 'choices' to bring about P. But such possible choice situations only extend the range of act-propositions if these are the *immediate* content of the relevant choices. And the most natural filling out of the scenarios are ones where God will bring about P if you indicate your choice to Him in some way—so this may be an act of bringing about P, but only via the more basic act of indicating-to-God. Likewise, I struggle to imagine what it would be to believe one could immediately act to bring about any proposition—as to believing one had magical powers so that one could bring this about by wishing it were so or wiggling one's fingers. But the wishing or wiggling would then feature as act propositions, not P itself. Thanks to Ed Elliott for pressure on this front!

[5] A slight generalization that will be useful later. If we have partitions γ and γ' of two spaces k and k'. We

3.4 Jeffrey values

Add to the probability p over the algebra Γ a Jeffrey-style *expected value* assignment (Jeffrey, 1965). The value assignment to propositions v, for Δ an arbitrary partition, must satisfy:

$$v(X) = \sum_{Y \in \Delta} p(Y|X)v(X \wedge Y)$$

We can take $\Delta = \gamma$, in which case no matter what X is, either $X \cap c_i = c_i$ or $X \cap c_i = \emptyset$ for any cell of γ. Thus if $X = c_1 \ldots c_n$, the value assigned to X is just a weighted average of the values assigned to the c_i, with weights given by $p(c_i|X)$.

First observation: determination of probability and value by assignments to cells.

Suppose we have two Jeffrey-style probability-value assignments over the whole of A, which agree on the probability and the values assigned to cells of γ. We earlier saw that the probabilities assigned by the two to evidence and action propositions, conjunctions thereof and conditional probabilities between evidence and action propositions. Now consider the value assigned to any action proposition. This will again be a weighted average of the values assigned to cells of γ by the respective assignments (and ex hypothesi the values of the cells coincide). But we've just argued that the conditional probabilities which gives the *weights* for the weighted average also coincide. So the value assigned to every action proposition is the same, on any two probability-value assignments so long as they agree on their assignments across γ.[6]

Second observation: preservation under updating by evidence.

Now suppose we have two probability-value assignments p_1, v_1 and p_2, v_2 over A. Suppose evidence has arrived, and so the probabilities are 'updated' by total evidence E, to form $p'_i : X \mapsto p_i(X|E)$. Correspondingly, v'_i is the value assignment determined by $v'_i(X) = \sum_w p'(w|X)v_i(w)$, where w ranges over possible worlds. The former corresponds to update of degrees of belief by conditionalization on total evidence; the latter corresponds to update of desirability by holding fixed the desirability of complete possible worlds, given Jeffrey's constraints on desirability.

The pair p'_1, v'_1 and p_2, v'_2 will agree on their assignments across γ if the original pair did.

To begin with, recall that the evidence E on which we're updating must be the direct union of cells drawn from γ, say $E = c_1 \dot{\cup} \ldots \dot{\cup} c_n$. The updated probability of $p'_i(c)$ will therefore be either zero (when c is not among the c_j) or given by $p_i(c)/\sum_{0 \leq j \leq n} p_i(c_j)$. Thus, since p_1 and p_2 agree on the c_i and so the RHS of this equation, p'_1 and p'_2 on the LHS of the equation agree also.

We need also to show that the values assigned to cells of γ by v'_1 and v'_2 will coincide under these assumptions.

The key fact here is that for any world w and cell c of γ that is contained within the evidence E

consider the algebras Γ and Γ' generated by γ and γ' respectively. Suppose we have probabilities p and p' over the larger algebras Ω and Ω' of k and k' respectively, and a bijection from cells of the γ into γ', which preserves probabilities. By the reasoning above, the induced bijection between Γ and Γ' preserves probabilities of direct unions thereof and conditional probabilities between such direct unions. The result in this paragraph is the special case of this where $k = k'$, $\gamma = \gamma'$ and the bijection is the identity. The more general version is useful later.

[6] To continue the generalization from an earlier footnote, we suppose that the bijection between γ and γ' preserves value as well as probability. By the reasoning here, the expected value of direct unions of such cells is preserved by the induced bijection between Γ and Γ'.

on which we're updating, $p_i(w|c) = p'_i(w|c)$. This is simply a special case of the general fact that conditionalizing a probability on Z leaves invariant any conditional probabilities $p(X|Y)$ where Y entails Z.

So now take any cell c compatible with the evidence E upon which we have updated.

$v'_i(c) = \sum_{w\in c} p'_i(w|c)v'_i(w) = \sum_{w\in c} p_i(w|c)v_i(w) = v_i(c)$.

