GEORGE BEALER

AN INCONSISTENCY IN FUNCTIONALISM*

1. THE TWO THESES OF FUNCTIONALISM

Behaviorism and naive physiological reductionism are the forerunners of functionalism in psychology and philosophy of mind. Like its forerunners, functionalism is not a single, unified theory. Rather, it is an intellectual movement. Consequently generalizations about functionalism run a risk of over-simplification. Bearing this in mind, I will distinguish two forms of functionalism – one oriented toward behavior and the other oriented toward physiology. In this discussion I will count a given doctrine as functionalistic (whether it be behavioristic or physiological in orientation) if an only if it is committed to a certain relevant pair of theses – one negative and one positive.

The negative thesis of behavioral functionalism is tantamount to the rejection of behaviorism itself. The thesis is that (terms which express) mental properties, relations or states do not have *ordinary explicit definitions* which appeal solely to (terms which express) behavioral properties, relations or states. The positive thesis of behavioral functionalism is that (terms which express) mental properties, relations or states do, by contrast, have purely behavioral *functional definitions*, i.e., they can be defined solely in terms of how they *function together* in (theories concerning) the typical psycho-behavioral causal manifold.

The negative and positive theses of physiological functionalism are just what one would expect: (terms which express) mental properties, relations or states do not have *ordinary explicit definitions* which appeal solely to (terms which express) physiological properties, relations or states; however, (terms which express) mental properties or states do have purely physiological *functional definitions*; i.e., they can be defined solely in terms of how they *function together* in (theories concerning) the typical psycho-physiological causal manifold.

If his positive thesis is correct, the functionalist is then free to uphold some form of physicalism despite the failure of behaviorism and naive

Synthese 38 (1978) 333–372. *All Rights Reserved.*

physiological reductionism alleged in the negative theses of functionalism.[1] It is this physicalism which attracts many followers to functionalism.

I will now say a word about the intuitive motivation for functionalism and, in particular, for behavioral functionalism. Both here and in what follows virtually everything I will say about behavioral functionalism applies, *mutatis mutandis*, to physiological functionalism. For convenience, therefore, I will throughout this paper focus primarily on the former.[2]

The negative thesis of behavioral functionalism is motivated by the following sort of considerations. Particular mental states do not, without regard to other mental states, have any particular elementary behavioral correlates, whether they be behavioral inputs or behavioral outputs. Consider the case of behavioral outputs. Typically, a particular mental state leads to a particular behavioral output only by virtue of its causal interaction with other mental states. For example, a desire for food might produce a trip to the refrigerator if it is believed that food is there but might not if it is believed that there is no food there. Likewise, a belief that there is food in the refrigerator might produce a trip there if there is a desire to eat but might not if there is no desire to eat. Belief does not have an ordinary explicit definition which appeals solely to behavioral outputs; appeal to the notion of desire is required. Likewise, desire lacks an ordinary explicit definition which appeals solely to behavioral outputs; appeal to the notion of belief is required. Thus, neither belief nor desire has an ordinary explicit definition which appeals solely to behavioral outputs.

The same sort of situation holds for behavioral inputs. Typically, behavioral inputs produce new mental states only by virtue of causal interaction with prior mental states. For example, food deprivation might produce a desire for food if it is believed that food will relieve hunger but not if it is believed that food will cause, e.g., painful cramps. Likewise, retinal stimulation by reflections from a relevant page of a medical text on food deprivation might produce a belief that food will cause painful cramps; that is, it might cause this belief if there is a desire to perform certain preliminary computations called for by the text but not if there is no such desire but instead a strong desire to eat no matter what the risk. Belief does not have an ordinary explicit definition which appeals solely to behavioral inputs; appeal to the notion of desire is required. Likewise, desire lacks an ordinary explicit definition which appeals solely to behavioral inputs; appeal to the notion of belief is required. Thus, neither belief nor desire has an ordinary explicit definition which appeals solely to behavioral inputs. In addition, appeal to, e.g., behavioral input/output *pairs* and like notions does not help matters for the behaviorist. The functionalist believes that comparable difficulties beset *every* ordinary explicit behavioral definition of belief and desire, even those definitions which use these more sophisticated behavioral notions.

But, what, then, are belief and desire? According to behavioral functionalism, belief and desire are those relations which, in precisely the ways just characterized, causally act upon one another to produce appropriate behavior. From this claim the behavioral functionalist concludes that belief and desire can be *functionally defined* solely in terms of behavior. He concludes, for example, that belief can be defined as a (the) relation which, in the way characterized above, causally interacts with another unnamed relation [i.e., desire] to produce the appropriate behavior; likewise desire can be defined as a (the) relation which, in the corresponding way, causally interacts with another unnamed relation [i.e., belief] to produce the appropriate behavior. These, then, are the sort of definitions – i.e., functional definitions – upon which the positive thesis of behavioral functionalism rests its case.

Let us consider some characteristic statements made by two leading functionalists. Expressing his support for what I have called the negative thesis of behavioral functionalism, Gilbert Harman asserts:

> There is no noncircular way to specify the relevant dispositions. For they are dispositions to act in certain situations; and the relevant situations essentially include beliefs about the situation and desires concerning it. What a man will do if he hits his thumb with a hammer depends on who he believes is watching and what desires he has concerning his relationship to the watchers. But beliefs are dispositions to act in certain ways only given certain desires, whereas desires are dispositions to act in certain ways only given certain beliefs. A belief that it will rain will be manifested in the carrying of an umbrella only in the presence of a desire not to get wet; and the desire for money will manifest itself in acts that tend to get one money only if one believes that those acts will get one money. Since even in theory there is no noncircular way to specify relevant dispositions in pure behavioral terms, behaviorism cannot provide an adequate account of mental processes and experiences.[3]

In connection with what I have called the positive thesis of behavioral functionalism, Harman states his support as follows:

I will defend a kind of functionalism which defines mental states and processes in terms of their roles in a functional system.[4]

> A psychological model represents a more or less rigorously specified device that is intended to be able to duplicate the relevant behavior of a person. If the device is sufficiently described, it should be realizable as a robot or as I shall say, an automaton.
>
> An abstract automaton is specified by its program. The program indicates possible reactions to input, how internal states plus input can yield other internal states, and how internal states and input can lead to various sorts of output. In a psychological model, input can represent the effect of perception and output can represent intentional action.
>
> As Aristotle pointed out, mental states and processes are to be functionally defined. They are constituted by their function or role in the relevant program. To understand desire, belief, and reasoning is to understand how desires, beliefs, and instances of reasoning function in a human psychology.[5]

In support of what I am calling the negative thesis of behavioral functionalism, David Lewis ascribes to his view an essential virtue which behaviorism lacks:

> ... it allows us to include other experiences among the typical causes and effects by which an experience is defined. It is crucial that we should be able to do so in order that we may do justice, in defining experiences by their causal roles, to the introspective accessibility which is such an important feature of any experience. For the introspective accessibility of an experience is its propensity reliably to cause other (future or simultaneous) experiences directed intentionally upon it.... The requisite freedom to interdefine experiences is not available in general under behaviorism; interdefinition of experiences is permissible only if it can in principle be eliminated, which is so only if it happens to be possible to arrange experiences in a hierarchy of definitional priority.[6]

Elucidating what I am calling the positive thesis of behavioral functionalism, Lewis goes on to say:

> We, on the other hand, may allow interdefinition with no such constraint. We may expect to get mutually interdefined families of experiences.... Whatever occupies the definitive causal role of an experience in such a family does so by virtue of its own membership in a [physical] causal isomorph of the family of experiences, that is, in a system of [physical] states having the same pattern of causal connections with one another and the same causal connections with states outside the family, viz., the stimuli and behavior. The isomorphism guarantees that, if the family is identified *throughout* with its [physical] isomorph, then experiences in the family will have their definitive causal roles.[7]
>
> The definitive causal role of an experience is expressible by a finite set of conditions

> that specify its typical causes and its typical effects under various circumstances. By analytic necessity these conditions are true of the experience and jointly distinctive of it.[8]

Both the negative thesis and the positive thesis of functionalism have considerable intuitive appeal. In one form or the other – behavioral or physiological – functionalism has received extremely widespread acceptance among philosophers of mind and philosophically minded cognitive psychologists. Indeed, the support for functionalism appears comparable in enthusiasm to the support received by behaviorism and naive physiological reductionism in their heyday.

Despite the intuitive appeal of the two theses of functionalism and despite their enthusiastic support, it has proven difficult to determine *in general* the genuine merits of these theses. The reason for this is that discussions – philosophical and psychological – have, in most instances, been quite *informal* in character. In particular, the fundamental distinction upon which functionalism is based – i.e., the distinction between *ordinary explicit definitions* and *functional definitions* – has, in most of the discussions, been treated in an imprecise manner which has prohibited decisive tests.

The purpose of the present paper is (1) to explicate the functionalist's distinction between ordinary explicit definitions and functional definitions and (2) to make some points which count against functionalism when the ordinary-explicit/functional distinction is so explicated. It will be my intention to explicate this distinction in a way which yields a natural interpretation of the leading informal versions of functionalism in the literature today.

My main conclusion will be that, given the suggested explication, functional definitions of mental predicates exist if and only if ordinary explicit definitions exist as well. Thus, the positive thesis of functionalism is true if and only if the negative thesis is false. In this, functionalism is inconsistent. Do functional definitions of mental predicates exist? If the functionalists' arguments against the existence of ordinary explicit definitions are sound arguments, then it follows via my main conclusion that functional definitions of mental predicates do not exist.

Before proceeding it should be noted that, since my discussion will be fairly general in form, much of what I will say will apply equally to functionalisms in (philosophy of) biology, (philosophy of) social science, and in (philosophy of) science, generally. However, I will not comment further on these applications.

2. THE DISTINCTION BETWEEN ORDINARY EXPLICIT AND FUNCTIONAL DEFINITIONS

a. Ordinary Explicit Definitions.

Let $T_1, \ldots, T_n$, $O_1, \ldots, O_m$ be predicates of some standard first-order theory A. $T_1, \ldots, T_n$ are to be thought of as mental predicates, and $O_1, \ldots, O_m$ are to be thought of as physical predicates, behavioural or physiological.

For greater generality, $T_1, \ldots, T_n$ may also be thought of as theoretical predicates, and $O_1, \ldots, O_m$, as observational predicates. For convenience, I will confine the discussion to *binary* predicates. No generality will be lost in so doing.

The aim of this portion of the paper is to characterize the class of formulas which, according to functionalists, qualify syntactically as ordinary explicit definitions of $T_1, \ldots, T_n$ in terms of $O_1, \ldots, O_m$. Of the formulas in this class, the most elementary type (henceforth called type-0) are simply the first-order formulas which have no non-logical constants beyond $O_1, \ldots, O_m$.[9] The reason why all such first-order formulas qualify syntactically as ordinary explicit definitions is clear. The interaction *among the several* theoretical relations expressed by $T_1, \ldots, T_n$ is in no way represented by first-order O-formulas. However, the representation of such interaction is what is distinctive about functional definitions.

In the next most elementary type of ordinary explicit definition (henceforth called type-1) we find certain second-order formulas, namely, second-order formulas which are formed from *first-order inductive definitions* by use of the Frege-Dedekind technique. The initial inductive definitions can, of course, have no non-logical constants beyond $O_1, \ldots, O_m$. If $Q(x, y)$ is an inductive definition of $T(x, y)$, the Frege-Dedekind direct definition of $T(x, y)$ is the following:

> $T(x, y)$ iff_{def}
> x, y satisfy every relation (and hence, the smallest relation) which satisfies the clauses of the inductive definition Q.

