# Analytic Functionalism*

Wolfgang Schwarz
30 Jan 2013

## 1 Overview

Lewis held that psychological states are individuated by their causal-functional profile. Pain, for example, may be characterized as whatever state is typically caused by burns and injuries, causes such-and-such signs of distress, a desire for the state to go away, and so on. If it turns out that some biological state, say C-fibre firing, uniquely plays this role, then it has turned out that C-fibre firing is pain. According to Lewis, the roles that characterize mental states can be extracted from *folk psychology*: our tacit but shared beliefs about how mental states interact with one another, what kind of behaviour they tend to cause, and how they change under the impact of perceptual stimulation. Folk psychology implicitly defines our mental vocabulary: 'pain', 'hunger' etc. *mean* 'whatever state plays this and that role'. Thus psychological truths are analytically entailed by non-psychological truths. If you know what 'pain' means, and you know that C-fibre firing plays the relevant role, then you can infer with certainty that people whose C-fibres fire are in pain.

Lewis's position, often called *analytic functionalism*, was inspired by Ryle's [1949] analytic behaviourism, which took psychological predicates to express complex sets of behavioural dispositions. On this view, to say that someone is hungry is to say that they would eat if offered food, that they would more likely go to a restaurant than to a bar, etc. Lewis instead identifies hunger with the inner state that provides the causal basis for these dispositions. (Lewis wasn't the first to make this proposal, see e.g. [Reichenbach 1938].) This vindicates the use of psychological predicates in causal explanations: you went to the restaurant *because* you were hungry; your hunger is part of what caused you to choose the restaurant over the bar. It can also accommodate the fact that mental states often have distinctive behavioural effects only in combination with other mental states. A desire for happiness can manifest itself in all kinds of behaviour, depending on what else the subject believes and desires. In addition, analytic functionalism can allow that paralyzed people (say) may have pain or hunger even though they lack the relevant behavioural dispositions. That's because folk psychology is full of *ceteris paribus* clauses: pain and hunger are defined as states that *typically* ground such-and-such dispositions.

* Thanks to Jochen Faseler, Alan Hájek, Daniel Nolan, Adam Pautz and Daniel Stoljar for helpful comment on an earlier version.

In this article, I will review some of the main tenets of Lewis's philosophy of mind. I will begin with some comments on the methodology Lewis employed in his analysis of psychological terms, which has become standard in functionalist accounts across philosophy. Then I discuss the difference between what are often called 'realizer functionalism' and "role functionalism", and argue that Lewis made the wrong choice. Section 4 presents Lewis's often misunderstood account of intentionality. In section 5, I end with a few pessimistic remarks on the prospect of analyzing phenomenal truths in terms of functional roles.

## 2 The Canberra Plan

Lewis subscribed to the familiar Fregean view that the meaning (in some sense of 'meaning') of referential expressions is given by conditions that specify their reference under hypothetical circumstances. The conditions are common knowledge among fully competent members of the relevant linguistic community, and elicited by intuitions of the kind familiar from Kripke and Gettier. Often the reference conditions associated with a term can be made explicit by looking at a certain body of statements – a "theory" – in which the term occurs. For the term 'entropy', the relevant theory would be thermodynamics or statistical mechanics. For more ordinary terms like 'pain' or 'water', the right starting point is our folk theory, "a generally shared body of tacit belief" about the relevant subject matter [1997b: 333]. The folk theory of water might say that water covers a large part of the Earth, quenches thirst, is typically transparent, and so on.

Imagine such a theory written as a single sentence $T$. The *matrix of* $T$ is the same sentence with all occurrences of the relevant term (or terms) replaced by variables (different variables used for different terms). The matrix of the above water theory, for example, would begin with '$x$ covers a large part of the Earth and $x$ quenches thirst', and so on. This expresses a condition, or *role*. If an entity satisfies the condition, it is called a *realizer* of the role. The chemical substance $H_2O$ is arguably a realizer of the water role, because $H_2O$ covers a large part of the Earth, quenches thirst, and so on. If the matrix of $T$ contains more than one free variables, then a realizer is a list of entities rather than a single entity.

Existentially binding the free variables in the matrix of $T$ yields the *Ramsey sentence of* $T$ (see [Ramsey 1931b]). The Ramsey sentence of the water theory says that there is an $x$ such that $x$ covers a large part of the Earth etc. $T$ is logically stronger than its Ramsey sentence, because it doesn't only say that *something* is so-and-so, but that *water* is so-and-so. Thus the water theory is false and its Ramsey sentence true if something other than water plays the water role, while water itself does not play that role. However, if 'water' is implicitly defined by the water role – i.e., if the reference conditions for 'water' are given by the matrix of the water theory – then this possibility can be ruled out: it

could not turn out that something plays the water role, but water doesn't. Hence the Ramsey sentence of $T$ is a priori equivalent to $T$. No possible discovery could reveal that one of them is true and the other false. The non-empirical part of $T$ that logically goes beyond the Ramsey sentence is captured by the *Carnap conditional of* $T$: the material conditional with the Ramsey sentence of $T$ as antecedent and $T$ itself as consequent (see [Carnap 1963]). The Carnap conditional isolates the analytic, definitional component of the theory.

People sometimes object that there are not enough analytic truths about things like water to make this story work: surely it isn't analytic that water covers a large part of the Earth, or that it quenches thirst. That is true, but irrelevant. The individual conjuncts of the water need not be analytic. What's supposed to be analytic is only the Carnap conditional of the entire theory. In general, the more empirical claims are added to a theory, the harder it gets to falsify its Carnap conditional. Recall that in order to falsify a material conditional, one has to find that the antecedent is true and the consequent false. Could it turn out that something plays the water role, but water does not? If the water role is merely to cover a large part of the Earth, it is easy enough to imagine some such discovery. On the other hand, suppose the water theory contains absolutely everything you believe about water. Could it turn out that all these things are true of *something*, while they are not true of water? How could you discover that this something isn't water if there is no feature which you think water has but this other substance lacks?