The second identity holds by the above key fact and the assumption that value-assignments are invariant on worlds under updating by evidence. The result shows that not only are value-assignments invariant in their assignments to worlds, but also on their assignments to cells of γ.

But since v_1 and v_2 agree on their assignments to these cells ex hypothesi, v'_1 and v'_2 will continue to agree.

Note that it's crucial to this argument that γ partitions any evidence proposition—the result is not valid for arbitrary 'updates', but only for updates by the propositions we picked out in fixing what ε and hence γ is to be.[7]

3.5 Applications

An ideally rational agent will be disposed to perform a_1 rather than $a_2, \ldots, a_n$ if $v(a_1)$ exceeds $v(a_i)$ for $i \neq 1$—where v is the value function that the agent would rationally hold in the situation faced with the choice.

Can the totality of the agent's choice dispositions determine what beliefs and desires the agent has? The argument above gives a limitative result. For suppose the agent's actual beliefs are the result of updating ur-prior P on evidence stream $e_1, \ldots e_n$ and her desires are the result of combining these in the Jeffrey-way with an invariant desirability assignment v to worlds. The beliefs and desires that she would rationally hold in alternative choice situations are then the result of updating P by some stream of evidence.

Monkey around with the ur-prior P and the value-assignment v to worlds howsoever you wish, under the sole constraint that the probability and value assigned to cells of γ be held fixed. Then the observations above tell us that the original and the twisted psychologies correspond to exactly the same dispositions to choose in the same circumstances.

How can you generate a twisted psychology? Here's one special case (the one that Lewis and Weatherson focus upon). Suppose w, w' are both in a cell c (they are "indiscriminable" by any evidence or action). Then we simply permute the probabilities and values assigned to w and to w', and let the new overall probability and value assignment to arbitrary propositions be determined by the new probability/values assigned to worlds.

There's plenty of other ways to twist a psychology other than the Lewis permutation, however. You could concentrate *all* probability devoted to the cell on a single 'representative' world, and let that world's value match the value of the cell as a whole. You could divide the probability of the cell equally over all worlds in the cell, and assign each a utility equal to the total value of the cell. You could let probability and utility go *undefined* at any finer grain than the cells

[7] Applied to our earlier generalized setting, the claim will be that updating p by some element of Γ, and updating p' by the corresponding element (under the induced bijection) of Γ' will result in the updated probability and value assignments to cells being preserved under the original bijection between γ and γ'—a result that by earlier arguments means that the induced bijection preserves probability and value over the whole of Γ and Γ'. The result replicates the reasoning above.

themselves. All these assignments, since they agree on probability and value at the cell-level, will be indistinguishable in terms of dispositions to choose in any evidential situation.

4 The bubble argument: synthesis

The first pass at the bubble argument for representational scepticism conveyed a sense of what deviant interpretations of a single action might be like, and gestured at how this could be generalized. The second pass works in a formal model of choice, evidence and their rational relations to belief and desire. In effect it shows that rationalization alone cannot fix attitudes to possibilities that are not already carved by observation or action. (In a sense this is the limitative counterpart of the representation theorems that serve as the proof-of-principle of the interpretationist strategy. They showed that if you assume that choices across a range of options are sufficiently rich, the degree of belief and desirability can be pinned down. To get that result, you need to have information about the choices an agent is disposed to make between *arbitrary* propositions. But the sphere of action of ordinary agents like us is limited—we cannot choose whether or not a fly on Alpha Centauri keels over or not. Our choices are mediated by interactions with our immediate environment. And what the formalism above does is pin down the consequences for rationalizing interpretations from limited spheres of agency and evidence.

The limits imposed by the formal result depend on what one takes the scope of evidence and action propositions to be. This is where we go back to the non-normal bubble worlds of our first pass. Let your bubble (at t,w) be a part of the w that is sufficient to fix what evidence you are impacted by at that t, and what actions you take then. Sitting in a windowless room, the bubble's boundaries might coincide with those of room, for example (or perhaps, if you adopt the view that part of your visual evidence is the existence of the *house* containing the room, then the bubble should contain the house). Likewise, the action you in fact take within the room are confined to the insides of the room, intuitively. The bubble of x at w is the fusion of x's bubbles at w for varying t.