Consider an example. The following is a first-order inductive definition of the ancestor relation given in terms of the parent relation:

(i) u is a parent of $v \supset u$ is an ancestor of v

(ii) $(\exists w)$ (u is a parent of w & w is an ancestor of v) $\supset u$ is an ancestor of v

Hence, the definiens in the following is a type-1 ordinary explicit definition of the ancestor relation formed, via the Frege-Dedekind technique, from the above inductive definition:

x is an ancestor of y iff_{def}

$(\forall F)\{[(\forall u, v)$ (u is a parent of $v \supset F(u, v)$ and

$(\forall u, v)((\exists w)$ (u is a parent of w & $F(w, y)) \supset$

$F(u, v))] \supset F(x, y)\}$.

The class of type-1 ordinary explicit definitions is defined as follows:

> Q is a type-1 ordinary explicit definitions of T given in terms of $O_1, \ldots, O_m$ if and only if (a) for some inductive definition B whose only non-logical constants are $O_1, \ldots, O_m$ and whose quantified variables are all first-order, Q is the result of turning B into a second-order direct definition via the Frege-Dedekind technique or (b) Q is a compound formula constructed out of (i) $O_1, \ldots, O_m$, (ii) other type-1 definitions and (iii) the logical connectives and first-order quantifiers.

Generalizing, we get the following hierarchy of ordinary explicit definitions: type-0, type-1, . . . , type-n . . . , where a type-n definition, $n > 1$, is just like some type-1 definition except that it contains one or more type($n - 1$) definitions.[10]

Using the above ideas, I propose to explicate the functionalist's notion of an ordinary explicit definition as follows:

> A formula Q qualifies syntactically as an *ordinary explicit definition* of the theoretical (mental) predicate T given in terms of the observational (physical) predicates $O_1, \ldots, O_m$, if and only if Q is somewhere in the above hierarchy of definitions.

How can we be assured that every formula which satisfies this explicans is indeed an ordinary explicit definition and not a functional definition? First, inductive definitions are paradigm non-functional definitions. Second, if the definitions $Q_1, \ldots, Q_n$ of $T_1, \ldots, T_n$, respectively, are in the above hierarchy, then these definitions do not characterize the *interaction* among the theoretical (mental) relations expressed by $T_1, \ldots, T_n$. For every pair i, j, where $1 \leqslant i, j \leqslant n$, either the relation expressed by T_i or the relation expressed by T_j (or both) is defined *wholly independently* of the other relation. The fact that $Q_1, \ldots, Q_n$ – and hence, the relations expressed by $T_1, \ldots, T_n$ – admit of this sort of *separation* makes them non-functional.

Perhaps there are some formulas which intuitively qualify as ordinary explicit definitions but which do not qualify as such according to the suggested explication. Even if this is so, this explication does not in any way prejudice the case against functionalism. Indeed, by circumscribing the class of ordinary explicit definitions, we are actually making more likely that the negative thesis of functionalism will be correct. At the same time, we do not lessen the chances that the positive thesis will be correct. For this reason, it is quite safe to adopt the explication as it stands.

b. *Functional Definitions*

Not surprisingly the following will be my explication of the functionalist's notion of a functional definition:

> Q qualifies syntactically as a *functional definition* of T given in terms of $O_1, \ldots, O_m$ if and only if (a) Q is identical to – or provably equivalent to – some second-order formula which contains no non-logical constants beyond $O_1, \ldots, O_m$ and (b) Q is not an ordinary explicit definition of T given in terms of $O_1, \ldots, O_m$.

I believe that this explication is adequate. However, since I have as yet given no concrete examples of functional definitions, it is no doubt difficult to see clearly what a functional definition is and to see why the explication is adequate. Therefore, I will go over in detail some paradigm sorts of functional definitions. There are, evidently, three paradigm sorts of functional definition. I will characterize each of these three sorts, and from this characterization it will become clearer what a functional definition is and why they fall within the scope of the suggested explication.

The first such paradigm is best understood against the background provided by (i) F. P. Ramsey's method for eliminating theoretical terms and (ii) the associated notion of a *Ramsey constant*, developed by R. M. Martin. In 'Theories'[11] Ramsey provided us with a failsafe method for *eliminating* the theoretical terms from a theory while preserving its observational content. In 'On Theoretical Constructs and Ramsey Constants'[12] R. M. Martin maintains that Ramsey's method can be used in the formulation of *definitions* of the theoretical terms of a theory, where these definitions use only the observational terms of the theory.

To see what these proposals amount to, consider a standard first-order theory A whose theoretical (mental) predicates are $T_1, \ldots, T_n$ and whose observational (physical) predicates are $O_1, \ldots, O_m$. In what follows I will, for convenience, confine the discussion to those theories A whose axioms are *finite* in number and, as before, whose predicates are all *binary*. No generality will be lost. Let

$$(\exists F_1) \ldots (\exists F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_m)]$$

be the result of (1) conjoining the axioms of A, (2) replacing the constants $T_1, \ldots, T_n$ with distinct predicate variables $F_1, \ldots, F_n$, and (3) existentially quantifying $F_1, \ldots, F_n$. This sentence is a *Ramsey sentence*. Notice that this sentence contains no theoretical predicates. For our purposes we may characterize Ramsey's discovery as follows: for any sentence Q whose non-logical constants are selected from $O_1, \ldots, O_m$,

$$A \vdash Q$$

if and only if

$$(\exists F_1) \ldots (\exists F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_m)] \vdash Q$$

This is to say, Q is provable in A if and only if Q is provable from the corresponding Ramsey sentence.[13] Thus, the observational content of A and of the Ramsey sentence are the same.

Now for R. M. Martin's method of Ramsey constants. According to Martin, the following, for any predicate T_i in any theory $A(T_1, \ldots, T_n, O_l, \ldots, O_m)$, is an adequate definition:

$$T_i(x, y) \text{ iff}_{def}$$
$$(\exists F_1) \ldots (\exists F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_m) \;\&\; F_i(x, y)].$$

Henceforth, $T_i^*(x,y)$ will be used as shorthand for the right-hand side. Expressions such as $T_i^*(x,y)$ are what Martin calls *Ramsey constants*. To see the intuitive content of these definitions, consider the case where $n = 1$ and $m = 1$:

$$T(x,y)\ \mathit{iff}_{\mathrm{def}}(\exists F)[A(F,O)\ \&\ F(x,y)].$$

This says that, by definition, x, y satisfies the theoretical relation T if and only if x, y satisfies some relation F which makes the theory $A(F, O)$ true.

In his elegant and ingenious 'Method in Philosophical Psychology' Professor H. P. Grice offers one of the very few precise statements in print of how behaviourally oriented functional definitions might be formulated.[14] The functional definitions of mental predicates ventured there are none other than *Ramsey constants*, although they are not explicitly identified either as functional definitions or as Ramsey constants. That is, from a formal point of view, when the mental predicates of a given psychological theory are treated like the 'theoretical predicates' $T_1, \ldots, T_n$, and the behavioral predicates are treated like the 'observational predicates' $O_1, \ldots, O_m$, any given mental predicate T_i, $1 \leqslant i \leqslant n$, is functionally defined by the associated Ramsey constant $T_i^*(x,y)$:

$$T_i(x,y)\ \mathit{iff}_{\mathrm{def}}\ (\exists F_1)\cdots(\exists F_n)\ [A(F_1,\cdots,F_n,O_1,\cdots,O_m)\ \&$$
$$F_i(x,y)].$$

Now although these definitions are not explicitly called functional by Professor Grice, they clearly qualify as such. To see why this is so, it is helpful to consider the simple example where $n = 2$ and $m = 2$:

$$x \text{ believes } y\ \ \mathit{iff}_{\mathrm{def}}\ (\exists B)(\exists D)[A(B,D,O_1,O_2)\ \&\ B(x,y)]$$

$$x \text{ desires } y\ \ \mathit{iff}_{\mathrm{def}}\ (\exists B)(\exists D)[A(B,D,O_1,O_2)\ \&\ D(x,y)]$$

The former definition says, in effect, that x believes y if and only if there is a relation B and a relation D whose behavior with respect to each other and to behaviors O_1 and O_2 is the same as the behavior of belief and desire with respect to each other and to the behaviors O_1 and O_2, and x stands in relation B to y. That is, there is a relation B and a relation D which *function* with respect to each other and to the behaviors O_1 and O_2 in the same way belief and desire function with respect to each other and to the behaviors O_1 and O_2, and x stands in the relation B to y.[15]

Thus, we arrive at the first paradigm sort of functional definition, namely, those Ramsey constants T_i^* which are constructed from any first-order theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$ wherein $n \geqslant 2$. The purpose of the restriction on n is, of course, to insure that these Ramsey constants represent the way – as determined by the whole theory $A(F_I, \ldots, F_n, O_1, \ldots, O_m)$ – in which the theoretical (mental) relation F_i *interacts with other* theoretical (mental) relations $F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_n$ to produce observable (physical) phenomena involving the relations $O_1, \ldots, O_m$.

We are now in a position to characterize the remaining two paradigm sorts of functional definitions. They are:

$$(1) \qquad T_i(x,y) \; iff_{\text{def}} \; (\forall F_1, \cdots, F_n)[A(F_1, \cdots, F_n O_1, \cdots, O_m) \supset F_i(x,y)]$$

$$(2) \qquad T_i(x,y) \; iff_{\text{def}} \; (\exists 1 F_1) \cdots (\exists 1 F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \;\&\; F_i(x,y)]$$

where, as before, $n \geqslant 2$. The definiens in (1) will be called a *Carnap constant* and represented by $T_i^{**}(x,y)$.[16] The definiens of (2) will be called a *Lewis constant* and will be represented with $T_i^{***}(x,y)$[17] To see the intuitive content of Carnap and Lewis constants, consider the case in which $n = 1$ and $m = 1$:

$$T(x,y) \; iff_{\text{def}} \; (\forall F)[A(F, O) \supset F(x,y)]$$

$$T(x,y) \; iff_{\text{def}} \; (\exists 1 F)[A(F, O) \;\&\; F(x,y)].$$

The former says that, by definition x, y satisfies the theoretical predicate T if and only if x, y satisfies *every* relation F which makes the theory $A(F, O)$ true. The latter says that, by definition, x, y satisfies the theoretical predicate T if and only if x, y satisfies the *unique* relation F which makes the theory $A(F, O)$ true.

Given the previously suggested explication of the notion of a functional definition, all three paradigm sorts of functional definition – Ramsey constants T_i^*, Carnap constants T_i^{**} and Lewis constants T_i^{***} formed from first-order theories $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$, wherein $n \geqslant 2$ – clearly qualify as functional definitions. Now, one might plausibly hold that *every* functional definition is either a Ramsey constant, a Carnap constant or a Lewis constant formed from some first-order formula $A(T_1, \ldots, T_n,$

$O_l, \ldots, O_m$) where $n \geqslant 2$.[18] However, I have chosen not to base my explication of the notion of a functional definition on this proposition. My reason for doing this is the following: It is difficult to be certain that all candidate functional definitions which functionalists might put forth would be covered by the resulting circumscribed explication. In this connection, consider the version of (physiologically oriented) functionalism held by Hilary Putnam:

> I am inclined to hold the view that psychological properties would be reduced not to physical$_2$ properties in the usual sense (i.e, first-order combinations of fundamental magnitudes), but to *functional states*, where crude examples of the kinds of properties I call 'functional states' would be (a) the property of being a finite automaton with a certain machine table; and (b) the property of being a finite automaton with a certain machine table *and* being in the state described in a certain way in the table. To say that a finite automaton has a certain machine table is to say that *there are properties* (in the sense of physical$_1$ properties) which the object has (i.e., it always has one of them), and which succeed each other in accordance with a certain rule. Thus, the property of having a certain machine table is a *property of having properties which* ... – although a property of the first-level (a property of things), it is of 'second order' in the old Russell-Whitehead sense, in that its definition involves quantification over (first-order) physical$_1$ properties. This is a general characteristic of a 'functional' properties, as I use the term: although physical$_1$ properties in a wide sense, they are second-order physical$_1$ properties.[19]

Notice that the functional properties indicated by Putnam in example (b) here are most naturally defined by Ramsey constants formed from formulas $A(T_1, \ldots, T_n,\ O_1, \ldots,\ O_m)$, where $n \geqslant 2$. However, Putnam states the functional properties indicated in examples (a) and (b) are *crude*, implying thereby that there are far more sophisticated kinds of functional properties and, hence, far more sophisticated kinds of functional definitions. Nevertheless, he does give a *necessary condition* for functional properties, namely, they (and hence, their definitions) must be *second-order*. Now let us assume that my explication of the notion of an ordinary explicit definition does not let in unwanted cases – and surely it does not. It follows that my explication of the notion of a functional definition (i.e., second-order non-ordinary explicit definitions) is guanranteed to cover all definitions which Putnam would deem to be functional.