The downside of adding a lot of empirical assumptions to a theory is that it may then easily turn out that there is no actual realizer of the theoretical role. The Carnap conditional then remains true, but it becomes useless for locating the relevant phenomenon in fundamental reality. We can't identify water with the chemical kind that realizes the water role if nothing realizes that role. But we also wouldn't say that water does not exist if it turns out that some of our water beliefs are false. We know that most of our theories about most things may well be mistaken, even in quite central respects. The matrix of our theories is therefore too strong to express the reference conditions for the relevant terms. Lewis's usual response is to say that things may count as referents of a term even if they fall somewhat short of realizing the associated role (see [1984: 59], [1994: 298], [1995], [1996: 58], [2004b: 280]). In [1966: 104] and [1970: 83] he makes a slightly different suggestion, which I like better: to weaken the theoretical roles. Let me illustrate a natural way how this can be down.

Let's begin with the Carnap conditional of our total water theory: if something satisfies all our water beliefs, then surely that something is water. What if nothing satisfies all our beliefs, but something satisfies everything except that it does not occur on Mars, or in cucumbers? Then that something is still water. At some point, as we drop or revise more and more of the original assumptions, it becomes unclear whether the thing that

realizes the revised role still deserves the name 'water'. If nothing comes even close to realizing the water role, then there is no water. This is how scientists once discovered that phlogiston or the planet Vulcan do not exist: they found that nothing comes even close to playing the role associated with those terms. Now we have a list of Carnap-style conditionals, with increasingly weakened or modified versions of the original theory $T$:

$$\exists x T(x) \supset T(\text{water}),$$
$$(\neg\exists x T(x) \wedge \exists x T'(x)) \supset T'(\text{water}),$$
$$(\neg\exists x T(x) \wedge \neg\exists x T'(x) \wedge \exists x T''(x)) \supset T''(\text{water})$$
$$\ldots$$

The conjunction of these conditionals is logically equivalent to the single Carnap conditional

$$\exists x T^*(x) \supset T^*(\text{water}),$$

where $T^*(x)$ is defined as

$$T(x) \vee (\neg\exists y T(y) \wedge T'(x)) \vee (\neg\exists y T(y) \wedge \neg\exists y T'(y) \wedge T''(x)) \vee \ldots.$$

$T^*(x)$ is the weakened theoretical role that does a better job at capturing the reference conditions for the relevant term than $T(x)$. Unlike $T(x)$, it also takes into account hypothetical situations in which our theory is false. If nothing realizes the weakened water role, then there is no water.

What if a role is realized by *several* things? In early works like [1970: 83] and [1972: 252], Lewis declared that the relevant term should then be treated as empty. A better response is to say that in such a case, the term is semantically indeterminate between the different candidates, as Lewis says in [1997b: 347], [2004b: 280] and [2009: 220, fn.9].

Let us return to psychological terms. Here the meaning-giving theory is folk psychology. Lewis sometimes suggests that the folk psychological role of mental states is a purely *causal* role. But theoretical roles don't need to be causal. In fact, Lewis himself mentions various folk psychological truths that don't concern causal roles – that toothache is a kind of pain ([1972: 258]), that people who have a conscious experience typically know the essential nature of their experience ([1995]), and that letter boxes are red ([1997b]).

The claim about letter boxes, Lewis suggests, might be part of a psychophysical theory that defines *red experience*. It is interesting not only because it isn't causal but also because it is clearly not shared by all competent speakers of English. According to Lewis, the relevant Carnap conditional is analytic only in a certain sub-group of the English-speaking community. Similarly, one might say – although Lewis does not say so – that various scientific findings have entered into the theoretical role of psychological terms as used by psychologists and neuroscientists. Allowing empirical scientific facts to

constrain a state's theoretical role leads to an account often called *psychofunctionalism*, which is supposed to be an alternative to Lewis's *analytic* functionalism. Lewis does insist that "scientific findings that go beyond common sense must be kept out, on pain of changing the subject" [1974: 112] (see also [1994: 311f.]). Lewis's "subject" presumably was the reference of terms like 'pain' as used outside of scientific circles. Even then, he arguably underestimates the deferential element of ordinary usage. Especially for somewhat technical terms like 'trauma' or 'depression', it may well be part of the folk understanding that these terms denote whatever experts say they denote. The gap between psychofunctionalism and analytic functionalism is therefore less wide than is often assumed.

Lewis never gave a concrete analysis of terms like 'pain'. In other cases, where he does offer an analysis – of causation, chance, conventions or moral values – he often rejects the idea that the analysis ought to precisely match our ordinary usage, or that the relevant theory must be common belief among all competent speakers. Part of the reason is that ordinary language is full of ambiguity, indeterminacy and context-dependence, which can stand in the way of a systematic philosophical analysis. Lewis's analyses are therefore better understood as Carnapian explications: the goal is not to precisely trace our ordinary understanding of the relevant words, but to isolate a theoretically interesting core in the vicinity of our more or less unstable and indeterminate ordinary usage. As he says about moral value: "The best I can hope for is that my [...] theory lands somewhere near the middle of the range of variation and indecision – and also gives something that I, and many more besides, could be content to adopt as our official definition [...], in the unlikely event that we needed an official definition" [1989: 86f.].

What is crucial for Lewis's brand of functionalism is that the relevant terms – whether in their ordinary or in some regimented sense – are really defined by their theoretical role. The statement that $X$ plays the pain role must analytically entail that $X$ is pain. This ensures that the corresponding psychological truths are analytically entailed by, and thus reducible to, non-psychological truths.