I submit that worlds w and u which have duplicate bubbles for x will not be distinguished by evidence-propositions or action-propositions. The content of your perceptual evidence is as true of any bubble-twin of the actual world as it is of the bubble world itself. The choices you make in the actual world are taken in just the same sense in any world that duplicates the bubble. Because of this, each world that duplicates x's bubble will be contained within the same 'cell' of x's γ. In particular, we can take our actual, presumably uniform world, and pair it with a counterinductive 'bubble-and-void' world which duplicates my bubble, but where there is literally nothing—just void—outside the bubble. For any world w that x takes to be doxastically possible, one can construct a bubble-and-void twin, and such pairs will always share a cell.

We can then use this pairing of worlds with bubble-and-void twins as the basis for any of the monkeying around with psychologies considered earlier. Personally, I put a lot of confidence in the hypothesis that the world is uniform rather than bubble-and-void-ish. But within each cell we could switch the confidence and values around to produce the same cell-by-cell assignments (this is what Weatherson and Lewis propose). We could project *all* the confidence of a given cell onto the bubble-and-void possibility, leaving none at all for uniform possibilities. We could distribute credence uniformly. We could leave all such questions undefined. All give the same results.

The exact way we draw can boundaries of the bubble can depend on one’s theory of immediate action and evidence. More externalist treatments of immediate evidence and action will force one to extend the bubble quite far to get the argument to run. Imagine throwing stones from a mountaintop while surveying the lands, or gazing at the stars—if the houses smashed by the avalanches you cause and the stars that fill your vision are part of your immediate actions and perceptual evidence, then the bubble covers vast distances (though it still has limits). More internalist treatments of immediate evidence and action (with the limiting case where immediate content concerns the distribution of sense-data and internal strivings) will allow us to wrap the bubble very closely around the agent . Everyone has a bubble puzzle, but in their own boutique version.

5 Representational scepticism: idealist and realist maneuvers

With the general scheme of bubble puzzles now established, let’s consider two opposing reactions, which each in their own way defuse the threat of representational scepticism.

Berkeley is the talismanic representational sceptic in the history of philosophy. We can construct a Berkeleyan challenge in this framework. First, adopt a metaphysics of perception and action on which both are individuated ‘internalistically’. Evidence propositions are in the first instance something like arrays of sense-data; actions are something like strivings or internal movements of the will. Both perhaps nomically- but not metaphysically-necessarily connected to the external environment. Your ‘bubble’ in this setting is wrapped tightly around: it shrinks to just you, your sense-data and strivings.

Now consider the rationalizing psychology floated above on which assignments below the cell level are *undefined*. Then the finest ‘resolution’ that your beliefs and desires distinguish between are patterns of sense data and strivings. This would be a psychology on which we do not represent differences between possibilities where one’s ideas are the upshot of external objects and ones where such objects are absent. Our result does not quite reach the Berkeleyan conclusion that no representation of external objects is possible. The conclusion that there cannot be a fact of the matter that we represent external objects is, however, within a hair’s breadth.

So what is the Berkeleyan reaction? The case illustrates is how one might defeat a representational sceptical argument without revising one’s metaphysics of representation. Berkeley recommends we adopt an idealist metaphysics of the world, on which there are no non-mental objects to be represented. In terms of the argument above, you can think of the Berkleyian denying that the space of genuine possibilities cuts any finer than γ itself—the patterns of possible sensory experience and strivings. If we stick with that possibility-space, and with the idealist regards ordinary things as constructs from mental items, the argument for *underdetermination* of representation goes away. On this way of looking at things, it is the realist posit of facts about the world that cut finer than the array of ideas and willings that would generate representational scepticism. (One can consider numerous parallels to this strategy—for example, one aspect of Benacerraf’s puzzle for abstract acausal mathematical objects is that, even where such things to exist, we could not think about them in the absence of causal interactions. The idealist reaction to representational scepticism just mentioned is the analogue of the nominalist reaction to Benacerraf's puzzle).[8]

The other way to eliminate representational scepticism is to adjust the metaphysics of representation. Interpretationists were searching for an account of 'correct representation', and what we've seen is that *rationalizing choices in the light of evidence* looks like it underdetermines belief-desire psychology. A natural thought is that there are further constraints. This was of course Lewis's reaction (Lewis, 1983, 1984). Some representational contents, said Lewis, are 'more eligible' than others, more apt to be the contents of attitudes. This has something to do with the ranking of properties as more or less natural (closer in definitional distance to the properties featuring in microphysics). The idea is most often developed in the context of linguistic content, where we assign properties as semantic values of predicates, and can rank candidate interpretations by the naturalness of the values assigned. It is much harder to see how to generalize this to mental content, if that is not assumed to be language-like (though see ANON for my discussion of the case) But the general theme is clear: further elaboration of the constraints under which source intentionality of choice and evidence determines the content of the mediating states of belief and desire.