In view of the foregoing, we may feel confident that the suggested explication of the notion of a functional definition is broad enough. In fact, the only plausible objection to this explication is that it might be too broad. However, this feature cannot in any way prejudice the case against functionalism. Indeed, without weakening the case for the negative thesis of functionalism, it actually strengthens the case for the positive thesis.

3. ASSESSMENT OF FUNCTIONAL DEFINITIONS

Using the suggested explications, I will now assess the adequacy and need for functional definitions. My conclusions will then be used in an assessment of the two theses of functionalism themselves. I begin by examining the adequacy of those functional definitions which are formed from whole theories A by the method of Ramsey constants, the method of Carnap constants and the method of Lewis constants. It should be noted that what I have to say about such functional definitions will be entirely general and will, therefore, apply to the three methods themselves. In this assessment we need not require of adequate definitions that the definiendum and the definiens be actual synonyms. If these functional definitions were to satisfy some weaker notion of definability, that would be impressive enough. The primary weaker notions of definability are (a) logical equivalence and (b) material equivalence. I will consider each of these notions in turn.

a. Logical Equivalence.

When I speak of logical equivalence, I will mean provable equivalence relative to a given theory. Thus, relative to this standard of definability, Q_i will be an adequate definition of T_i relative to the theory A if and only if

$$A \vdash T_i(x, y) \equiv_{xy} Q_i(x, y)$$

That is, Q_i is an adequate definition of T_i relative to A if and only if Q_i and T_i can be proven in A to be equivalent.

Notice that, relative to the theory $A(T_1, \ldots, T_n, O_1, \ldots O_m)$, if Q_i is a provably adequate definition of T_i, then Q_i and T_i have the same observational import (or alternatively, physical import). That is, if

$$A(T_1, \ldots, T_n, O_1, \ldots, O_m) \vdash T_i(x, y) \equiv_{xy} Q_i(x, y)$$

then, for all sentences $\mathfrak{D}$ whose non-logical constants are selected from $O_1, \ldots, O_m$,

$$A \vdash \mathfrak{C} \text{ iff } A' \vdash \mathfrak{C}$$

where A' results from A by substituting Q_i for T_i throughout. Thus, one way of testing whether a definition is provably adequate is to determine whether the observational (physical) import of the definiendum and the definiens are the same. To this end, I will begin my assessment by determining whether, in general, the observational (physical) import of a theoretical (mental) predicate T_i coincides with the observational (physical) import of an associated Ramsey constant T_i^* formed from a first-order theory A. Let the theory A^* result from A by substituting (for each i, $1 \leqslant i \leqslant n$) T_i^* for T_i. My question is this:

$$A \vdash \mathfrak{C} \text{ iff } A^* \vdash \mathfrak{C}?$$

It is true that, for any A and any $\mathfrak{C}$, if $A \vdash \mathfrak{C}$ then $A^* \vdash \mathfrak{C}$. To see this note that the Ramsey sentence, $(\exists F_1) \ldots (\exists F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_n)]$, is derivable from A^* by existential generalization; as we previously noted, A and this Ramsey sentence yield the same O-sentences as theorems. Despite this, the converse is false. That is, it is false that, for any A and any $\mathfrak{C}$, if $A^* \vdash \mathfrak{C}$, then $A \vdash \mathfrak{C}$. Thus, the answer to our question is negative. Consider an example. Let A be the miniature theory whose non-logical axioms are:

$$(\forall x, y)(T_1(x,y) \supset O_1(x,y))$$

$$(\forall x, y)(T_2(x,y) \supset O_1(x,y))$$

$$-(\forall x, y)(T_1(x,y) \equiv T_2(x,y)).$$

Note that the sentence $(T_1^*(x,y) \equiv T_2^*(x,y))$ is provable in first-order logic. Therefore, since $A^* \vdash -(\forall x)(\forall y)(T_1^*(x,y) \equiv T_2^*(x,y))$, A^* is inconsistent. Thus, for *every* $\mathfrak{C}$, $A^* \vdash \mathfrak{C}$ But since A is consistent, it is not the case that, for every $\mathfrak{C}$, $A \vdash \mathfrak{C}$. Therefore, relative to a given theory A, the observational (physical) import of T_i is not, in general, the same as the observational (physical) import of the Ramsey constant T *.

The same example can also be used to show that the analogous conclusion holds for Carnap constants T_i^{**} and Lewis constants T_i^{***}.

Since Ramsey constants, Carnap constants and Lewis constants do not, in general, preserve observational (physical) import, they do not, in general, make provably adequate definitions. That is, none of the following holds, generally:

$$A \vdash T_i(x,y) \equiv_{xy} T_i^*(x,y)$$

$$A \vdash T_i(x,y) \equiv_{xy} T_i^{**}(x,y)$$

$$A \vdash T_i(x,y) \equiv_{xy} T_i^{***}(x,y).$$

There are, however, certain theories A such that T_i and T_i^* do have the same observational (physical) import relative to A. This fact gives raise to the following question: For such theories A, must T_i and T_i^* be provably equivalent relative to A? Again, the answer is negative.

To see what the problem is here, consider the tiny theory A whose only non-logical axiom is:

$$(\forall x,y)(T_1(x,y) \supset O_1(x,y)).$$

The only purely observational sentences (or alternatively, physical sentences) which are theorems of A or A^* are *logical truths*. Thus, A and A^* have the same provable observational (physical) import. Nevertheless, T_1 and T_1^* cannot be proven to have the same extension relative to A. For, whereas T_1 could be any single relation which is included in O_1, T_1^* – i.e.,

$$(\exists F)[(\forall u,v)(F(u,v) \supset O_1(u,v)) \,\&\, F(x,y)]$$

is the union[20] of all relations included in O_1. Hence, T_1^* is identical in extension to O_1 itself. Since T_1 might be *properly included* in O_1 (and since first-order quantification theory is sound) T_1 and T_1^* cannot be proven to be equivalent in A.

The above example can also be used to show that, even if T_i and the Carnap constant T_i^{**} – and T_i and the Lewis constant T_i^{***} – have the same provable observational (physical) import, they cannot, in general, be proven to have the same extension relative to A.[21]

Which theories A are such that, for each i, $1 \leqslant i \leqslant n$, T_i and the Ramsey constant $T_i{}^*$ are provably equivalent? The answer is this: all and only those theories A which are such that, for any universe of discourse and any extensional interpretation of $O_1, \ldots, O_m$, there is *at most one* extensional interpretation of each $T_1, \ldots, T_n$ such that A comes out true, i.e., all and only those theories A such that:

$$(*) \qquad A(F_1, \cdots, F_n, O_1, \cdots, O_m) \,\&\, A(G_1, \cdots, G_n, O_1, \cdots, O_m)$$
$$\models (F_1(x,y) \equiv_{xy} G_1(x,y)) \,\&\, \cdots \,\&\, (F_n(x,y) \equiv_{xy} G_n(x,y))$$

Why must *all* theories which satisfy condition $(*)$ be such that T_i and T_i^*, $1 \leqslant i \leqslant n$, can be proven in A to be equivalent?

Consider the union of all relations F_i which, for some $F_1, \ldots, F_{i-1}$, $F_{i+1}, \ldots, F_n$, make A true. Now suppose that A satisfies condition $(*)$, i.e., suppose that relative to any universe of discourse and any interpretation of $O_1, \ldots, O_m$, there is *at most one* interpretation of each $T_1, \ldots, T_n$ which makes A true. Then, if A is true, T_i and the above union of relations F_i must be identical. However, the extension of the Ramsey constant T_i^* – i.e., $(\exists F_1, \ldots, F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_m)\ \&\ F_i(x,y)]$ – is precisely this union of relations F_i.[22]

Now consider the claim that *only* those theories A which satisfy condition $(*)$ are such that T_i and T_i^*, $1 \leqslant i \leqslant n$, can be proven in A to be equivalent. This claim follows directly from Padoa's method[23] for proving undefinability:

> If, for some universe and some interpretation of $O_1, \ldots, O_m$, there is more than one interpretation of the predicate T which makes the theory $A(T, O_1, \ldots, O_m)$ true, then there does not exist a formula $Q(O_1, \ldots, O_m, x, y)$ which, given A, is logically equivalent to T.

By repeated applications of Padoa's method, we get the following:

> If, for some universe and some interpretation of $O_1, \ldots, O_m$, there is more than one interpretation of the predicate T_i, $1 \leqslant i \leqslant n$, which makes the theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$ true, then there does not exist a formula $Q_i(O_1, \ldots, O_m, x, y)$ which, given A, is logically equivalent to T_i.

By contraposition, we get:

> If, for some formula $Q_i(x,y)$ whose non-logical constants are selected from $O_1, \ldots, O_m$,
>
> $$A(F_1, \cdots, F_{i-1}, T_i, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models T_i(x,y) \equiv_{xy} Q_i(x,y),$$
>
> then
>
> $$A(F_1, \cdots, F_{i-1}, G, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m)\ \&\ A(F_1, \cdots, F_{i-1}, H, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models G(x,y) \equiv_{xy} H(x,y).[24]$$

Given soundness, the antecedent can be replaced with:

$$A(F_1, \cdots, F_{i-1}, T_i, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m)$$
$$\vdash T_i(x, y) \equiv_{xy} Q_i(x, y).$$

To get out claim – i.e., that only those theories A which satisfy condition $(*)$ are such that, for each i, $1 \leqslant i \leqslant n$, T_i and T_i^* are provably equivalent in A – simply substitute $T_i^*(x, y)$ for $Q_i(x, y)$.

By similar arguments it is easily shown that, for every i, $1 \leqslant i \leqslant n$, the Carnap constants T_i^{**} are also provably equivalent to T_i relative to a given theory A if and only if A satisfies condition $(*)$. It is also easy to show that the same thing holds for Lewis constants T_i^{***}. It follows, by the way, that, for every i, $1 \leqslant i \leqslant n$ the Ramsey constant T_i^*, the Carnap constant T_i^{**} and the Lewis constant T_i^{***} are provably equivalent relative to a given theory A if and only if A satisfies condition $(*)$.