Why care about analytic entailment? Many formulations of physicalism do not require analytic or a priori entailment of psychological truths by physical truths. On some accounts, it is enough if the physical and the psychological are connected by brute "metaphysical laws" or by some kind of "grounding" relation. Another popular formulation, which Lewis himself often uses, invokes supervenience or necessitation: every possible world that exactly matches the actual world in all physical respects without containing anything else also matching it in every other respect. This would allow the entailment of the psychological by the physical to be necessary a posteriori, like the entailment of truths about Hesperus by truths about Phosphorus. Lewis never took this possibility seriously. Following Frank Jackson, he argued that when it comes to the entailment of all truths by the fundamental truths, the difference between metaphysical

and semantic or epistemic necessitation disappears. His argument is based on the "two-dimensionalist" premise that every a posteriori necessity is a priori entailed by some ordinary contingent truth about the actual world. For example, the necessary truth that Hesperus = Phosphorus is a priori entailed by the contingent a posteriori truth that one and the same planet plays both the Hesperus role and the Phosphorus role. Now let $P$ be the complete truth about the distribution of fundamental properties and relations in our world, and suppose for reductio that $A$ is some truth for which the conditional $P \supset A$ is necessary a posteriori. By the above premise, there is a further fact $Q$ such that $Q \supset (P \supset A)$ is a priori. But $Q$ itself is made true by $P$. Hence $P \wedge Q$ is equivalent to $P$, and so $Q \supset (P \supset A)$ is equivalent to $P \supset A$. Since the former is a priori, so is the latter (see [Lewis 1994: 296f.], [Lewis 2002], [Jackson 1998a: 93]). The problem with this argument is that if truth-making is only a matter of necessitation, then the fact that $Q$ is made true by $P$ does not entail that $P \wedge Q$ is a priori equivalent to $P$: the link from $P$ to $Q$ may itself be necessary a posteriori. The argument from two-dimensionalism doesn't work. It could be repaired by adding a further premise to the effect that the fundamental truths are "semantically stable" so that primary and secondary intension coincide, but it is not clear to me whether Lewis would have endorsed this premise. At any rate, Lewis's ambition was to show how psychological truths are a priori or analytically entailed by physical truths – merely "metaphysical" connections are not enough.

This form of reductionism presupposes an analytic–synthetic distinction, but only a comparatively mild form. Remember that Lewis accepts that ordinary usage is often too shifty and indeterminate to allow for precise tracing: there may be no fact of the matter whether a particular Carnap-style conditional for 'water' or 'pain' is analytic in English, or in some sub-community of English. In addition, Lewis's account is neutral on the existence and nature of "concepts", understood as psychological entities. It is not assumed that our concept of pain, for example, is in some sense decomposable into more basic concepts. Finally, it is not assumed that predicates like 'pain' can be analyzed by a simple list of individually necessary and jointly sufficient predicates. The analysis takes the form of a rather complicated Carnap conditional.

## 3 Contingent identity

Suppose the folk theory for pain looks something like this:

(1) Pain is typically caused by injuries, it tends to cause distress, etc.

Suppose further that physiological investigations reveal that the role characterized by (1) is realized by C-fibre firing:

(2) C-fibre firing is (the only state that is) typically caused by injuries, tends to cause distress, and so on.

(1) and (2) logically entail that pain is C-fibre firing. More generally, as long as the role characterized by (1) is uniquely realized by some brain state or other, it follows that pain is identical to that brain state. This is Lewis's "argument for the identity theory" [Lewis 1966].

Lewis is confident that some premise along the lines of (2) is true, although this is of course empirical speculation: for all we know in the armchair, it could turn out that the role of pain is occupied by non-physical perturbations of ectoplasma. Premise (1) is also subject to empirical tests: it might turn out that nothing occupies the folk psychological role of pain – even if the role is weakened in the way I suggested in the previous section. In order to separate the empirical and non-empirical premises, it may be advisable to replace premise (1) by the corresponding Carnap conditional:

(1′) *If* some state is typically caused by injuries etc., *then* pain is typically caused by injuries etc.

(1′) and (2) still logically entail (3), but (1′) no longer contains the empirical assumption that something plays the pain role.

One might be suspicious about the parenthetical uniqueness clause in (2). Science can tell us that C-fibre firing plays the pain role, but can it also tell us that the role is not also played by something else, something non-physical (see [Block and Stalnaker 1999])? The answer depends on the details of the role. If part of a role is that $x$ is the only thing that does so-and-so, then it can't turn out that two different things fully realize the role. But Lewis is anyway not committed to the view that premise (2) can be conclusively established by science. The reasons for believing (2) may include considerations of theoretical simplicity or parsimony. What's important for Lewis is that the totality of physical truths entails the psychological truths, not that we can actually derive the psychological truths from our present knowledge about physics. In this context, the totality of physical truths must be understood as including a "that's all" clause, since otherwise all kinds of negative truths will be left out. In this sense, the physical truths do rule out that the pain role is realized by some non-physical state along with C-fibre firing.

On the other hand, it might turn out that no unique physical state occupies the pain role (even if we focus on a single species or individual). For one thing, there are many kinds of pain: toothache, headache, etc. What if these correspond to very different biological states? Even a psychologically determinate type of pain might involve different neural mechanisms on different occasions. There could also be more or less inclusive ways of identifying a realizer: perhaps C-fibre firing is an equally good candidate for playing the pain role in a given individual as some much wider brain state that includes the C-fibre firing. In these cases, Lewis's account would probably say either that 'pain'

is indeterminate between different candidates or that it does denotes a state that is physically and biologically rather disjunctive (compare [1994: 305]).

In Lewis's argument for the identity theory, the identity of mental states with biological states follows logically from folk-psychological definitions and broadly physical facts. There are no gaps to be filled by further considerations of simplicity and parsimony. There is also no logical room for *role functionalism*: the view that pain isn't the realizer of the pain role, but the higher-level property of being in some state or other that realizes the role. According to Lewis, this flatly contradicts the folk psychological characterization of pain. If pain is defined as the state that does so-and-so, and C-fibre firing is the state that does so-and-so, then we aren't free to say that pain isn't actually the state that does so-and-so, but rather the property of being in some state or other that does so-and-so (see [1994: 307f.]).

This means that Lewis faces the stock objection to the identity theory: if pain is C-fibre firing, then only creatures with C-fibres can have pain; but the folk understanding of psychological terms surely doesn't rule out that creatures of radically different kinds can have pain. Lewis offers two replies. First, he suggests to distinguish *pain* from *having pain*: while pain is defined by its causal-functional profile and must therefore be identified with the realizer, having pain is the higher-level property that is common among all creatures whose state occupies the role of pain in their respective cognitive architecture (see [Lewis 1966: 101f.], [Lewis 1994: 307]). Lewis's second reply is that the identity of pain with a particular realizer state is contingent and kind-relative: *in humans, at the actual world*, pain is C-fibre firing (or whatever); in other creatures and at other worlds, pain may be something else (see [1969: 25], [1980a], [1983b: 43-45], [1986b: 267f.], [1994: 305–308]; this position is also defended in [Braddon-Mitchell and Jackson 1996], which is generally an excellent introduction to a theory of mind very close to Lewis's).