Whether you jump the Berkeleyian way or the Lewisian way, you have hard work to do in spelling out the details of the revised metaphysics—an idealist metaphysics of the world around us in Berkeley's case, or an elaborated metaphysics of content in Lewis's. Evaluating the responses here is a matter for another day (to put my cards on the table, I'm sympathetic to a version of the Lewis strategy, but I think it's much harder to avoid falling back into representational scepticism than has been acknowledged). In the remainder of this paper, focused as it is on the core representational sceptical argument that generates the puzzle, I want to examine whether there are loopholes in the argument.

6 Extension to Jeffrey conditionalization

The Bayesian model that underpinned the earlier results takes for granted a model of learning from experience where some proposition (the total immediate evidence proposition) is learned with certainty. That is certainly a familiar model, but seems a very strong assumption to make. Indeed, presumably most of what we ordinarily take to be the content of experience we do not learn with certainty (e.g. I am not certain that I see a dagger before me on the basis of my visual experience, though I come to be extremely highly confident of it). So the model seems to have built-in pressure towards the more radical models on which it is some Cartesian theatre of sense-data which characterise the evidence we most immediately have.

Now, Bayesians haven't wanted to be saddled with conditionalization as the only model of update. So they have searched for a more flexible account of updating, to offer to those who

[8] Compare this strategy to another kind of representational scepticism, associated with Putnam. Putnam notoriously argues that sceptical possibilities are self-refuting. If one were a brain in a vat, one's thoughts and talk would be about Vat-objects, rather than putative real-world objects (likewise, it is not the case that you are having the illusion of acting upon real objects; rather, you are really acting on vat-objects). Rather than a causal metasemantics, I offer an interpretationist reconstruction. The hypothesis will be that the evidence and actions of the original partition do not discriminate between (a) you being out of the vat and interacting with a real cherry tree; and (b) you being in the vat and interacting with a vat cherry tree. If the space of genuine possibilities does distinguish between worlds where you are in or out of the vat, then the conclusion of the above arguments will be that all the options and twists of psychology are available—there is no fact of the matter whether you are representing yourself as envatted or out of the vat, or simply having attitudes that are undefined below a level that quotients out the envatment issue (your attitudes would have truth-conditional content roughly like 'that vat-or-non-vat cherry tree is flowering').

think that the evidential impact of a learning episode might be to fix high-but-not-complete confidence in a certain proposition (e.g. that there is a dagger before me). Jeffrey conditionalization is one such strategy. I'll explore the hypothesis that the right way to respond to perceptual episodes is to update by means of Jeffrey conditionalization—and we'll see what becomes of our argument for representational scepticism.

Jeffrey conditionalization is a way of updating probabilities given as input, not a proposition on which you become certain, but a partition whose elements have fixed (new) probabilities. Thus it could be A and $\neg A$, with the constraint that $p(A) = 0.9$ and $p(\neg A) = 0.1$. Note that the kind of evidence on which updating by conditionalization is designed to operate can be viewed as the special case where the constraint is that $p(A) = 1$ and $p(\neg A) = 0$. But more generally we have a partition Θ and a constraint c that maps elements of Θ to real numbers totaling 1.

The Jeffrey-update of a prior P by Θ, c is then given by $p(X) = \sum_{\theta \in \Theta} c(\theta) P(X|\theta)$, i.e. a weighted average of what you'd get by learning each θ with certainty, with weights given by the constraints.

To adapt the argument for representational scepticism, we need to revisit the characterization of ε. This was a partition chosen so that every evidence proposition was the direct union of its cells. The natural generalization is to take all the cells of any partition involved in a possible learning experience as 'evidence propositions'. We then run through the reasoning as before. The first observation in 3.4 goes through just as before. The reasoning under the second observation shows that two psychologies agreeing on probability and utility over cells of γ will continue to agree after conditionalizing on any evidence proposition. But a learning episode now may not take that form. Still, since a Jeffrey update is simply a mixture at fixed weights of conditionalizations on evidence propositions (in our new sense), and we know the latter are the same for our two prior probabilities, we'll get that the updated probability assigned to a given cell of γ will be the same for each.

The argument for the invariance of value-assignments under updating can be run exactly as before, with the same assumptions.