Now consider E. W. Beth's theorem on definability in first-order theories:[25]

> Let A be a first-order theory whose predicates are
>
> $T, S_1, \cdots, S_k$ (where $k \geqslant 1$) such that
>
> $A(F, S_1, \cdots, S_k)$ & $A(G, S_1, \cdots, S_k) \vdash F(x, y) \equiv_{xy} G(x, y).$
>
> Then, there is a *first-order* formula $Q(x, y)$ all of whose predicates are selected from $S_1, \ldots, S_k$ such that
>
> $A(T, S_1, \ldots, S_n) \vdash T(x, y) \equiv_{xy} Q(x, y).$

That is, if A is a first-order theory such that $S_1, \ldots, S_k$ uniquely determine the extension of T, then there is a *first-order* formula $Q(x, y)$ containing at most $S_1, \ldots, S_k$ such that, relative to A, $Q(x, y)$ is provably equivalent to $T(x, y)$. Hence, for such theories A, T is definable in terms of $S_1, \ldots, S_k$ without recourse to higher-order quantifiers and, hence, without recourse to, e.g., Ramsey constants.

Now what about condition $(*)$? From Beth's theorem and the completeness of first-order predicate logic, it follows directly that each first-order theory A which satisfies condition $(*)$ is such that each predicate T_i, $1 \leqslant i \leqslant n$, is definable in terms of $O_1, \ldots, O_m$ without recourse to higher order quantifiers and, hence, without recourse to Ramsey constants.

With this conclusion in mind let us reconsider our previous question: which theories A are such that, for each theoretical (mental) predicate T_i, the extension of T_i must coincide with the extension of the associated Ramsey constant (i.e., with the associated functional definition) T_i^*? The answer is now clear. The functional definition based on the Ramsey constant T_i^* has the same extension as T_i for those and only those theories A for which T_i has a *first-order* definition – i.e., a type-0 ordinary explicit definition. Hence, given the suggested explication of the ordinary-explicit/functional distinction, functional definitions based on Ramsey constant T_i^* have the same extension as T_i for those and only those theories A for which T_i has an *ordinary explicit definition*, i.e., a non-functional definition. That is, the method of Ramsey constants and the associated functional definitions work for exactly those first-order extensional theories for which it is not *in principle* needed. Moreover, as a matter of practice, the way in which we would typically attempt to show that a given theory A satisfies condition $(*)$ – or the antecedent condition in Beth's theorem – would be to produce *explicit* first-order definitional equivalents. Seldom, if ever, would we do things the other way around. To the extent that this is so, the method of Ramsey constants and the associated functional definitions are *practically* superfluous as well.

Now we generalize on this conclusion. Suppose that A is a first-order theory such that, for each i, $1 \leqslant i \leqslant n$, the theoretical (mental) predicate T_i is functionally definable in some way or other (including, e.g., the method of Carnap constants or the method of Lewis constants). That is, suppose that A is such that, for each i, $1 \leqslant i \leqslant n$ there is a formula $Q_i(x, y)$ which satisfies our conditions for what counts as a functional definition and, further, that

$$A(T_1, \ldots, T_n, O_1, \ldots, O_m) \vdash T_i(x, y) \equiv_{xy} Q_i(x, y).$$

From this it follows via soundness and Padoa's method, that A satisfies condition $(*)$. But if A satisfies condition $(*)$, it follows via Beth's theorem that each theoretical (mental) predicate T_i in A has a first-order definition. Hence, T_i has an *ordinary explicit definition*.

Therefore, if the functionalist's distinction between functional definitions and ordinary explicit definitions can be explicated in the way suggested earlier, then the following conclusion is obtained. For any first-order theory, there exist provably adequate functional definitions of the theoretical (mental) predicates if and only if there also exist provably adequate ordinary explicit definitions. Thus, if logical (i.e., provable) equivalence is taken as the standard of definability intended by functionalists, then the positive thesis of functionalism is true *if and only if* the negative thesis is false. In this, functionalism is logically inconsistent.

b. Material Equivalence

Let $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ be an interpretation of the predicates $T_1, \ldots, T_n, O_1, \ldots, O_m$. $\mathscr{D}$, of course, is the universe of discourse. According to the standard of definability presently under consideration, $Q_i(x, y)$ defines T_i if and only if

$$T_i(x, y) \equiv_{xy} Q_i(x, y)$$

is a true sentence on interpretation $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$.

Relative to a given first-order theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$, are the Ramsey constants T_i^*, the Carnap constants T_i^{**}, or the Lewis constants T_i^{***} materially adequate definitions of the theoretical (mental) predicate T_i? The arguments given early in Section (3a) show that these definitions are not in general materially adequate. Given this fact, it is natural, therefore, to wonder under what conditions materially adequate functional definitions exist. In this connection, it is tempting to seek, for the present standard of definability, a result analogous to the result obtained in Section (3a). Of the results one might expect to find, the following is the strongest:

> For all interpretations $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ and for each predicate T_i, $1 \leqslant i \leqslant n$, T_i has relative to interpretation $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ a materially adequate functional definition in terms of $O_1, \ldots, O_m$ if and only if there is a first-order formula $Q_i(O_1, \ldots, O_m, x, y)$ – i.e., a type-0 ordinary explicit definition – which, relative to interpretation $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$, is a materially adequate definition of T_i.

This proposition is. of course, false. First-order number theory, e.g., provides readily accessible counterexamples. For example, it is easily proven that the *less-than* relation and the *addition* relation do not have materially adequate first-order definitions in terms of zero and the successor function.[26]

Nevertheless, there are interpreted first-order formulas $B(<, \mathrm{Add}^3, o, ')$ which implicitly define the less-than relation and the addition relation in terms of zero and the successor function. Thus, the Ramsey constants

$$(\exists F_1, F_2)[B(F_1, F_2, o, ') \,\&\, F_1(x,y)]$$

$$(\exists F_1, F_2)[B(F_1, F_2, o, ') \,\&\, F_2(x,y)]$$

are materially adequate functional definitions of the less-than relation and the addition relation, respectively. The associated Carnap and Lewis costants are also materially adequate functional definitions of these relations.

Despite the above false start, it will be noticed that $<$ is inductively definable in terms of o and $'$. Thus, $<$ has a materially adequate (second-order) type-1 ordinary explicit definition in terms of o and $'$. At the same time, Add^3 is inductively definable in terms of $<$, o, $_1$. Therefore, Add^3 has a materially adequate type-2 ordinary explicit definition in terms of o and $'$. Generalizing, one might be tempted to accept the following:

> for any interpretation $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ and for each predicate T_i, $1 \leqslant i \leqslant n$, T_i has relative to $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ a materially adequate functional definition in terms of $O_1, \ldots, O_m$ if and only if, for some k, T_i also has relative to $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ a materially adequate type-k ordinary explicit definition in terms of $O_1, \ldots, O_m$.

Again, this proposition is too strong.[27]

This failure, however, suggests that there might be a significant class of interpretations $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ for which the above proposition holds. This is indeed so. Let $\langle \mathscr{D}, \mathscr{T}_1, \ldots, \mathscr{T}_n, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$ be such that (a) the universe of discourse $\mathscr{D}$ is countable (i.e., finite or denumerable) and (b) some relation $\mathscr{R}$ which well-orders the countable domain $\mathscr{D}$ (i.e., which arranges the elements of $\mathscr{D}$ into an order: first, second, third, . . .) can be given an ordinary explicit definition in terms of $O_1, \ldots, O_m$ (i.e., for some k, a type-k definition in terms of $O_1, \ldots, O_m$). For terminological convenience, I will say of such interpretations themselves that they *well-order countable domains*. Likewise, if such an interpretation is the standard interpretation of a theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$, I will say of the theory A itself that it well-orders a countable domain. Our conclusion is the following:

> For any interpretation $\langle \mathcal{D}, \mathcal{T}_1, \ldots, \mathcal{T}_n, \mathcal{O}_1, \ldots, \mathcal{O}_m \rangle$ which well-orders a countable domain and for any T_i, $1 \leqslant i \leqslant n$, T_i, has, relative to $\langle \mathcal{D}, \mathcal{T}_1, \ldots, \mathcal{T}_n, \mathcal{O}_1, \ldots, \mathcal{O}_m \rangle$, a materially adequate functional definition in terms of $O_1, \ldots, O_m$ if and only if (for some k) T_i also has, relative to $\langle \mathcal{D}, \mathcal{T}_1, \ldots, \mathcal{T}_n, \mathcal{O}_1, \ldots, \mathcal{O}_m \rangle$, a materially adequate (type-k) ordinary explicit definition in terms of $O_1, \ldots, O_m$.

To see this, consider the 'if' part first. If ordinary explicit definitions Q_i exist, then *ad hoc* functional definitions of T_i can always be constructed. For example, given the several ordinary explicit definitions $Q_1, \ldots, Q_n$, T_i can be functionally defined as follows:

$$T_i(x, y) \equiv_{xy} (\exists F_1, \cdots, F_n)[(F_1(x, y) \equiv Q_1(x, y)) \mathbin{\&} \cdots \mathbin{\&}$$
$$(F_n(x, y) \equiv Q_n(x, y)) \mathbin{\&} F_i(x, y)]$$

For the 'only if' part, consider any given interpretation

$$\langle \mathcal{D}, \mathcal{T}_1, \ldots, \mathcal{T}_n, \mathcal{O}_1, \ldots, \mathcal{O}_m \rangle$$

which well-orders a countable domain. $\mathcal{D}$ is either finite or denumerable. Suppose $\mathcal{D}$ is finite, then clearly every relation $\mathfrak{R}_i$ on $\mathcal{D}$ has a first-order – hence, type-0 – ordinary explicit definition. Thus, if T_1 has a materially adequate functional definition, it has a materially adequate ordinary explicit definition. Thus, we must consider only the case in which $\mathcal{D}$ is denumerable.

The desired result follows directly from a proposition which is an adaptation of the Tree Theorem from Recursion Theory. The statement of this proposition requires a few preliminaries. Suppose that $O_1, \ldots, O_m$ are relations on the natural numbers and that all recursive functions have first-order explicit definitions in terms of $O_1, \ldots, O_m$. Let $A(O_1, \ldots, O_m)$ be an open-sentence containing at least one predicate quantifier. Every such formula A can be *normalized*, i.e., converted into one of the following 'alternating quantifier' forms:

(1) $$(\forall F_{i_k})(\exists F_{i_{k-1}})(\forall F_{i_{k-2}}) \cdots (\exists F_{i_3})(\forall F_{i_2})$$
$$(\exists F_{i_1})(\forall v_j)[B(O_1, \ldots, O_m)]$$

(2) $$(\exists F_{i_k})(\forall F_{i_{k-1}})(\exists F_{i_{k-2}}) \cdots (\forall F_{i_3})(\exists F_{i_2})$$
$$(\forall F_{i_1})(\exists v_j)[B(O_1, \ldots, O_m)]$$

where $k \geqslant 1$ and B contains no quantifiers. A relation is said to be Π_k^1 in $\mathcal{O}_1, \ldots, \mathcal{O}_m$ if it has a definition $A(O_1, \ldots, O_m)$ which, when normalized, has the form of (1). A relation is said to be Σ_k^1 in $O_1, \ldots, O_m$ if it has a definition which, when normalized, has the form of (2). Given these concepts, the previously mentioned proposition may now be stated:

PROPOSITION. For any $k \geqslant 1$, every relation which is Σ_k^1 in $\mathcal{O}_1, \ldots, \mathcal{O}_m$ or Π_k^1 in $\mathcal{O}_1, \ldots, \mathcal{O}_m$ has a type-k ordinary explicit definition in terms of $O_1, \ldots, O_m$.[28]

The recursive functions, of course, have first-order explicit definitions in terms of +, $\cdot$, $<$. Moreover, + and $\cdot$ have inductive-turned-direct definitions in terms of $<$. Suppose that $<$ – which well-orders the denumerable domain $\mathscr{D}$ of natural numbers – has an ordinary explicit definition in terms of $O_1, \ldots, O_m$. Then, for every predicate T_i, $1 \leqslant i \leqslant n$, which expresses a relation on the natural numbers, if T_i, has a functional definition in terms of $O_l, \ldots, O_m$, it also has an ordinary explicit definition in terms of $O_1, \ldots, O_m$. This conclusion is identical to the result we are attempting to prove except that it is confined to predicates which express relations on the natural numbers. However, nothing in the argument hinges on this fact. All that is required is (a) that the domain $\mathscr{D}$ is denumerable and (b) that some relation $\mathscr{R}$ – which well-orders $\mathscr{D}$ – has an ordinary explicit definition in terms of $O_1, \ldots, O_m$. Thus, the desired result holds too.