How can an identity be relative to a kind or world? First of all, whether or not something realizes a theoretical role can depend on a world, a time or other factors: $H_2O$ satisfies '$x$ covers a large part of the Earth' at the actual world, but not at other possible worlds; Barack Obama satisfies '$x$ is president of the US' in 2012, but not in 2021; Lake Burley Griffin satisfies '$x$ is the closest lake' relative to my present location, but not relative to other locations. When a term is implicitly defined by a matrix like this, we have a choice of either fixing the relevant parameters once and for all or letting the term inherit the referential shiftiness of the matrix. Definite descriptions are usually shifty: 'the closest lake' refers to different things in different contexts and under different embeddings. According to Lewis, mental terms are equally shifty: 'pain' behaves just like 'the state that plays the pain role'. What this picks out depends on the contextually salient type of creature; in humans, the state that plays the pain role may be C-fibre firing, in Martians it may be something else. Thus 'pain' denotes different states in different contexts and under different embeddings.

I find it rather implausible that expressions like 'pain' behave in this manner. In my view, the semantically more plausible choice is to say that mental terms rigidly denote the relevant higher-level property. But then what about Lewis's argument against this proposal? It is true that 'pain' must denote the realizer of the pain role. But the higher-level property may actually qualify as a realizer. This is obscured by formulations like (1), in which 'pain' appears to pick out a particular node in a causal structure, but perhaps this is not the best regimentation of folk psychology. When a sharp pain causes me to withdraw my hand, then what does the causing is arguably not the universal pain, but a concrete *occurrence* of *instantiation* of pain. Suppose we rewrite the folk psychological definition along these lines, characterising pain as a property $x$ such that occurrences of $x$ are typically caused by injuries, cause distress, etc. Now whenever we are in pain, we instantiate the higher-level property of being in some state or other with such-and-such typical causes and effects in creatures of our type. What brought it about that we instantiate this property? Often the cause might be an injury. And often instantiations of this property lie causally upstream of various signs of distress. So the higher-level property can realize the rewritten pain role.

Lewis notes this possibility in [1994: 307], but objects that the higher-level property is too disjunctive, "and therefore no events are essentially havings of it". He also complains that admitting both the lower-level and the higher-level property as causally efficacious "would lead to absurd double-counting of causes".

But who said that the events that are caused by injuries must be *essentially* havings of the property $x$? The modal individuation of events is notoriously murky, and the folk can hardly be assumed to have a settled opinion on this matter (see [Bennett 1988]). Moreover, if a particular pain event is contingently an occurrence of the higher-level property, then the very same event can also be an occurrence of a lower-level property. There is no double counting. This is exactly what Lewis says in [1997b: 341f.] in response to the closely related question whether dispositional properties or their categorical bases should be regarded as causally efficacious. "The very same event", he writes, "that is essentially a having of some causal basis of a certain disposition is also accidentally a having of the disposition itself. So an effect of this event is caused by a having of the basis, and caused also by a having of the disposition. But since these havings are one and the same event, there is no redundant causation." (Compare also [1997a: 142–144], [1986a: 223f.].) I see no reason why the same can't be said for functional properties like pain and their lower-level bases.

One might fear that this solution is overly permissive. To use an example of Lewis's, suppose Mary dies because she put an aluminium ladder against a power line. We want to say that the ladder's electrical conductivity causally contributed to Mary's death, but not the ladder's opacity. However, the causal basis of electrical conductivity in aluminium is the same as the basis of its opacity. So the event that caused Mary's death is just as much

an occurrence of conductivity as an occurrence of opacity. We somehow want to say that the event caused the death only in one of its two guises: Mary died because her ladder conducts electricity, not because it is opaque. One difference that seems relevant here is emphasized in [Jackson and Pettit 1990]: Mary's death is counterfactually invariant under substantial variations of its cause as long as the ladder's electrical conductivity is held fixed. Thus if Mary had used a different ladder, made of different material, she would still have died, as long as her ladder conducts electricity. By contrast, she wouldn't have died if she had used an equally opaque ladder made of wood. These considerations also explain why causal explanations in terms of higher-level, dispositional or functional properties are often better than explanations in terms of their lower-level bases: the higher-level explanation says less about how the effect actually came about, but more about what would have happened under counterfactual circumstances. This is in part why psychological explanations are often more useful than explanations in terms of underlying neurobiological events.

Jackson and Pettit go on to suggest that higher-level properties are not involved in *real* causation at all, because all the causal work is done by more lower-level properties. This is reminiscent of Jaegwon Kim's "causal exclusion argument" against the efficacy of functional properties (see e.g. [Kim 1998]). Such arguments generally rely on a strongly anti-Humean and anti-Lewisian account of causation as a special kind of "production". If causation is something like counterfactual dependence or influence, then it isn't hard to see how higher-level properties can be involved in genuine causation. If Mary hadn't used a ladder that conducts electricity, she wouldn't have died. (It may also be worth pointing out that even if one believes in an anti-Humean force of production, one should accept that not all causation is by production, given the pervasiveness of double prevention, see [Schaffer 2000], [Hall 2004].)

Now I suggested that it isn't really properties that cause, but instantiations of properties. Our problem was that in Mary's case, there is a single event $C$ which is an instantiation both of opacity and of electrical conductivity. Following [Lewis 2004a], let's say that $C$ is a cause of Mary's death $D$ iff counterfactual variations of $C$ go along with variations of $D$. To get the desired outcome, we might then suggest that in a context where $E$ is described as an instantiation of electrical conductivity, the relevant counterfactual variations are variations with respect to conductivity. (See [Lewis 2004b] on the context-dependence of our individuation of events across worlds.) This might explain why Mary's death is caused by the instantiation of conductivity and not by the instantiation of opacity, although these instantiations are one and the same.