So the extension to Jeffrey conditionalization goes through. I do think it affects the significance of the bubble argument, but in a way already prefigured in our discussion. We emphasized earlier that there were substantive philosophical issues in identifying the evidence propositions and action propositions that form the basis of the interpretation. The point that these may be argued to be externalistically individuated—so that the immediate evidence of my senses gives me that *that is a dagger in front of me* rather than *there is a cluster of thus-and-such colours and shapes in front of me*—is something we've already acknowledged. The impact of such models on the earlier considerations is to stretch out the boundaries of the bubble with respect to which the representational sceptical argument is posed, to encompass any objects that feature in evidential content. Nevertheless, if we were stuck with conditionalization as our model of update, the combination of these more externalist views with the model of rational update would be implausible. What the generalization to Jeffrey updates does, in the first instance, is give us a way of representing the more externalist content and showing that indeed, the argument can be run as before, with the only issue being to choose the bubble carefully.

There are other loopholes that Jeffrey updates allow for, but it's hard to see how they could be exploited. For example, any cells involved in a Jeffrey update will count as evidence propositions in the relevant sense. Perhaps the direct input from experience, for a person who gets testimony from friends that she is actually in a counterinductive world, should be to set credence in that scenario to 0.0001, while letting her credence in it being a normal world to be the difference between that and 1. But that would make *being in a counterinductive world* an evidence proposition. And so your evidence propositions and any partition refining these would separate the inductive from counterinductive worlds, and so we wouldn't be able to construct a 'bubble twin' of our actual experience that lets things go counterinductive (e.g. voidlike) outside the bubble.

The trouble with evaluating all this is that the characterization of the partition involved in a Jeffrey update and the numbers assigned to cells is notoriously under-theorized. So one who wants to resist representational scepticism in this way has an uphill struggle defending some particular theory of rational updates that gives a role to the kind of propositions that would cause trouble for the argument. On the other hand, there is at least one attractive story about Jeffrey updates, due to Wolfgang Schwarz, that is very friendly to the representational sceptic. According to Schwarz (ms.), every Jeffrey update is in fact produced by an underlying conditionalization—a conditionalization you can think of as a matter of learning a sense-data distribution with certainty. The twist is that Schwarz urges us to think of these sense-data propositions as 'virtual' propositions that don't describe genuine ways for the world to be. In essence, they are introduced just to code up various patterns of Jeffrey-style updates over genuine propositions. The visual system works as if it received information about sense data (which explains why we may be under the illusion that there is such a thing as the 'sense data impacting me right now'), but the real action is in the way genuine propositions are updated, and typically there is no genuine proposition updating on which would get us that effect—restricting attention to genuine propositions we get what looks like Jeffrey update.

The issues here are complex, but I wanted to emphasize that if Schwarz's intriguing picture is the correct way to handle these issues, then we can run the original arguments with A being the fine-grained space of 'virtual worlds' which are specified to include sense-data distributions. Relative to that space, the evidence propositions are the virtual sense-data propositions, and the bubble that we construct is correspondingly local. We can then run the original argument, to produce bubble-and-void twins of our ordinary psychology. And these will rationalize behaviour in the light of experience. Now, although we've constructed this twisted psychology by making appeal to virtual propositions, the twists in the psychology do have impacts for genuine propositions—-shifting credence across to worlds with lots of void some distance from me, for example. And so we get the result that when we restrict attention back to credences and utilities across genuine propositions, we'll have a psychology that's both (a) crazy and (b) rationalizes all action in the light of experience.

In sum, the formal argument does generalize to Jeffrey updates. It requires a new understanding of evidence propositions in order to do so. And to assess the impact of that for our argument, we'll need to have detail on what sorts of Jeffrey updates experience prompts (compare: to run the original argument we need to have detail on what sorts of propositions experience requires us to conditionalize upon). Finally, at least one worked-out story of that will take us back to the most extreme kind of representational scepticism, even though the cells of the Jeffrey update can be as externalist and world-involving as you like.

7 Coda: Causal Decision Theory

I think Jeffrey conditionalization, if anything, makes the argument for representational scepticism more robust. But there is a more difficult obstacle that might really undercut the formal underdetermination argument. The model of rational constraints on degrees of belief and utility involved in the formal arguments above was a version of Jeffrey's Evidential Decision Theory (EDT). But EDT suffers from counterexamples—it recommends one-boxing in Newcomb puzzles where many hold one should two-box. It recommends choosing to receive good news over promoting good results (or so its detractors maintain, and I'll go along with that for now).