Our conclusion, then, is this: Granted it is not true, for every interpreted theory, that its theoretical (mental) predicates have materially adequate functional definitions if and only if they also have materially adequate ordinary explicit definitions. Nevertheless, for every interpreted theory which well-orders a countable domain, the theoretical (mental) predicates have materially adequate functional definitions if and only if they also have materially adequate ordinary explicit definitions. For such theories, therefore, the positive thesis of functionalism is true if and only if the negative thesis is false. Thus, if cognitive psychology is such a theory, functionalism turns out to be inconsistent.

This conclusion, of course, leaves us with the question of whether cognitive psychology well-orders a countable domain. This question will be addressed in Section (4b).

Two final observations are in order. We have seen that, for theories which well-order a countable domain, the method of functional definitions works for exactly those theories for which it is *in principle* not needed. It should be noted, moreover, that *as a matter of practice*, the material adequacy of candidate functional definitions is commonly checked by no other means that the construction of an equivalent ordinary explicit definition.[29] Second, for any structure $\langle \mathscr{D}, \mathscr{O}_1, \ldots, \mathscr{O}_m \rangle$, if $\mathscr{D}$ is infinite, then there are *uncountably* many relations $\mathscr{T}$ on $\mathscr{D}$ which do not have definitions – ordinary explicit or functional – given in terms of $O_1, \ldots, O_m$. Or put another way, the relations on $\mathscr{D}$ which are undefinable far outnumber the relations on $\mathscr{D}$ which are definable. Thus in particular, if cognitive psychology has a denumerable domain, there is at this stage of our discussion no assurance that there exist any of the functional definitions promised by functionalism.

4. TWO PHILOSOPHICAL PROPOSITIONS

a. The Foregoing Results Hold for Causal Languages

It will have been noticed that so far no mention has been made of *causality*. There is a temptation to think that the foregoing negative conclusions might fail to obtain if the first-order theory A were equipped to represent causal necessity. After all, the positive thesis of functionalism is the doctrine that (predicates which express) mental states or mental properties are definable in terms of how they function in (theories concerning) the typical psychophysical *causal* manifold.[30] Despite this temptation, however, each of our negative conclusions stands even with the addition to A of apparatus for representing causal necessity.

In this connection I will sketch two methods by which causal necessity can be represented, methods which sustain the foregoing conclusions. First, there are techniques for contextually defining within a first-order extensional language the operation of intensional abstraction.[31] 'That'-clauses can thereby be represented in a first-order extensional language. Consider a sample causal-necessity sentence:

It is causally necessary that Φ.

According to this method for representing causal necessity, this sentence is parsed as follows:

$$\underbrace{\text{It-is-causally-necessary}}\ \underbrace{\text{that-}\Phi\,.}$$

Where the property of being a causally necessary proposition is expressed by a 1-place non-mental predicate C^1, this sentence is represented as follows:

$$C^1\ (\text{that-}\Phi\,)$$

Since $\ulcorner$that-$\Phi\urcorner$ can be contextually defined in a first-order extensional language, the above sentence may be treated as an abbreviation for a longer first-order extensional sentence. Hence, all the foregoing negative conclusions concerning functionalism stand without modification.

For a second approach to the representation of causal necessity, suppose that a non-extensional causal necessity operator '$\boxdot$' and appropriate principles characterizing its logical behavior are added to A and that the semantics for the resulting theory $A_{\boxdot}$ is done along the lines of 'causally possible worlds.'

Except in those places where we relied on Beth's theorem or the Tree theorem, all our above reasoning goes through substantially unchanged. Concerning Beth's theorem, analogues have been proved for a wide variety of modal systems.[32] Although this is not the place to discuss the philosophical issue of which modal system(s) best represent the 'must'- and 'can'-of-causality in natural language, we should note that analogues of Beth's theorem do hold for a variety of reasonable candidate modal systems. From this fact we can see that it is quite likely that the associated negative conclusion holds for the relevant first-order theories with the non-extensional causal necessity operator. It is not unreasonable to expect that comparable results can also be obtained for analogues of the Tree Theorem.

b. Cognitive Psychology Well-Orders a Countable Domain

In Section (3b) we saw that, if cognitive psychology well-orders a countable domain, then there exist materially adequate functional definitions if and only if there also exist materially adequate ordinary explicit definitions. Is it true that cognitive psychology well-orders a countable domain? That is, on the standard interpretation of cognitive psychology, is the universe of discourse $\mathscr{D}$ countable? And is there a relation $\mathscr{R}$ which well-orders $\mathscr{D}$ such that $\mathscr{R}$ has an ordinary explicit definition in terms of the non-mental predicates of cognitive psychology?

Before I attempt to answer this question, a terminological note should be made. When I say that cognitive psychology well-orders countable domain, I of course do not mean that *every* formulation of cognitive psychology well-orders a countable domain. Rather, I mean simply that there exists at least *one* adequate formulation of cognitive psychology which well-orders a countable domain.[33]

I will now sketch my reasons for thinking that cognitive psychology does indeed well-order a countable domain. I will first consider the question of countability. Then, I will consider the well-ordering property.

Concerning the matter of countability, I will consider (i) the *subjects* of psychological relations, (ii) the *objects* of psychological relations, (iii) the representation of *continuous* psychological processes and (iv) the mathematics used by cognitive psychology.

The matter of *subjects* is relatively easy. Inasmuch as every subject studied by cognitive psychology has a unique position in a universal genealogical tree (consisting of all ancestors and descendants), the subjects of psychological relations are countable in number.

The matter of the *objects* of psychological relations is more complicated, for there is notorious disagreement concerning just what these objects are. For the present discussion I will make a simplifying assumption, namely, that cognitive psychology can be adequately formulated in such a way that the objects of our psychological relations can be identified with *linguistic entities*, namely, the well-formed expressions of a given canonical language (e.g., 'mentalese'). In the present context, at least, it is appropriate to make this assumption, for this assumption is espoused by the majority of functionalists themselves. I will leave it to the reader to consider the situation that would obtain if, in order to be adequate, each formulation of cognitive psychology is forced to identify the objects of psychological relations with platonic objects (e.g., propositions). Now, if indeed the objects of psychological relations can be identified with the well-formed expressions in a given canonical language, then of course they are countable in number.[34]

Next we come to the matter of *continuity*. Some cognitive psychological processes appear to be continuous rather than discrete. This fact suggests that cognitive psychology might be forced beyond the countable. I submit, however, that, if appropriately small units are chosen for the formulation of our cognitive psychological theories, such apparently continuous processes can be successfully characterized as discrete. In psychology continuity is a convenience, not a necessity.

Finally, we come to the matter of the *mathematics* required by cognitive psychology, notably, portions of probability theory and measure theory. It might be thought that these portions of mathematics carry with them an unavoidable commitment of the uncountable. There are two ways to avoid this outcome. The first is to show that the probability theory and measure theory required by psychology can be captured by an *approximate mathematics* which is formulable in a countable setting.

The second – and formally more pleasing – way to avoid the above outcome is to show that the requisite probability theory and measure theory can actually be constructed within a set theory which posits only countably many sets. Such a set theory does indeed exist. The countable set theory developed by Charles Chihara on the basis of a set theory first constructed by Hao Wang comfortably provides the probability theory and measure theory used by cognitive psychology.[35] This set theory can be given a first-order formulation[36] and, hence, does not upset the negative results obtained earlier.

From this fact an important corollary follows. Suppose that, contrary to the claim I made earlier, cognitive psychology cannot avoid positing continuous (as opposed to discrete) psychological processes. Given that the measure theory used by cognitive psychology can be constructed in a countable setting, such continuous processes need not lead us beyond the countable. With this conclusion in hand, we may safely conclude that cognitive psychology does indeed have a countable domain.

We come now to the question of whether there is a relation $\mathscr{R}$ which well-orders the domain $\mathscr{D}$ of cognitive psychology such that $\mathscr{R}$ has an ordinary explicit definition in terms of the non-mental predicates of cognitive psychology. To see that such a relation exists, consider the following. First, 'nodes' in the previously mentioned universal genealogical tree are well-ordered by various relations which are inductively definable in terms of purely biological relations. Second, the well-formed expressions in a formal language are well-ordered by various relations which are inductively definable in terms of purely syntactic relations. Third, the sets in the Chihara-Wang countable set theory are well-ordered by various relations which are inductively definable in terms of the primitive vocabulary of that set theory.[37] Fourth, in terms of these relations it is possible, in turn, to define relations which well-order the special objects (e.g., times) essential to the representation of continuous psychological processes (assuming that the latter indeed exist).

Given the foregoing definitions, it is easy to then construct an inductive-turned-direct definition of a non-psychological relation which well-orders the union of (i) the set of subjects of psychological relations, (ii) the set of objects of psychological relations, (iii) the set of mathematical objects required by psychology and (iv) the set of special objects essential to the representation of continuous psychological processes. These four classes of objects, however, constitute the entire domain of any adequate formulation of cognitive psychology. Thus, I conclude that there is an adequate formulation of cognitive psychology which well-orders a countable domain.

5. CONCLUSIONS

In Section (3a) we saw that there exist provably adequate functional definitions of mental predicates if and only if there also exist provably adequate ordinary explicit definitions. In Section (3b) we saw that, if cognitive psychology well-orders a countable domain, the analogous result holds for the case of materially adequate definitions.

In Section (4b) I argued that cognitive psychology does indeed well-order a countable domain. Given that conclusion, it is clear that on both standards of definability – provable adequacy and material adequacy – functional definitions of mental predicates exist if and only if ordinary explicit definitions also exist. Thus, on both standards of definability, the positive thesis of functionalism is true if and only if the negative thesis is false. On both standards of definability, therefore, functionalism is inconsistent.

We have also seen that there is in general no guarantee that there exist either provably adequate or materially adequate functional definitions of the standard mental predicates. Indeed, the relations on the domain of cognitive psychology which fail to have provably adequate or materially adequate functional definitions actually *outnumber* those which do have such functional definitions. Clearly, the burden of proof for the existence of functional definitions rests with the proponents of the positive thesis of functionalism. However, to a surprising extent, proponents of functionalism seem to be oblivious to this requirement. Instead, the existence of functional definitions – whether provably or materially adequate – is more like an article of faith for a majority of the functionalists.

And how might functionalists set out to show that either provably or materially adequate functional definitions exist? Our primary conclusion shows that the existence of adequate functional definitions is not one bit more likely than the existence of ordinary explicit (behavioristic and/or physiological) definitions. This, moreover, is not just an *in principle* conclusion; it bears on *practice* as well. The primary strategy by which functionalists might attempt to meet the above requirement is to show that there exist precisely those things whose existence they deny, i.e., provably adequate or materially adequate ordinary explicit definitions. Ironically, the very proofs used in Sections (3a) and (3b) actually suggest directions for research into this matter.

In my closing remarks I will apply the foregoing conclusions to the consideration of the question: do there in fact exist either provably adequate or materially adequate functional definitions of the standard mental predicates? I do not propose to answer this question categorically. Instead, I will sketch a position which merits further study. I do not wish to be committed to this position.

