9. David Lewis on Psychophysical and Theoretical Identifications

Martín Abreu Zavaleta

June 9, 2014

1 The view

Lewis believes that mental states can be identified with some physical state or other. In this, Lewis can be seen as an identity theorist, who claims that every mental state is a physical state. In this, he seemingly defends the same view as Smart or Feigl. However, his argumentative strategy is very different. Smart argued for identity theory partly on the basis of the theoretical parsimony of identity theory vs. the complexity of dualism.

Lewis, on the other hand, claims that we don't need to appeal to simplicity in order to discover theoretical identifications (e.g. that water is H_2O or that pain is stimulation of C-fibers). Instead, such identifications *follow* from the meaning of the terms by which we refer to the things to be identified, together with some empirical facts. We'll examine the reasons why he thinks this is the case. In order to do this, we must first explain his account of theoretical terms.

His general argument is this:

1 Mental state M = the occupant of causal role R (by definition of M)

2 Neural state N = the occupant of causal role R (by the physiological theory); therefore,

3 Mental state M = neural state N (by the transitivity of =)

2 The Ramsey-Lewis definition of theoretical terms

In a paper called "How to define theoretical terms", Lewis introduces a way of defining theoretical terms that uses only logical vocabulary and terms that we understood beforehand. Theoretical terms are simply terms introduced by a theory that were not in our vocabulary before. If we call our theory *T*, we may call the terms introduced by it *T-terms*. We can call *O-terms* the terms in the vocabulary that preceded the introduction of the *T-terms*. The only requirement on the *T-terms* is that they be names. There is no requirement on the *O-terms*, except for the fact that we must grasp them independently of the recently introduced theoretical terms.[1]

The first step to define the T-terms is to write T in a way that exhibits the occurrences of T- terms in it: $T(t_1, ..., t_n)$, where 't_1', ..., 't_n' are T-terms. Given this enunciation of T, we now substitute free

[1] This doesn't prevent us from introducing terms that denote properties, but will merely make a difference with respect to the way we talk about these properties. For instance, if we want to define a term like 'red', instead of using the predicate 'red', we use the name that denotes the property expressed by that predicate, e.g. 'redness'.

variables for all the T-terms. Thus, we get what Lewis call the *realization formula* of T: $T(x_1, ..., x_n)$. Note that the realization formula of T contains only O-terms and free variables.

Keeping the interpretation of the O-terms fixed, any n-tuple that satisfies the realization formula of T will be said to *realize* T, or to be a *realization* of T. Lewis makes a case for adopting the following meaning postulates:

(1) If there is a unique n-tuple that realizes T, then the T-terms name, respecitvely, the components of that n-tuple.

(2) If there is more than one realization of T, then the T-terms name nothing.

(3) If there is no realization of T, then the T-terms name nothing.

With the meaning postulates in place, we can determine the denotation of each term introduced by the theory: the *i*th T-term will denote the *i*th member of the n-tuple that uniquely realizes T, if T has a unique realization, and will name nothing otherwise. Thus,

t_1 = the first member of the unique n-tuple that realizes T

t_2 = the second member of the unique n-tuple that realizes T

And so on for the rest of the T-terms.

At some points, it will be easier to refer to the denotations of the T-terms as the entities that play or realize a certain *causal (or functional) role* in the unique realization of T. By a causal role I mean the property of standing in such and such causal relations to other things, properties, classes, et cetera, as determined by T. For instance, T determines a causal role for t_1. Such causal role can be obtained by taking all the sentences of T in which 't_1' occurs and substituting free variables for them. The property expressed by the resulting open sentence is a causal role.

We can talk about the role that T assigns to t_i as the t_i-role. Instead of the definitions for the T-terms used above, we may say that t_i is the member of the unique realization of T that realizes the t_i role. Note that a functional role is not the same as the thing that realizes the role: playing a functional role is a property of the property that realizes the role.

According to Lewis's method, the terms introduced by theories that don't have a unique realization are denotationless. This may be the right result for theories that we take to be hopeless, like the theory of phlogiston. But there are theories that, though unrealized, are *nearly* realized. That is, there is a unique n-tuple of entities that realizes a weaker version of the theory, even though it doesn't realize the original theory.

In these cases, we would like to say that the terms introduced by T denote the elements of the nearest realization of T, as long as this realization comes close enough to realizing the original theory.