The rival, at the same level of generality, is causal decision theory or CDT. The characterization of (causal) expected utility u, in the version of causal decision theory I will consider, takes the following form:

$$u(X) = \sum_{k \in \kappa} p(k) \sum_{Y \in \Delta} p(Y|X \wedge k) v(X \wedge Y \wedge k)$$

The key change is that a special role is given to a causal background partition κ. Relative to each cell k of this partition, expected value is calculated as in Jeffrey's formalism, but a weighted average of these conditional expected values is taken, with the weights being the subject's credence that they are in the relevant cell. (I will assume for simplicity that cells of κ have a non-empty intersection with cells of γ).

Consider the following four worlds:

1. Bubble-and-hellfire, announce progressive taxation. 0.01 credence. -100 utility.
2. Bubble-and-hellfire, fail to announce progressive taxation. 0.01 credence. -100 utility.
3. Uniform world, announce progressive taxation. 0.48 credence. 10 utility.
4. Uniform world, fail to announce progressive taxation. 0.48 credence. 0 utility.

We further assume that worlds 1 and 3 are in the same cell of the partition that refines the evidence and action partitions (the causal background is: a local bubble of normality, with hellfire outside), and likewise worlds 2 and 4 (the causal background is mundane, uniform inside and outside the agent's bubble). Our earlier results for Jeffrey's decision theory assured us that the expected utility of actions would be invariant, as long as we kept the expected utility of the cells of such a partition invariant. One way to do that would be to permute the probability and utility assigned to worlds 1 and 3 respectively, keeping the rest fixed.

We can show that the result for Jeffrey's decision theory doesn't generalize to causal decision theory by showing that this specific permutation instance of it won't leave causal expected utility invariant. We are assuming that κ-partition has a cell containing worlds 1 and 2, and a cell containing worlds 3 and 4 (intuitively, a bubble world is one causal backdrop, and a uniform world is another). On the original unpermuted assignment above, the causal expected utility of announcing progressive taxation will then be the probability attached to the bubble cell, multiplied by the utility of the unique bubble-world at which the act is performed, plus the probability of the uniform cell, multiplied by the utility of the unique uniform world at which the act is performed, i.e. $0.02 \times -100 + 0.98 \times 10 = 7.8$. The permuted assignment on the other hand would be:

1. Bubble-and-hellfire, announce progressive taxation. 0.48 credence. 10 utility.
2. Bubble-and-hellfire, fail to announce progressive taxation. 0.01 credence. -100 utility.

3. Uniform world, announce progressive taxation. 0.01 credence. -100 utility.
4. Uniform world, fail to announce progressive taxation. 0.48 credence. 0 utility.

The impact of the permutation is to equalize the probability of the bubble cell and the uniform cell. Although the utilities of the two outcomes being averaged is the same (just appearing in different spots) this means that the weights attached to the averaging are altogether different. The same recipe as before gives us: $0.5\times 10+0.5\times -100=-45$. So as promised, this illustrates that in the causal expected utility framework we cannot monkey around within the cells of the action/evidence partition as we could in the earlier setting.

It is straightforward to check that there is a generalized result available. If one holds fixed the expected utility attached to the cells of a common refinement of action, evidence *and causal background partition* then the line of reasoning from earlier goes through. So within *those* cells we have our result. But this generalization doesn't allow our original argument to be reconstructed. Uniform and Bubble worlds now live in different cells of the relevant partition, and so one cannot revive in this fashion the original underdetermination result whereby an ordinary believer in an uniform external world could be systematically reinterpreted as a bubble-believer.

There is a way to get the argument going again (I thank ANONYMIZED for this idea). As before, let α be our action-partition, ε be our evidence partition, and γ be a common refinement. κ is the causal background partition. Now take an arbitrary $g\in\gamma$, and let $g_i=g\cap k_i$. Note the collection of all g_i for $g\in\gamma$, discarding any null sets, is a partition of k_i.

Let σ be a permutation on κ. Call probability/utility pairs p,u and p',u' σ-variants iff the utility and probability of g_i on the first pair is the same as that of $g_{\sigma(i)}$ on the second pair for each $g\in\gamma$, and $p(k_i)=p'(k_\sigma(i))$. We claim that σ-variant interpretations hold the causal expected utility of acts fixed.