Let us first consider the case where provable adequacy is taken as the standard of definability.

For our purposes, first-order cognitive psychologies can be separated into two epistemological types: (a) those which psychologists might in fact construct – and be epistemologically justified in so doing – on the basis of data assembled in the course of actual psychological research and (b) those which psychologists in fact would be unable to so construct and epistemologically justify.

An example of a type-(b) theory will help to clarify this distinction. There are uses of the terms 'belief' and 'desire' according to which each normal adult human being has an infinite number of beliefs and desires.[38] How does this come about? A plausible answer goes as follows. Each human being has a *finite* number of basic beliefs and desires, which either he possesses innately or intuitively or he acquires in perception or introspection. All remaining beliefs and desires are 'derived' in a rational (or nearly rational) way from these basic beliefs and desires or from other derived beliefs and desires. On the assumption that the conscious creatures (past, present, and future) are finite in number, there exists a true, finitely statable cognitive psychology *B* whose axioms include specifications of each of the basic beliefs and desires of each conscious creature (past, present, and future). If *B* also includes a suitable representation of the processes by which each conscious creature obtains its derived beliefs and desires, then it is conceivable that, relative to *B*, belief and desire do have provably adequate functional definitions (and, hence, provably adequate ordinary explicit definitions as well). *B* however, is a type-(b) theory. To see this, note that it is in fact impossible for any of us to know all of his own future basic beliefs and desires, not to mention (i) the basic beliefs and desires of other conscious creatures or (ii) the past or the future variations in the processes by which future derived beliefs and desires are obtained by ourselves or other conscious creatures.

If type-(b) theories are the only true theories which validate the positive thesis of functionalism, and physicalism itself, then these doctrines cease to be of much interest. The reasons for this are quite the same as the reasons why, e.g., phenomenalism has ceased to be of much interest. Clearly, it is of crucial importance to the positive thesis of functionalism – and to physicalism itself – that there exist appropriate type-(a) theories. Do any exist?

For the moment let us suppose that such a theory – call it *A* – exists. More specifically, let *A* be a true type-(a) theory such that, for each mental predicate, there exists a functional definition which, relative to *A*, is a provably adequate definition. In view of the conclusion reached in Section (3a), it follows that, for each mental predicate, there also exists a *first-order* formula which, relative to *A*, is a provably adequate definition. Now consider the arguments given by functionalists to show that there do not exist ordinary explicit definitions of belief and desire.[39] These arguments, it will be recalled, turn on phenomena such as (i) the fact that belief and desire interact with each other to produce physical outputs and (ii) the fact that physical inputs give rise to new beliefs and desires only by virtue of interaction with old beliefs and desires. It is possible that these arguments, together with sophisticated variations, can be made to yield the conclusion that there do not exist any first-order physicalistic formulas[40] which, relative to *A* (or any other true type-(a) theory), are logically equivalent to 'believe' and 'desire'. Let us assume that this is indeed so. From this it follows, via the fact that there is functional definability only if there is ordinary explicit definability, that there are standard mental predicates which do not, relative to A, have provably adequate functional definitions (or indeed any provably adequate physicalistic definitions whatsoever). If the above assumption holds, therefore, we arrive at the following conclusion: at best only type-(b) theories validate that version of the positive thesis of functionalism – and for that matter, physicalism itself – which takes provable equivalence to be the standard of definability; in this, this version of the positive thesis of functionalism – and physicalism – cease to be of interest.

A similar situation holds for the case in which material adequacy is adopted as the standard of definability. According to one of the major conclusions reached earlier, if the standard mental predicates have materially adequate functional definitions, they also have materially adequate ordinary explicit definitions. For our purposes, ordinary explicit definitions can be separated into two epistemological types: (a) those which, if they were materially adequate, we could in fact *know* to be materially adequate and (b) those which, if they were materially adequate, we could *not* in fact know to be materially adequate. Consider again the arguments used by functionalists to show that there do not exist any adequate ordinary explicit definitions of belief and desire. It is possible that these arguments, together with sophisticated variations, could be made to yield the following conclusion: there do not exist any materially adequate (first-or second-order) ordinary explicit definitions of belief and desire which are of type-(a), i.e., which we could in fact know to be materially adequate.

Earlier in the present section it was indicated that the primary strategy by which one might attempt to show that there exist adequate functional definitions of mental predicates is to show that there exist adequate ordinary explicit definitions. It is equally true that the primary strategy of showing that a given functional definition is adequate involves the construction of an equivalent ordinary explicit definition. Consider the following proposition: the above strategy provides the *only* route by which we might in fact come to know of any given functional definition of belief or desire that it is materially adequate. For the purpose of discussion, let us assume that this proposition is true.

Concerning the existence of materially adequate functional definitions of the standard mental predicates, three situations are possible: (1) such definitions simply do not exist; (2) such definitions exist and it is in fact possible for us to identify them, i.e., to know what they are; (3) such definitions exist but it is in fact impossible for us to identify them. Recall the two propositions considered in the two preceding paragraphs, respectively: (i) there do not exist any type-(a) materially adequate ordinary explicit definitions of belief and desire (i.e., materially adequate ordinary explicit definitions which we could in fact know to be materially adequate), and (ii) the actual construction of an equivalent ordinary explicit definition constitutes the only way by which we could in fact come to know of a given functional definition of belief or desire that it is materially adequate. If these two propositions are indeed true, it follows that possibility (2) is ruled out. Thus, if these two propositions are true, we get the following conclusion: *either* adequate (functional or ordinary explicit) physicalistic definitions of the standard mental predicates do not exist *or* such definitions exist but cannot be identified by us, i.e., cannot be known by us to be materially adequate. In this, the version of the positive thesis of functionalism which takes material adequacy as its standard of definability ceases to be of much interest. In similar fashion, the associated version of physicalism itself ceases to be of much interest.

So what is wrong with functionalism? First, its two theses are outright inconsistent. Second, insofar as the arguments for its negative thesis show at least that there are no epistemologically justifiable ordinary explicit definitions of the mental, then epistemologically justifiable functional definitions of the mental also fail to exist.

Reed College

NOTES

*I am indebted to Professors William Craig, John Addison and Charles Chihara and Mr. Mark Bedan for valuable discussions and to the Mellon Foundation and Reed College for financial assistance. Drafts of the paper were presented at the Northwest Conference on Philosophy, November 1976, and at the Reed College Philosophy Colloquium, February 1977.

[1] I hasten to add that if this form of physicalism is correct, it does not obviously follow that any interesting form of materialism is also correct. (See note 15) Several followers of functionalism, however, appear to have the opposite opinion.

[2] The functionalist's attack on naive physiological reductionism derives from the basic insight that, for most psychological states, there is an open-ended list of dissimilar physiological states which could give rise to that state, and there is an open-ended list of physiologically dissimilar states which could be causal effects of that state.

[3] Gilbert Harman, *Thought*, Princeton, New Jersey, 1973, p. 41.

[4] *Ibid.*, p. 34.

[5] *Ibid.*, pp. 44, 45. It should be noted, incidentally, that the historical claim is arguable.

[6] David Lewis, 'An Argument for the Identity Theory,' *Journal of Philosophy* 63, (1966), 17–25.

[7] Ibid., p. 21.

[8] *Ibid.*, pp. 19–20.

[9] It is, of course, understood here and throughout the discussion that the observational (physical) predicates $O_1, \ldots, O_m$ do not include copulas such as 'ϵ', the 'is'-of-predication, 'has-as-a-property', 'stand-in-the-relation', etc. As we will see, such copulas would make it possible to give functional definitions with first-order formulas. If such expressions occur in the theory A, they are to be treated neither as observational terms nor as theoretical terms but rather as *auxiliary parameters.* Cf., e.g., p. 364, S. C. Kleene, *Mathematical Logic*, New York, 1967. Despite this restriction, a good amount of the definitional power customarily afforded by such copulas is provided by second-order variables occurring in ordinary explicit definitions of type-i, $i > o$.

[10] For example,

$$(\forall F)\{[(\forall u, v, w)((v = 0 \;\&\; u = w) \supset F(u, v, w)) \;\&$$
$$(\forall u, v, w)\,(F(u, v, w) \supset F(u, v', w'))] \supset F(x, y, z)\}$$

is a type-1 definition (i.e., definiens) of the addition relation on the natural numbers given in terms of zero (0), identity (=) and the successor function ('). Let $Q(x, y, z)$ be shorthand for this type-1 definition of the addition relation. Then,

$$(\forall G)\{[(\forall u, v, w)((v = 0 \;\&\; w = 0) \supset G(u, v, w)) \&$$
$$(\forall u, v)(\exists w, z)(G(u, v, w) \;\&\; Q(w, u, z)) \supset Gu, v', z)] \supset G(x, y, z)\}$$

is a type-2 definition of the multiplication relation on the natural numbers given in terms of zero, identity and the successor function.

[11] 'Theories,' F. P. Ramsey, *The Foundations of Mathematics*, London, 1931, pp. 212–236.

[12] R. M. Martin, 'On Theoretical Constants and Ramsey Constants,' *Philosophy of Science* 31 (1966), 1–13.

[13] To see this, suppose first that

$$(\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m)] \vdash Q$$

By the deduction theorem,

$$\vdash (\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m)] \supset Q$$

By quantification theory,

$$\vdash (\forall F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \supset Q]$$

By universal instantiation,

$$\vdash A(T_1, \cdots, T_n, O_1, \cdots, O_m) \supset Q$$

Finally, by *modus ponens*,

$$A(T_1, \cdots, T_n, O_1, \cdots, O_m) \vdash Q$$

For the converse, suppose

$$A(T_1, \cdots, T_n, O_1, \cdots, O_m) \vdash Q$$

By the deduction theorem,

$$\vdash A(T_1, \cdots, T_n, O_1, \cdots, O_m) \supset Q$$

By universal generalization on $T_1, \ldots, T_n$,

$$\vdash (\forall F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \supset Q]$$

But since Q contains no occurrences of $F_1, \ldots, F_n$, we get

$$\vdash (\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m)] \supset Q$$

by quantification theory. Finally, by *modus ponens*,

$$(\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m)] \vdash Q.$$

[14] See § 2, H. P. Grice, 'Method in Philosophical Psychology,' *Proceedings and Addresses of the American Philosophical Association*, Newark, New Jersey, 1975, pp. 23–53. Incidentally, it should be noted that in *Thought* Harman does *not* provide a precise method for giving functional definitions. Even though in the section 'A modified Ramsey method' (pp. 41–43) a logically sound method for *eliminating* mental predicates is given, no method for *defining* mental predicates is given. It is worth noting, however, that the 'Ramsey constants' associated with this modified Ramsey method are subject to the same sort of criticisms offered below.

[15] For a concrete ẽxample, see Grice, *op. cit.*, pp. 32–36. Incidentally, it should now be clear why the form of physicalism is entailed by functionalism – i.e., that form of physicalism which asserts that science can be expressed in an exclusively physicalistic vocabulary – does not clearly entail an interesting version of materialism. According to the above functional definition, if x believes y, there exist relations B and D which satisfy $A(B, D, O_1, O_2)$. However, nothing at all has been said to indicate that the values of B and D are *physical relations*.

[16] To understand the motivation for the term 'Carnap constant,' see the discussion of Carnap's treatment of theoretical terms in D. Lewis, 'How to Define Theoretical Terms,' *Journal of Philosophy* 67 (1970), 427–466. Note the similarity between this technique for turning first-order inductive definitions into second-order ordinary explicit definitions. The difference lies in the form of the 'kernal,' i.e., the matrix.