With role functionalism as a genuine option, the terminology gets a bit confusing, because we now have two different "roles" with correspondingly different "realizers". First of all, there are the folk psychological roles for states like pain, expressed by a matrix like 'occurrences of $x$ are typically caused by injuries, cause distress', etc. This matrix

expresses a second-order property, a property of properties. On the present proposal, the matrix is realized by the first-order property of being in some state or other occurrences of which are (in creatures of the salient kind) typically caused by injuries, cause distress, etc. This is a first-order property because it applies to individuals rather than properties. It is nevertheless higher-level in the sense that it is largely neutral on the physical or biological constitution of the relevant individuals. The higher-level property is a realizer of the pain role. On the other hand, relative to a particular type of individual, this higher-level property determines another role – another second-order property: to be in pain is to be in some state $y$ that typically does such-and-such in creatures of the relevant kind. For humans, the state that does such-and-such might be C-fibre firing. One might therefore say that C-fibre firing is a "realizer of the pain role", but this pain role is not the role that defines 'pain'.

It may seem odd that the realizer of the pain role (the one that defines 'pain') can effectively be read off from the role itself. We don't have to wait for science to find out that pain is the property of being in some state or other that does such-and-such. But this happens quite often in Canberra planning. Consider the role expressed by '$x$ is a property which applies to all and only unmarried adult men', which we might have retrieved from a somewhat simplistic folk theory of bachelorhood. Again, this is a second-order condition, and we can immediately name a property that satisfies the matrix: being an unmarried adult man.

Does it matter whether psychological properties are identical to higher-level functional properties or to lower-level biological properties? Both accounts can agree on what there is, and even on the truth-conditions of sentences like 'Fred has pain'. Their disagreement only concerns the reference of singular terms like 'pain'. In his later writings, Lewis therefore suggests that the disagreement is superficial and unimportant (see e.g. [1994: 307], [1997a: 142-144], [2004b: 281]). On the other hand, his identification of psychological states with brain states creates follow-up problems that may not arise on the alternative proposal. For example, according to Lewis, folk psychology says that people who are in pain typically know what state they are in: their evidence rules out any possibility where they are not in pain. If pain is C-fibre firing, this would mean that their evidence rules out every possibility where their C-fibres aren't firing. That seems false. Lewis concludes that this part of folk psychology must be rejected (see [1995: 327f.]). If pain is a higher-level property, then knowledge that you are in pain is not knowledge that you are in a certain physiological state. Rather, it is knowledge about the functional profile of your current state. And it is not too implausible that when you introspectively recognise a state as pain, then you recognise it for example as a state that people are generally inclined to avoid (see [Armstrong 1968: 96–99]; see also Daniel Stoljar's article in this handbook).

## 4 Beliefs, desires, decisions

A central part of folk psychology concerns the interaction of beliefs, desires and choices. Crudely put, people typically do what they believe will bring about what they desire. Why does Mary play the cello at 3am? Because she wants to annoy her neighbour and believes that her musical performance is a good way to achieve that. Other facts about Mary's attitudes are also involved, although we would rarely bother to mention them: Mary's desire to annoy her neighbour is not trumped by other desires, she believes that her neighbour will hear the cello, that he won't respond by setting fire to her apartment, and so on. Mary's behaviour is explained not by a single belief and desire, but by a whole system of beliefs, desires, and possibly further attitudes.

There is an infinite number of possible systems of attitudes. Not only are there infinitely many things a person could in principle believe and desire, these attitudes also come in many different degrees: Mary can be more or less certain about how her neighbour will react, and her desire to annoy him may be stronger or weaker. According to Lewis, this part of folk psychology, when systematized, "should look a lot like Bayesian decision theory" [1979: 149]. Bayesian decision theory represents a system of beliefs and desires by a pair of a probability function $P$ and a utility function $U$. Both functions assign numbers to propositions. The $P$ value assigned to a proposition represents the agent's degree of confidence that the proposition is true. (It is not assumed that the agent knows the proposition's objective probability.) The $U$ value represents the degree to which she would like it to be the case that the proposition is true. (It is not assumed that this is a matter of material well-being.) Given a choice between some actions, decision theory then says that the agent, if rational, makes true whatever option $A$ has the highest expected utility, defined as

$$EU(A) = \sum_{S \in W} P(S)U(S\&A),$$

where $W$ a suitable partition of propositions. What exactly makes a partition "suitable" is a controversial matter. Lewis [1981a: 11] suggests that each member of $W$ should be a "maximally specific proposition about how the things [the agent] cares about do and do not depend on his present actions".

Decision theory thus describes a simple connection between any system of (coherent) beliefs and desires and a corresponding set of choice dispositions. Hence one can to some extent read off a rational agent's attitudes from the choices she is disposed to make when confronted with a given set of options. This approach was already developed by Ramsey [1931c], who also proved a *representation theorem* showing that if an agent's choice dispositions satisfy certain qualitative constraints, then there is *unique* system of beliefs and desires that matches her dispositions. (For utilities, "unique" means unique up to positive affine transformation, since utility scales have arbitrary unit and zero.)

More streamlined results to the same effect have since been established for example in [Savage 1954]. However, the qualitative constraints assumed in these theorems are extremely strong. On Lewis's account, this is somewhat ameliorated by the fact that the agent's actual choice dispositions are less important than the typical dispositions of other (perhaps merely possible) agents in the same state (see [1974: 119–121], [1994: 321f., 324 fn.42]). Nevertheless, Lewis clearly deemed the constraints too strong: choice dispositions, he says, do not fully determine an agent's system of beliefs and desires. Further constraints must be invoked (see [1983b: 50–52], [1986c: 107f.]).

Some of these further constraints concern the way systems of belief and desire typically change under the impact of perceptual stimulation. Folk psychology says that under normal circumstances, people who are falling down a crevasse realize that they are falling; that is, they come to assign high probability to the hypothesis that they are falling. Similarly, when people with functioning eye sight confront a red wall, they typically come to believe that there is something red in front of them (see [1980b: 274], [1979: 514,534], [1983b: 50], [1983a: 380], [1994: 299f.], [1997b]).