To see that this is the case, consider the two instances of the causal expected utility equation, for an original p,u and its σ-variant u',p', for an arbitrary act proposition X:

$$u(X)=\sum_{k_i\in\kappa}p(k_i)\sum_{Y\in\gamma}p(Y|X\wedge k_i)v(X\wedge Y\wedge k_i)$$

$$u'(X)=\sum_{k_i\in\kappa}p'(k_{\sigma(i)})\sum_{Y\in\gamma}p'(Y|X\wedge k_{\sigma(i)})v'(X\wedge Y\wedge k_{\sigma(i)})$$

We want to show that $u(X)=u'(X)$. It will suffice to show that the inner sums are identical, i.e. for each i:

$$\sum_{Y\in\gamma}p(Y|X\wedge k_i)v(X\wedge Y\wedge k_i)=\sum_{Y\in\gamma}p'(Y|X\wedge k_{\sigma(i)})v'(X\wedge Y\wedge k_{\sigma(i)})$$

This suffices since the fact that p,u and p',u' are σ-variants already tells us that $p(k_i)=p'(k_{\sigma(i)})$ for each i. The needed result follows by earlier arguments (in fact, the slight generalization thereof outlined in accompanying footnotes). In more detail: (1) we earlier showed that when we have partitions γ_1 and γ_2 over two spaces k_1 and k_2 related by a bijection which preserves probability and value, the expected value of any direct union of cells of the partition is preserved. (2) The set γ_i whose members are the g_i, i.e. intersections of $g\in\gamma$ with k_i, is a partition of k_i; likewise for $\gamma_{\sigma(i)}$ and $k_{\sigma(i)}$. The mapping from g_i to $g_{\sigma(i)}$ is a bijection between these partitions (this relies on our starting assumption that the intersection of cells of κ and of γ are always non-empty), and by the assumption that p and p' are σ-variants, their restriction to k_i and $k_{\sigma(i)}$ respectively meet the conditions of the earlier result. Hence we know that the (probability and) expected utility of any direct sum of cells is preserved under that mapping. The LHS of the identity above is the expected utility of $X \wedge k_i$, and the RHS is the expected utility of $X \wedge k_{\sigma(i)}$, and these restrictions of X are direct unions of cells of the respective partitions which are images of each other under the induced bijection. Therefore the preservation result just quoted tells us the LHS and RHS are identical, as required.[9] Similar results follow for conditionalizing and Jeffrey-conditionalizing on evidence propositions.

Heuristically, think of the agent's doxastic space as split into distinct slices, each corresponding to distinct elements of the causal background partition. Each slice receives some probabilistic weight from the agent. σ-variant interpretations swap around what slice receives what weight, so e.g. the bubble-slice may receive the weight previously attached to the uniform-slice, and vice versa. There is a uniform grid provided by γ that can be applied to each such slice, and we can cross-identify locations in this grid from slice to slice. The grids are the γ_i and the 'cross-identification' associates g_i and g_j. A σ-variant interpretation doesn't just swap the credence that a given slice is actual—it also swaps the credence and expected value of corresponding grid-locations between the two slices.

Causal decision theory requires that we respect the dividing lines that separate different causal background hypotheses (cells in κ). However, the above result shows that the identities of the background conditions themselves are a matter of complete indifference to rationalizing behaviour. Sensible people give very low credence (0.00001) to the bubble background hypothesis, and high weight (0.9) to uniform ones. But so far as fitting with experience and action is concerned, you can equally well switch these around (so long as compensating adjustments elsewhere are made, as detailed above). So we can, after all, rationalize actions in light of experience in bubble-believing manner even if we are causal decision theorists.

8 Conclusion

The message of the bubble puzzle is that *you can't get out more than you put in*—the constraints on correct interpretation from rationalizing action in the light of experience are strictly limited by the contents of action, on the one hand, and experience, on the other. At least on the EDT model of rationality, we get no fix on the contents of belief and desire that goes beyond the agent's local sphere of interaction with the world, and on the CDT model, we have no fix on how the agent divides their credence across background hypotheses. A Berkleyian might embrace the conclusion of the bubble puzzle argument, and pursue the phenomenalist project of reconstructive analysis of everyday 'external world' content out of the materials thus provided. But most of us are not ready to pursue this line of defence, and must go searching for saving constraints.

[9] That we have a bijection to work with here relies on the non-nullity of g_i, i.e. that cells of κ have non-empty intersection with cells of γ. Notice that even if this assumption holds for just two cells of κ (for all cells of γ) then a restricted form of the result holds—switch those two slices of the agent's doxastic and value profile, leaving the rest fixed, and run the above argument. Further, the bubble vs. uniform scenarios are designed exactly to be consistent with all action and evidence propositions, so this restricted result covers the case of principal interest here. Though I believe the construction can be extended to the case where we relax the starting assumption, and some of the g_i or $g_{\sigma(i)}$ are empty—since the definition of σ-variants guarantees that the cells they are paired with get zero probability—due to the above points I won't explore the issue further here.