[17] The motivation for the term 'Lewis constant' comes from David Lewis, *ibid.* Lewis' method makes use of a system of logic, designed by Dana Scott ('Existence and Description in Formal Logic,' in Schoenman, R. (ed.), *Bertrand Russell: Philosopher of the Century*, London, 1967), which admits vacuous names and vacuous descriptions such that:

(1) if α and β are both vacuous, $\alpha = \beta$ is true;

(2) if some α_i is vacuous, the atomic formula $R^k(\alpha_1, \ldots, \alpha_k)$ may either be true or false, depending on the chosen method of interpretation.

To state his method, Lewis first converts the first-order theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$ into the first-order theory $A_L(t_1, \ldots, t_n, o_1, \ldots, o_m)$ whose theoretical and observational expressions are the *singular terms* $t_1, \ldots, t_n$ and $o_1, \ldots, o_m$, respectively, and whose only predicates are 'copulas,' e.g., 'stand-in-relation'. In particular, the atomic formula $\ulcorner T_i(\alpha, \beta)\urcorner$ of A is converted into the atomic formula $\ulcorner \alpha, \beta$ stand-in-relation $t_i\urcorner$ of A_L and the atomic formula $\ulcorner O_i(\alpha, \beta)\urcorner$ of A is converted into the atomic formula $\ulcorner \alpha, \beta$ stand-in-relation $o_i\urcorner$ of A_L. Now Lewis' method advances the following definitions:

$$t_i =_{\text{def}} (\iota x_i)(\exists 1 x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)[A_L(x_1, \cdots, x_n, o_1, \cdots, o_n)]$$

:or any i, $1 \leqslant i \leqslant n$.

Such definitions are clearly materially adequate if, relative to the standard interpretation of its o-terms, A_L is uniquely realized. Therefore, these definitions could run aground materially only if, relative to the standard interpretation of its o-terms, A_L is not uniquely realized. Lewis sidesteps this problem with a *philosophical theory*: if, relative to the standard interpretation of its o-terms, A_L is not uniquely realized, then A_L is false – particular, the t-terms of A_L are all vacuous and formulas of the form $\ulcorner \alpha, \beta$ stand-in-relation $t_i\urcorner$ are false for all assignments to α and β. Given this (rather radical) philosophical theory and given the fact that in Scott's system $\ulcorner \alpha = \beta\urcorner$ is true when α and β are both vacuous, the definitions can be seen to hold in all cases.

Lewis' presentation leaves the reader with the mistaken impression that his method is dependent for its statement upon Scott's system. Let the theoretical and observational expressions of A all be *predicates* – surely first-order theories A can always be formulated so that they are. Then, when Lewis' definitions are converted back into formulas of A, we obtain what I am calling *Lewis constants*:

$$T_i(x, y) \; \mathit{iff}_{\text{df}}(\exists 1 F_1, \ldots, F_n)[A(F_1, \ldots, F_n, O_1, \ldots, O_m) \;\&\; F_i(x, y)].$$

Consider the *predicate*-analogue of Lewis' radical philosophical theory: if relative to the standard interpretation of its O-predicates theory A is not uniquely realized, then the extension of each T-predicate of A is null. Given this philosophical theory, our Lewis constants – like Lewis' original definitions – provide materially adequate definitions. In this way, then, Lewis' method is not dependent on Scott's system.

It should be noted that, even if Lewis' philosophical theory were correct, functional definitions of mental predicates which are based on Lewis constants would still be subject to all the negative conclusions to be reached in our discussion. The reason for this is that mental predicates are not 'theoretical' in the sense covered by Lewis' philosophical theory. On the theory, theoretical terms are simply *new* terms. Mental predicates are for the most part very *old*.

Incidentally, Grice (op. cit.) has proposed a second method for defining mental predicates. The definitions resulting from this method appear to be what I am calling Lewis constants.

[18] Accordingly, we might arrive at the following explication of the notion of a functional definition:

> A formula Q qualifies syntactically as a *functional definition* of the theoretical (mental) predicate T_i relative to the predicates $O_1, \ldots, O_m$ if and only if
> (1) Q is identical to – or provably equivalent to – a Ramsey constant T_i^*, a Carnap constant T_i^{**}, or a Lewis constant T_i^{***} formed from some first-order formula $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$,
> (2) $n \geqslant 2$,

(3) Q is not an ordinary explicit definition of T_i given in terms of $O_1, \ldots, O_m$.

Even if this explication were adopted, however, my major criticism of functionalism – i.e., that the positive thesis is true only if the negative thesis is false – would still hold.

Incidentally, there is yet another explication of the notion of a functional definition according to which *whole theories* $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$, where $n \geqslant 2$, are counted as functional definitions of the theoretical (mental) predicate $T_1, \ldots, T_n$. However, my major criticism of functionalism survives even on this explication.

[19] p. 244, Hilary Putnam, 'On Properties,' in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Dordrecht, Holland, 1970, pp. 235–254.

[20] This point is due to William Craig.

[21] Both T_i^{**} and T_i^{***} must be the null-relation; however, T_i clearly can be a non-null relation. Thus extension need not be preserved. Notice, incidentally, that 'intersection' takes the place of 'union' in the characterization of the extension of T_i^{**} and T_i^{***}.

[22] The following is a formal proof of the claim thàt all theories which satisfy condition $(*)$ are such that T_i and T_i^* are provably equivalent in A. By completeness of first-order quantifier logic, if A satisfies condition $(*)$, then

$$A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} A(G_1, \cdots, G_n, O_1, \cdots, O_m)$$
$$\vdash (F_1(x,y) \equiv_{xy} G_1(x,y)) \mathbin{\&} \cdots \mathbin{\&} (F_n(x,y) \equiv_{xy} G_n(x,y))$$

By the deduction theorem,

$$\vdash [A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} A(G_1, \cdots, G_n, O_1, \cdots, O_m)]$$
$$\supset [(F_1(x,y) \equiv_{xy} G_1(x,y)) \mathbin{\&} \cdots \mathbin{\&} (F_n(x,y) \equiv_{xy} G_n(x,y))]$$

Hence, by propositional calculus,

$$\vdash A(G_1, \cdots, G_n, O_1, \cdots, O_m) \supset$$
$$([A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} F_i(x,y)] \supset G_i(x,y))$$

By universal generalization of $F_1, \ldots, F_n$,

$$\vdash (\forall F_1, \cdots, F_n)(A(G_1, \cdots, G_n, O_1, \cdots, O_m) \supset$$
$$([A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} F_i(x,y)] \supset G_i(x,y))$$

By quantification theory,

(i) $$\vdash A(G_1, \cdots, G_n, O_1, \cdots, O_m) \supset$$
$$((\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} F_i(x,y)] \supset_{xy}$$
$$G_i(x,y))$$

For the converse of the consequent of (i), we use

$$\vdash A(G_1, \cdots, G_n, O_1, \cdots, O_m) \supset (G_i(x,y) \supset [A(G_1, \cdots, G_n, O_1, \cdots, O_m$$
$$\mathbin{\&} G_i(x,y)])$$

from the propositional calculus. By existential generalization,

(ii) $$\vdash A(G_1, \cdots, G_n, O_1, \cdots, O_m) \supset (G_i(x,y) \supset_{xy}$$
$$(\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \mathbin{\&} F_i(x,y)])$$

Substituting T_i for G_i in (i) and (ii), we get

$$\vdash A(T_1, \cdots, T_n, O_1, \cdots, O_m) \supset (T_i(x,y) \equiv_{xy} (\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \;\&\; F_i(x,y)])$$

By *modus ponens,*

$$A \vdash T_i(x,y) \equiv_{xy} (\exists F_1, \cdots, F_n)[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \;\& \; F_i(x,y)]$$

i.e.,

$$A \vdash T_i(x,y) \equiv_{xy} T_i^*(x,y).$$

[23] A. Padoa, 'Logical Introduction to any Deductive Theory,' translated and printed in part in J. van Heijenoort (ed.), *From Frege to Gödel,* Cambridge, Massachusetts, 1967. See, e.g., Kleene, *ibid.*, pp. 362–365.

[24] For a formal proof of this proposition, assume the antecedent:

$$A(F, \cdots, F_{i-1}, T_i, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models T_i(x,y) \equiv_{xy} Q_i(x,y).$$

By substituting G for T_i, we get

(i) $$A(F_1, \cdots, F_{i-1}, G, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models G(x,y) \equiv_{xy} Q_i(x,y).$$

And by substituting H for T_i, we get

(ii) $$A(F_1, \cdots, F_{i-1}, H, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models H(x,y) \equiv_{xy} Q_i(x,y).$$

From (i) and (ii) and the symmetry and transitivity of $\equiv$ we arrive at:

$$A(F_1, \cdots, F_{i-1}, G, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \;\& \; A(F_1, \cdots, F_{i-1}, H, F_{i+1}, \cdots, F_n, O_1, \cdots, O_m) \models G(x,y) \equiv_{xy} H(x,y)$$

[25] E. W. Beth, 'On Padoa's Method in the Theory of Definition,' *Koninklijke Nederlandse Akademie van Wetenschappen* 56 (1953), 330–337. The easiest known proof of Beth's Theorem is given by Craig (cf., e.g., Kleene, § 57, 'Beth's Theorem on Definability,' *Mathematical Logic*, New York, 1967). The proof begins with Craig's Interpolation Lemma:

> In first-order predicate calculus, if
> $\vdash B \supset C$ then there is a formula D whose predicates are common to B and C such that $\vdash B \supset D$ and $\vdash D \supset C$.

From the hypothesis of Beth's Theorem, the Deduction Theorem and the propositional calculus, we get:

$$\vdash (A(S, O_1, \cdots, O_m) \,\&\, S(x, y)) \supset$$
$$(A(T, O_1, \cdots, O_m) \supset T(x, y))$$

From this and Craig's Interpolation Theorem, we know there is a formula $D(O_1, \ldots, O_m, x, y)$ such that

(i) $\vdash (A(S, O_1, \cdots, O_m) \,\&\, S(x, y)) \supset D$

and

(ii) $\vdash D \supset (A(T, O_1, \cdots, O_m) \supset T(x, y))$

Since the proof of the formula in (ii) would go through if T were replaced throughout by S, we get

(iii) $\vdash D \supset (A(S, O_1, \cdots, O_m) \supset S(x, y))$.

From (i), (ii) and the propositional calculus we get

$$\vdash A(S, O_1, \cdots, O_m) \supset (S(x, y \equiv D).$$

By modus ponens, we get the consequent of Beth's Theorem.

[26] The proof uses quantifier elimination plus the fact that both the less-than relation and its complement are infinite and also the fact that both the addition relation and its complement are infinite. See, e.g., Enderton, § 3.1, 'Natural Numbers with Successor,' and 3.2, 'Other Reducts of Number Theory,' *A Mathematical Introduction to Logic*, New York, 1972.

[27] For example, there are certain infinite structures in which the notion of well-foundedness is (a) implicitly defined by a first-order formula and (b) not, for any k, explicitly defined by a type-k formula. Yiannis Moschovakis, example in conversation.