Yet further (and evidently non-causal) compartments of folk psychology constrain the kinds of things for which people have non-instrumental desires, and the kinds of hypotheses they find a priori plausible or implausible (see [1974: 112f.], [1986c: 38f., 107], [1994: 320]). Lewis suggests that objective naturalness might play a role here, but it remains unclear what exactly this amounts to (see [1983b: 52–54], [1986c: 38f., 107], [1994: 320]).

The total picture then looks as follows. Folk psychology assigns a complex role to entire systems of belief and desire, as represented by a pair of a probability function and a utility function. The characteristic role of such a pair is, in the first place, to cause certain kinds of behaviour under certain conditions. In addition, folk psychology says that systems of beliefs and desires tend to change in a certain way in response to perceptual input, and that they tend to satisfy some general constraints of rationality. If a brain state comes sufficiently close to playing this role, then it can be identified with the relevant system of beliefs and desires.

One can think of folk psychology as a set of interpretation rules: if a state typically gives rise to such-and-such behavioural dispositions, has such-and-such typical causes, etc., then it can be interpreted as being a system of belief and desire with such-and-such content. But it would be wrong to conclude that for Lewis, an agent's beliefs and desires are somehow dependent on, or relative to, an external interpreter. The rules of folk psychology define *what it is* to have certain beliefs and desires. Whether an agent satisfies the definition is then a perfectly objective matter. An agent's beliefs and desires are no less objective and real than her mass and height.

A few aspects of Lewis's account may deserve special emphasis. One is the built-in holism about beliefs and desires. Decision theory, as outlined above, says nothing about

individual beliefs and desires. This may be one reason for Lewis's suggestion that 'beliefs' may be a "bogus plural" [1994: 311]. According to Lewis, it is an open empirical question whether our brains store information holistically, like a map or a connectionist network, or in discrete units. In the latter case, he argues, one might regard the individual units as individual beliefs. A more natural suggestion, I think, is to identify the property of believing a particular proposition $p$ with (roughly) the property of having a system of beliefs and desires which assigns sufficiently high probability to $p$. A belief that $p$ would then be an instantiation of this property. Following the remarks in the previous section, one could even say that such individual beliefs can occupy causal roles: Mary's belief that her neighbour would get annoyed if she played the cello is a state which, on the background of such-and-such other states (and absences), causes her to play the cello. Unlike for example Mary's belief that Napoleon died on St. Helena, the belief about her neighbour is a causal difference-maker for Mary's behaviour.

Many naturalistic accounts of mental content assume not only that there are individual beliefs, but that they really are stored in discrete units, perhaps further decomposed into individual "concepts". These concepts are then said to have their content in virtue of being normally "activated" in the presence of horses or other external objects. Lewis complains that this not only relies on contentious empirical assumptions about the architecture of the brain, but also largely ignores the folk psychological role of mental content (see [1994: 310–324]). Notice that from a decision-theoretic perspective, what needs to be naturalized is not so much the content of beliefs and desires but their strength. That's because what primarily varies from agent to agent, and thus what needs to be explained by physical or functional differences between agents, are the degrees of belief and desire associated with any given proposition. I am not aware of any even remotely plausible answer to this question in terms of causal origins (or conscious phenomenology, for that matter).

A second fact I want to highlight is that on the Lewisian (or Ramseyan) account, the norms of decision theory are not so much normative or descriptive but *constitutive* of agents with beliefs and desires. It is often claimed that ordinary people systematically violate the norms of decision theory, for instance by cooperating in a prisoner dilemma or by rejecting offers in the ultimatum game. Such claims often rest on overly simplistic assumptions about the agents' beliefs and desires, such as an identification of utilities with monetary payoff. (For discussion, see e.g. [Blackburn 1998] and [Joyce 1999: ch.2], but note that while Blackburn endorses the Ramseyean account, Joyce does not.) If an agent's beliefs and desires are *defined* in part as whatever probabilities and utilities make their choices come out rational, then it not easy to establish that the choices people make are generally not rational by the light of their beliefs and desire.

Nevertheless, standard decision theory includes idealizations that aren't part of folk psychology. In particular, it leaves little room for reasoning and a priori inquiry: by the

axioms of the probability calculus, the tautologous proposition which is true at every possible world has probability 1, and whenever a proposition $A$ has probability $x$, then any proposition entailed by $A$ has probability greater than or equal to $x$. Contrary to popular belief, this does not have the absurd consequence that everyone should be certain that Hesperus is Phosphorus, or that Fermat's Last Theorem is true. This would only follows if one were to read '$x$ is certain that $S$' as saying that the relevant agent assigns high probability to the proposition expressed by $S$. But there are good reasons to reject this hypothesis about attitude reports in English. The real "problem of logical omniscience" is that whatever belief is attributed with expressions like '$x$ believes that Fermat's Last Theorem is true', it looks like the relevant proposition (for example about the meaning of certain mathematical symbols) is entailed by other propositions which the agent knows. Solving this problem arguably requires going beyond the assumption that beliefs can be modeled by a standard probability distribution. (See [Lewis 1982: 103], [Stalnaker 1984], [Stalnaker 1991], [Stalnaker 1999b] for some attempts in this direction.)

A third and final aspect of Lewis's account that I want to mention is his insistence that mental content is narrow: that it never differs between intrinsic duplicates within the same world. This is partly due to his emphasis on the connection to behavioural dispositions. In Putnam's [1975] Twin Earth scenario, Oscar and his Twin Earth counterpart Twoscar are causally connected to chemically different substances, but they are disposed to display the exact same responses when put in the same situations. Moreover, on Lewis's account, an agent's attitudes are determined not only by the actual causes and effects of their inner state, but also by the role this state is disposed to play under other actual or hypothetical circumstances. The fact that on our planet, a certain type of belief state is usually caused by the presence of $H_2O$ therefore doesn't entail that the state's content somehow involves $H_2O$. After all, the very same state is usually caused by the presence of XYZ on Twin Earth.