As promised in the introduction, our discussion has shown that the bubble puzzle requires careful handling. The sketch in (Lewis, 1984) works well enough in the context of EDT, and we have now seen the general results of which Lewis's sketch was an instance. But Lewis's sketch would be fallacious if CDT were the background theory of rational choice—what he proposed was exactly the kind of swap of normal and counterinductive worlds across cells of the κ-partition that will not in general preserve causal expected utility (ironically, given that those CDT constraints are the constraints of practical rationality that he (Lewis, 1981) endorsed). I have argued that nevertheless, a version of the bubble puzzle can be developed in the CDT setting—but it takes a different form.

The bubble puzzle speaks engages directly with the decision-theoretic representation theorems which are the inspiration for interpretationist metaphysics of content. Unlike many underdetermination arguments, it targets underdetermination in the coarse-grained truth-conditional content. Unlike many, it targets mental content explicitly. From the realist perspective, content underdetermination arguments—from Quine's gavagai, through permutation arguments to Kripkenstein—are tests we run on theories of content to check whether they are constraining enough–unacceptable underdetermination would show our theory is missing some saving constraint. It is frustrating and distorting when these diagnostic tools force us to use the metaphysics of linguistic content as a proxy for the metaphysics of content in general. With the bubble puzzle added to the toolkit, we can test our metaphysics of mental content directly.

References

ALLAIS, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole Americaine. *Econometrica*, **21**(4), 503–546.

DRETSKE, FRED I. 1981. *Knowledge and the Flow of Information*. The David Hume Series. Stanford: CSLI Publications.

ELLIOTT, EDWARD. forthcoming. Probabilism, REpresentation Theorems, and Whether Deliberation Crowds out prediction. *Erkenntnis*.

ELLSBERG, DANIEL. 1961. Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, **75**(4), 643–669.

JEFFREY, RICHARD C. 1965. *The Logic of Decision*. 2nd edn. Chicago and London: University of Chicago Press. Second edition published 1983.

JOYCE, JAMES M. 1999. *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

KRIPKE, SAUL A. 1982. *Wittgenstein on rules and private language*. Oxford: Blackwell.

LEWIS, DAVID K. 1974. 'Radical Interpretation'. *Synthese*, **23**, 331–44. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 108–18.

LEWIS, DAVID K. 1981. 'Causal Decision Theory'. *Australasian Journal of Philosophy*, **59**, 5–30. Reprinted with postscript in Lewis, *Philosophical Papers II* (Oxford University Press, 1986) 305-36.

LEWIS, DAVID K. 1983. 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy*, **61**, 343–377. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 8–55.

LEWIS, DAVID K. 1984. 'Putnam's paradox'. *Australasian Journal of Philosophy*, **62**(3), 221–36. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 56–77.

MAHER, PATRICK. 1993. *Betting on theories*. Cambridge University Press.

MEACHEM, CHRIS, & WEISBERG, JONATHAN. 2011. Representation theorems and the foundations of decision theory. *Australian Journal of Philosophy*, **89**(4), 641–663.

MILLIKAN, RUTH. 1984. *Language, Thought, and Other Biological Categories*.

PAUTZ, ADAM. 2013. Does Phenomenology Ground Mental Content? *In: Phenomenal intentionality*. OUP.

PUTNAM, HILARY. 1980. 'Models and Reality'. *The Journal of Symbolic Logic*, **45**(3), 421–444. Reprinted in Benacerraf and Putnam (eds.) *Philosophy of Mathematics: Selected readings*, second edition (Cambridge University Press, Cambridge: 1983).

PUTNAM, HILARY. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

QUINE, W. V. 1964. ‘Ontological Reduction and the world of numbers’. *Journal of Philosophy*, **61**. Reprinted with substantial changes in Quine, *The Ways of Paradox and Other Essays: Revised and enlarged edition* (Harvard University Press, Cambridge, MA and London, 1976) pp.212—220.

SCHWARZ, WOLFGANG. ms.. Imaginery Foundations.

WALLACE, J. 1977. ‘Only in the context of a sentence do words have any meaning’. *In:* FRENCH, P.A., & T.E. UEHLING, JR. (eds), *Midwest Studies in Philosophy 2: Studies in the Philosophy of Language*. Morris: University of Minnesota Press.

WEATHERSON, BRIAN. 2013. The role of naturalness in Lewis’s theory of meaning. *Journal for the History of Analytic Philosophy*.

ZYNDA, LYLE. 2000. Representation theorems and realism about degrees of belief. *Philosophy of Science*, **67**, 45–69.