Here is a simple example. Suppose that 'pain' was introduced by the following pain-theory:

> (T) For any x, if x suffers tissue damage and *is normally alert*, x *is in pain*; if x is awake, x tends to be *normally alert*; if x *is in pain*, x winces and groans and *goes into a state of distress*; and if x is not *normally alert* or x is *in a state of distress*, x tends to make more typing errors. (Kim, p. 170)

The theoretical terms introduced by this pain theory are in italics. **Question:** Can you define these terms using the Ramsey-Lewis method? What would the definitions look like?

Identifying mental and physical states

It's easy to see how, once we adopt this method for defining theoretical terms, we may explain our discovery that the entities denoted by the T-terms are the same as the entities denoted by some other terms—including old terms and terms introduced by some theory other than T. All we need to find out is that the entities that we purport to identify with the denotations of the T-terms satisfy the descriptions that we used to define each of those terms, respectively.

Here's a brief example on how to define 'gene' using this method. Merely for illustrative purposes, let's suppose that Mendel's original genetic theory introduced only one new term, 'gene'. Let's further suppose that Mendel's theory just says that genes are passed from one generation to the next and determine hereditary traits according to the laws of segregation and independent assortment. Let's introduce the term 'genehood' to denote the property of being a gene, and modify the theory accordingly—thus getting the sentence that all the things which have genehood are such and such. Following Lewis's method, we substitute free variables for the occurrences of 'genehood'. In this case, we get the following open formula: all the things that have x are passed from one generation to the next and determine hereditary traits according to the law of segregation and the law of independent assortment. Since the theory introduces only one new term, we can define 'genehood' as follows:

> Genehood = the unique property had only by the things that are passed from one generation to the next and determine hereditary traits according to the laws of segregation and independent assortment.

According to Lewis, if there is no unique such property, the term 'genehood' names nothing. However, there may be a unique property that *nearly realizes* Mendel's theory. For instance, perhaps nothing realizes Mendel's original theory formulated in terms of the laws of segregation and independent assortment, but his theory entails the weaker theory that all the things that have genehood are passed from one generation to the next and determine hereditary traits—no mention of Mendel's laws in this case.

If there is a unique property that realizes the weaker theory, then the term 'genehood' denotes that property. In this case, 'genehood' denotes the property had only by the things that are passed from one generation to the next and determine hereditary traits. How can we discover the physical property that genes are identical to? Well, we may discover that the property of being a sequence of nucleic acids of a certain kind is the unique property had only by the things that are passed from one generation to the next and determine hereditary traits. This is all we need in order to discover that genehood is the property of being a sequence of nucleic acids of that kind.

3 Varieties of functionalism

If we follow Lewis's method to the letter, we're going to need to say which theory we are using to define mental terms. Here, two alternatives have been salient in the literature. The first is the one advocated by Lewis, usually called *folk* or *commonsense* functionalism. The other is *psychofunctionalism*:

Commonsense-functionalism: In our common use of mental terms, we are committed to a certain theory about mental states. The theory is supposed to be an a priori theory, made up of

platitudes that everyone competent in the use of mental terms is supposed to recognize as true. Lewis puts it this way: “Think of common-sense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses." The result should be a theory that we can use to define mental terms.

Psychofunctionalism: The theory with which we should define mental terms is an a posteriori theory, which we only learn as a result of scientific investigation of how our minds work (the sort of investigation they do in cognitive science labs).

Using Lewis’s method for defining theoretical terms, there are two further choices, which we’ll illustrate with the mental state of being in pain:

Role functionalism: According to role functionalists, being in pain is having a property that has the higher-order property of satisfying the pain-role—as defined by some psychological theory. Thus, something can be in pain as long as it has a property that realizes the pain-role, whether it be stimulation of C-fibers (for humans) or inflammation of M-nodes (for martians).

Realizer functionalism: According to realizer functionalists, being in pain is having the property that realizes the pain-role—again, as defined by some psychological theory. For instance, if what realizes the pain-role in humans is stimulation of C-fibers, then being in pain is to have stimulation of C-fibers.

It’s important to notice a characteristic of Lewis’s version of functionalism: he claims that whatever plays the roles determined by folk psychology must be unique. This seems to block the possibility of multiple realization. One way of doing this is by weakening our definitions of mental terms: instead of saying that pain, say, is the first member of the unique n-tuple that realizes T, we might say that *something is in pain* if it has the property that plays the pain-role in one of the realizations of T.