What is revealed by the Twin Earth thought experiment (as well as for example Burge's arthritis example) is that ordinary-language attitude reports can be sensitive to differences in the agent's environment: at least in some contexts, one can truly say that Oscar believes that water covers a large part of his planet, while Twoscar does not. Lewis's explanation is that *de re* belief statements, of the form '$x$ believes that $y$ is $F$', mean that $y$ satisfies some condition $G$ such that $x$ assigns high probability to the proposition that the $G$ is $F$. Oscar, for example, assigns high probability to the proposition that the watery stuff in his surroundings covers a large part of his planet; this stuff is in fact water, i.e. $H_2O$; hence we can truly say that Oscar believes that water covers a large part of his planet (see [1979: §13], [1981b: 412–414], [1986c: 32–34], [1994: 318f.], [Cresswell and von Stechow 1982]). The condition $G$ captures the way the relevant object is "presented" to the agent, which often makes a difference to the agent's behaviour (see [1983b: 50], [1979: 142f.], [1981b], [1994: 323f.]).

The fact that Lewis attributes the same beliefs to Oscar and Twoscar does not mean that the content of their belief state is determined without reference to the external world. [Stalnaker 2004] draws this conclusion and concludes that Lewis must have endorsed a kind of conceptual role account on which the content of a belief state is fixed by its syntactical structure together with a naturalness constraint on its interpretation. But that was not Lewis's view. As we have seen, for Lewis, what determines the content of a system of beliefs and desires is not its inner structure, but its typical perceptual inputs and behavioural outputs.

## 5 Phenomenal character

Many mental states have a distinctive qualitative or phenomenal character: pain has typical causes and effects, but it also has a typical feel. Lewis argues, plausibly enough, that it is not a contingent empirical discovery that pain feels painful: if something doesn't (typically) have the phenomenal character of pain, then it isn't pain. If pain can be analyzed by its functional profile, it follows that the same is true for the phenomenal character of pain. Indeed, according to Lewis, the phenomenal character of pain is simply the property which is satisfied by a state iff the state plays the pain role. But is the phenomenal character of pain really determined by the state's functional profile? Doesn't information about a state's causal or functional role leave it open how that state feels to its subject – and whether it feels like anything at all? Can't one imagine "zombies" that are physically and functionally just like us but lack phenomenal consciousness? Thus goes the conceivability argument against analytic functionalism (see [Chalmers 1996: 93–171]).

It is important to get the argument right. Analytic functionalism does not rule out the possibility that there are creatures physically just like us but without mental states. Suppose it turns out that non-physical perturbations of ectoplasma play the mental roles – which is perfectly compatible with analytic functionalism. Duplicating our physical bodies without duplicating the ectoplasma would then leave out the mental states. Chalmers therefore defines zombie worlds not as worlds physically like the actual world but without consciousness, but as worlds where $P \wedge \neg Q$ is true, where $P$ is the totality of the actual physical truths and $Q$ is (say) the claim that someone experiences pain. But again, a friend of analytic functionalism need not deny that there are worlds where $P \wedge \neg Q$ is true. If non-physical states turn out play the role of mental states, and the terms in $Q$ rigidly denote the realizers, then $P \wedge \neg Q$ describes a genuine possibility. What analytic functionalism, combined with physicalism, has to deny is not the conceivability of a zombie world, nor the conceivability of there being a $P \wedge \neg Q$ world, but the conceivability of $P \wedge \neg Q$ itself, as a hypothesis about the actual world. Since $P$ includes all truths about the typical functional roles of our brain states, $P \wedge \neg Q$ entails something like the

following.

> In creatures like us, there is a (unique) state that is typically caused by injuries, causes distress and a desire for the state to go away etc., but people who are in this state never experience pain.

According to analytic functionalism, this – understood as a hypothesis about our actual situation – is subtly incoherent. It is certainly hard to imagine how we could find out that the claim is true, or how we could ever *have* found that out. For my part, I think there is a reading on which the hypothesis is indeed coherent. But the issue is much less straightforward than many people seem to think.

Perhaps a better argument in support of the same anti-physicalist conclusion looks at the information we receive when we have an experience. Suppose one morning you feel an unusual twitch in your leg. What do you learn when you notice this sensation? Unless you happen to be a neuroscientist, you presumably don't learn that you are in such-and-such a physiological state. If you've never had that sensation before and don't know what it is called, you arguably also don't learn that you're in a state with such-and-such typical causes and effects. What then is the information you acquire? It looks like what you learn is not, or not just, that you have certain physical or functional properties (including indexical properties). But then you seem to have information which is not analytically entailed by the totality of all physical truths.

This line of thought is related to Frank Jackson's [1982] knowledge argument, but it doesn't concern a peculiar situation in which someone knows all relevant physical facts. Moreover, I explicitly asked what information you acquire, not what you learn or come to know, nor how to understand "knowing what it's like". Lewis suggests that when Jackson's Mary comes to know what it's like to see red, she primarily acquires a new set of abilities (as well as indexical information and perhaps new forms of representation, see [1988: 268ff., 278f., 287, 290], [1983c: 131f.], [1994: 294]): Mary learns to visually recognise and classify colours, to imagine red triangles, etc. ([1983c: 131], [1988: 285–288], [1995: 326f.]). She can't learn anything else, because she already knows all the physical facts. By contrast, when you notice your twitch, you are obviously unaware of many physical facts. The problem is that none of these seem to be good candidates for the information you acquire, but it is highly implausible to say that you don't acquire any information at all.

## References

David M. Armstrong [1968]: *A Materialist Theory of the Mind*. London: Routledge

Jonathan Bennett [1988]: *Events and Their Names*. Oxford: Clarendon Press

Simon Blackburn [1998]: *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press

Ned Block and Robert Stalnaker [1999]: "Conceptual Analysis, Dualism, and the Explanatory Gap". *The Philosophical Review*, 108: 1–46

David Braddon-Mitchell and Frank Jackson [1996]: *Philosophy of Mind and Cognition*. Oxford: Blackwell

Rudolf Carnap [1963]: "Replies and Systematic Exposition". In P.A. Schilpp (Ed.) *The Philosophy of Rudolf Carnap,* La Salle (Ill.): Open Court, 859–1016

David Chalmers [1996]: *The Conscious Mind*. New York: Oxford University Press

John Collins, Ned Hall and Laurie A. Paul (Eds.) [2004]: *Causation and Counterfactuals*. Cambridge (Mass.): MIT Press