[28] The proof goes as follows: If a relation $\mathscr{P}$ is Σ^1_k in $\mathscr{C}_1, \ldots, \mathscr{C}_m$, it is the *complement* of some relation which is Π^1_k in $\mathscr{C}_1, \ldots, \mathscr{C}_m$. Therefore, it suffices to prove the proposition for the case of relations which are Π^1_k in $\mathscr{C}_1, \ldots, \mathscr{C}_m$. The proof is by induction on k. Suppose a relation $\underline{\mathscr{P}}$ is Π^1_{k+1} in $\mathscr{C}_1, \ldots, \mathscr{C}_m$. Then there is a relation $\mathscr{S}$ which is Σ^1_k such that $\mathscr{P}$ is Π^1_1 in $\mathscr{C}_1, \ldots, \mathscr{C}_m, \mathscr{S}$. But $\mathscr{S}$ is just the complement ofsome relations $\mathscr{S}'$ which is Π^1_k in $\mathscr{C}_1, \ldots, \mathscr{C}_m$. By the induction hypothesis $\mathscr{S}'$ has a type-k ordinary explicity definition in terms of $O_1, \ldots, O_m$. Thus, $\mathscr{S}$ has a type-k ordinary explicit definition in terms of $O_1, \ldots, O_m$. Therefore, all that must be shown is that if a relation $\mathscr{P}$ is Π^1_1 in $\mathscr{C}_1, \ldots, \mathscr{C}_m, \mathscr{S}$, it has a type-1 ordinary explicit definition in terms of $O_1, \ldots, O_m, \mathscr{S}$. However, this follows by an adaptation of the *Tree Theorem*. (For a proof of the Tree Theorem itself, see, e.g., Shoenfield, *Mathematical Logic*, Reading, Massachusetts, 1967, p. 180.) Let $\mathscr{U}$ be the arguments of $\mathscr{P}$. The Tree Theorem goes as follows:

> If $\mathscr{P}$ is Π^1_1 in any sequence of relations $\mathscr{M}_1, \ldots, \mathscr{M}_j$ there is a relation $V(x, \mathscr{U})$ which is recursive in $\mathscr{M}_1, \ldots, \mathscr{M}_j$ such that, for all $\mathscr{U}$,
>
> $\mathscr{P}(\mathscr{U})$ if and only if the set x's which satisfy $V(x, \mathscr{U})$ forms a tree.

What is a tree? Gödel developed a technique for assigning a unique natural number to each finite sequence of natural numbers $\langle a_1, \ldots, a_h \rangle$ for each $h \geqslant 1$. Such a number is called a *sequence number*. Consider any pair of finite sequences $\langle a_1, \ldots, a_h \rangle$, $\langle a_1, \ldots, a_h, b_1, \ldots, b_{h'}$. We say that the sequence number of $\langle a_1, \ldots, a_h \rangle$ precedes the sequence number of $\langle a_1, \ldots, a_h, b_1, \ldots b_{h'} \rangle$. This relation of *preceding* is recursive and, hence, has a first-order definition. Now consider any infinite sequence of finite sequences:

$$\langle a_1, \ldots, a_h \rangle, \langle a_1, \ldots, a_h, b_1, \ldots, b_{h'} \rangle, \langle a_1, \ldots, a_h, b_1, \ldots, b_{h'}, c_1, \ldots, c_{h''} \rangle, \ldots$$

Such a sequence is called an *infinitely descending sequence of finite sequences.* The set of sequence numbers associated with an infinitely descending sequence of finite sequences is called an *infinitely descending sequence*. A *tree* is a set of sequence numbers which does not include any infinitely descending sequence. Put graphically, trees are objects all of whose 'branches' are finite. Now, the right-hand side of the bi-conditional in the Tree Theorem is equivalent to:

> $(\forall_y)[V(y, \mathscr{U}) \supset y$ is not an element of any infinitely descending sequence which is a subset of the x's which satisfy $V(x, \mathscr{U})]$.

This expression will have a type-1 ordinary explicit definition in terms of $\mathscr{M}_1, \ldots, \mathscr{M}_j$ if its consequent has an inductive-turned-direct definition in terms of $\mathscr{U}_1, \ldots, \mathscr{M}_j$. However, the following is just such a definition:

$$\begin{aligned}&(\forall F)\{[(\forall x)(V(x, \mathscr{U}) \,\&\, -(\exists v)(V(v, \mathscr{U}) \,\&\, x \text{ precedes } v) \,.\!\supset F(x)) \,\&\\ &\quad (\forall x)(V(x, \mathscr{U}) \,\&\, (\forall v)((V(v, \mathscr{U}) \,\&\, x \text{ precedes } v) \supset F(v)) \,.\!\supset F(x))]\\ &\quad \supset F(y)\}.\end{aligned}$$

[29] In this connection, consider an interpreted first-order theory $A(T_1, \ldots, T_n, O_1, \ldots, O_m)$ which *implicitly defines* (cf., note 18) the predicates $T_1, \ldots, T_n$ in terms of $O_1, \ldots, O_m$, i.e.,

$$\begin{aligned}&(\forall F_1, \cdots, F_n, G_1, \cdots, G_n)\{[A(F_1, \cdots, F_n, O_1, \cdots, O_m) \,\&\\ &\quad A(G_1, \cdots, G_n, O_1, \cdots, O_m)] \supset [(F_1(x, y) \equiv_{xy} G_1(x, y)) \,\&\, \cdots \,\&\\ &\quad (F_n(x, y) \equiv_{xy} G_n(x, y))]\}.\end{aligned}$$

For such theories A, the Ramsey constants T_i^* – as well as the Carnap constants T_i^{**} and the Lewis constants T_i^{***} – are materially adequate functional definitions of T_i, $1 \leqslant i \leqslant n$. Suppose that A well-orders a countable domain. Then, by an adaptation of an analogue of Beth's Theorem, i.e., the *Souslin-Kleene Characterization Theorem* (see, e.g., Shoenfield, *ibid.*, § 7.10), we know that each T_i expresses a relation $\mathscr{J}_i$ which is hyperarithmetical in $\mathscr{O}_1, \ldots, \mathscr{O}_m$. That is, $\mathscr{J}_j$ is a member of the class $\mathscr{H}$ of relations, where $\mathscr{H}$ is inductively defined as follows: (i) if $\mathscr{M}$ is a relation which is recursive in $\mathscr{O}_1, \ldots, \mathscr{O}_m$, then $\mathscr{M}$ is in $\mathscr{H}$; (ii) if σ is an effectively determined sequence of relations which are in $\mathscr{H}$ and if $\mathscr{M}$ is either the union or intersection of the relations in σ, then $\mathscr{M}$ is in $\mathscr{H}$. However, every relation which is hyperarithmetical in $\mathscr{O}_1, \ldots, \mathscr{O}_m$ has a rather natural second-order ordinary explicit definition in terms of $O_1, \ldots, O_m$. In fact, this perspective provides a very useful way for determining whether $T_1, \ldots, T_n$ are indeed implicitly defined by the theory A.

For a discussion of analogues of Beth's Theorem in recursion theory and effective descriptive set theory, see John Addison, 'The Theory of Hierarchies,' in *Logic, Methodology and Philosophy of Science, Proceedings of the 1960 International Congress*, Amsterdam, 1961, pp. 26–37.

[30] In the method for defining mental predicates which is suggested by Grice (i.e., a method of Ramsey constants), no such emphasis is placed on causality. It must be noted, however, that in a later section of that paper, it is indicated that '*ceteris paribus*'-clauses must be prefixed to statements of psychological laws. Since the logic of the affected contexts is non-standard and since this logic has not been formalized, we cannot at this stage judge either way whether our negative conclusions hold for the functional definitions constructed out of psychological laws so-stated. Nevertheless, definitions in which such '*ceteris paribus*'-clauses occur most likely do *not* qualify as fully physicalistic, for such '*ceteris paribus*'-clauses themselves do not strictly qualify as physical-object expressions. The covert circularity here is analogous to that created by the 'normal observation conditions'-clauses sometimes used to 'rescue' phenomenalism. For a discussion of this and related points, see Roderick Chisholm, 'The Problem of Empiricism,' *Journal of Philosophy* 45 (1948), 512–517.

It should also be noted that Grice (*op. cit.* p. 46) suggestes a way to define belief in terms of desire. When the details of this definition are examined, however, it is seen that the definition is not in conflict with the negative thesis of functionalism. (Although necessary, a 'hierarchy of definitional priority' – see the passage quoted from Lewis on page 336, above – is not sufficient for the vindication of behaviorism.) In any event, the success of the proposed definition turns on the presence of a '*ceteris paribus*'-clause and, with it, covert circularity.

[31] See, e.g., 'The Thesis of Extensionality' in my book, *Properties, Relations, and Propositions*, D. Reidel Publishing Company; Dordrecht, Netherlands, forthcoming.

[32] For example, Gabbay, 'Craig's Interpolation Theorem for Modal Logics,' *Conference in Mathematical Logic – London '70*, Lecture Notes in Mathematics, No. 255, Springer, 1972, pp. 111–127. Bowen, K. A., 'Normal Modal Model Theory,' *Journal of Philosophical Logic* 4 (1975), 1–131. Incidentally, in 'The Interpolation Lemma Fails for Quantified S5' (photocopied), Kit Fine has recently shown that analogues of neither Craig's interpolation theorem nor Beth's definability theorem hold for S5 with a Kripke-style semantics. It is very unlikely that this result, which evidently contradicts one of Bowen's results on this topic (*op. cit*), invalidates our conclusion, for it is very unlikely that the causal analogue of S5 (the system in which '⊡' and '⟐' replace '□' and '◇') accurately represents the logical behavior of the 'must'- and 'can'-of causality. For example, 'if, for all x, x can lift his own weight, then it must be so that, for all x, x can lift his own weight' is, intuitively, invalid. Example due to C. David Reeve.

[33] If this condition is met and if the following conjecture is true, that will be enough to warrant the conclusion I will seek to draw in Section (5). Let A be an adequate formulation of cognitive psychology which well-orders a countable domain, and let A' be another adequate formulation of cognitive psychology which does *not* well-order a countable domain. Suppose that A' is such that the standard mental predicates are functionally definable in terms of its physical vocabulary. The conjecture I have in mind is this: if A and A' are as above, then the standard mental predicates are also functionally definable in terms of the physical vocabulary of A. This conjecture is extremely

plausible, for surely the functional definability of the standard mental predicates does not stand and fall with the indicated difference between A' and A.

It is worth noting, incidentally, that even if certain parts of science (e.g., physics) should require uncountably many objects, this fact would not cast doubt upon the proposition that there exists a wholly adequate formulation of cognitive psychology which has a countable domain. This situation would hold true even if the 'unity of science' doctrine proved to be correct. For the 'unity of science' picture does not prohibit a given science from restricting its domain of inquiry to something less than the entire domain of science.

[34] This is, of course, not to say that subjects of psychological relations do not stand in psychological relations to sentences the *truth* of which would require the existence of uncountably many objects. The determination of the truth of such sentences, however, does not lie within the province of psychology. The province of psychology includes only the description and explanation of our mental states and behavior. For more on the topic of propositions as the objects of psychological relations, see my book, *ibid.*

[35] Charles Chihara, Y. Lin and T. Schafter, 'A Formalization of a Nominalistic Set Theory,' *Journal of Philosophical Logic* 4 (1975), 155–170.

[36] For techniques, see, e.g., §37, in Quine, *The Logic of Set Theory*, Cambridge, Massachusetts, 1963.

It should be noted, incidentally, that the 'ϵ' of a first-order Chihara-Wang set theory may be added as a non-mental primitive predicate to the language of psychology without affecting my previous explication of the explicit/functional distinction (cf. note 9). The reason for this is that on its platonistic interpretation this 'ϵ' relation holds only between sets and sets, not between individuals and sets; the sets studied by this set theory are *purely mathematical.*

[37] Indeed, it is possible to inductively define relations which actually enumerate the sets in the Chihara-Wang set theory. See, Wang, 'The Formalization of Mathematics,' *Journal of Symbolic Logic* 19 (1954), 241–266.

[38] For a brief defense of this thesis, see, e.g., Stephen Schiffer, *Meaning*; London, 1972, p. 36.

[39] See Section (1) above.

[40] Appropriate restrictions are, of course, in effect; see note 9.