Max Cresswell and Arnim von Stechow [1982]: "*De Re* Belief Generalized". *Linguistics and Philosophy*, 5: 503–535

Ned Hall [2004]: "Two Mistakes about Credence and Chance". *Australasian Journal of Philosophy*, 82: 93–111

Frank Jackson [1982]: "Epiphenomenal Qualia". *Philosophical Quarterly*, 32: 127–136. In [Jackson 1998b]

— [1998a]: *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press

— [1998b]: *Mind, Method and Conditionals: Selected Essays*. London: Routledge

Frank Jackson and Philip Pettit [1990]: "Causation in the Philosophy of Mind". *Philosophy and Phenomenological Research Supplement*, 50: 195–214

James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

Jaegwon Kim [1998]: *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge (Mass.): MIT Press

David Lewis [1966]: "An Argument for the Identity Theory". *Journal of Philosophy*, 63: 17–25. Reprinted with extensions in [Lewis 1983c]

— [1969]: "Review of *Art, Mind, and Religion*". *Journal of Philosophy*, 66: 22–27

— [1970]: “How to Define Theoretical Terms”. *Journal of Philosophy*, 67: 427–446. Reprinted in [Lewis 1983c]

— [1972]: “Psychophysical and Theoretical Identifications”. *Australasian Journal of Philosophy*, 50: 249–258. Reprinted in [Lewis 1999]

— [1974]: “Radical Interpretation”. *Synthese*, 23: 331–344. Reprinted in [Lewis 1983c]

— [1979]: “Attitudes *De Dicto* and *De Se*”. *The Philosophical Review*, 88: 513–543. Reprinted in [Lewis 1983c]

— [1980a]: “Mad Pain and Martian Pain”. In Ned Block (Hg.), *Readings in the Philosophy of Psychology* Bd.1, Cambridge (Mass.): Harvard University Press, 216–222. Reprinted in [Lewis 1983c]

— [1980b]: “Veridical Hallucination and Prosthetic Vision”. *Australasian Journal of Philosophy*, 58: 239–249. Reprinted in [Lewis 1986d]

— [1981a]: “Causal Decision Theory”. *Australasian Journal of Philosophy*, 59: 5–30. Reprinted in [Lewis 1986d]

— [1981b]: “What Puzzling Pierre Does Not Believe”. *Australasian Journal of Philosophy*, 59: 283–289. Reprinted in [Lewis 1999]

— [1982]: “Logic for Equivocators”. *Noûs*, 16: 431–441

— [1983a]: “Individuation by Acquaintance and by Stipulation”. *The Philosophical Review*, 92: 3–32. Reprinted in [Lewis 1999]

— [1983b]: “New Work for a Theory of Universals”. *Australasian Journal of Philosophy*, 61: 343–377. Reprinted in [Lewis 1999]

— [1983c]: *Philosophical Papers I*. New York, Oxford: Oxford University Press

— [1984]: “Putnam’s Paradox”. *Australasian Journal of Philosophy*, 61: 343–377. Reprinted in [Lewis 1999]

— [1986a]: “Causal Explanation”

— [1986b]: “Events”

— [1986c]: *On the Plurality of Worlds*. Malden (Mass.): Blackwell

— [1986d]: *Philosophical Papers II*. New York, Oxford: Oxford University Press

— [1988]: “What Experience Teaches”. *Proceedings of the Russellian Society*, 13: 29–57. Reprinted in [Lewis 1999]

— [1989]: "Dispositional Theories of Value". *Proceedings of the Aristotelian Society*, Suppl. Vol. 63: 113-137. Reprinted in [Lewis 2000]

— [1994]: "Reduction of Mind". In Samuel Guttenplan (Hg.), *A Companion to the Philosophy of Mind*, Oxford: Blackwell, 412–431. Reprinted in [Lewis 1999]

— [1995]: "Should a Materialist Believe in Qualia?" *Australasian Journal of Philosophy*, 73: 140–144. Reprinted in [Lewis 1999]

— [1996]: "Desire as Belief II". *Mind*, 105: 303–313. Reprinted in [Lewis 2000]

— [1997a]: "Finkish Dispositions". *Philosophical Quarterly*, 47: 143–158

— [1997b]: "Naming the Colours". *Australasian Journal of Philosophy*, 75: 325–342. Reprinted in [Lewis 1999]

— [1999]: *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press

— [2000]: *Papers in Ethics and Social Philosophy*. Cambridge: Cambridge University Press

— [2002]: "Tharp's Third Theorem". *Analysis*, 62: 95–97

— [2004a]: "Causation as Influence". In [Collins et al. 2004], 75–107

— [2004b]: "Void and Object". In [Collins et al. 2004], 277–291

— [2009]: "Ramseyan Humility". In D. Braddon-Mitchell and R. Nola (Eds.) *Conceptual Analysis and Philosophical Naturalism,* Cambridge (Mass.): MIT Press, 203–222

Hilary Putnam [1975]: "The Meaning of 'Meaning' ". In *Language, Mind, and Knowledge,* 131–193

Frank Ramsey [1931a]: *Foundations of Mathematics and other Essays*. London: Routledge & P. Kegan

— [1931b]: "Theories". In [Ramsey 1931a]

— [1931c]: "Truth and Probability". In [Ramsey 1931a]

Hans Reichenbach [1938]: *Experience and Prediction*. Chicago: University of Chicago Press

Gilbert Ryle [1949]: *The Concept of Mind*. Chicago: University of Chicago Press

Leonard Savage [1954]: *The Foundations of Statistics*. New York. Wiley

Jonathan Schaffer [2000]: "Causation by Disconnection". *Philosophy of Science*, 67(2): 285–300

Robert Stalnaker [1984]: *Inquiry*. Cambridge (Mass.): MIT Press

— [1991]: "The Problem of Logical Omniscience I". *Synthese*, 89. In [Stalnaker 1999a]

— [1999a]: *Context and Content*. Oxford: Oxford University Press

— [1999b]: "The Problem of Logical Omniscience II". In [Stalnaker 1999a], 255–273

— [2004]: "Lewis on Intentionality". *Australasian Journal of Philosophy*, 82: 199–212