EXPLANATION REVISITED*

DAVID KAPLAN[1]

University of California, Los Angeles

In 'Hempel and Oppenheim on Explanation', (see preceding article) Eberle, Kaplan, and Montague criticize the analysis of explanation offered by Hempel and Oppenheim in their 'Studies in the Logic of Explanation'. These criticisms are shown to be related to the fact that Hempel and Oppenheim's analysis fails to satisfy simultaneously three newly proposed criteria of adequacy for any analysis of explanation. A new analysis is proposed which satisfies these criteria and thus is immune to the criticisms brought against the earlier analysis.

Hempel and Oppenheim propose in [2] an analysis of the relation which holds between a theory T and a singular sentence E when E is explainable by T.[2] They also propose certain criteria of adequacy for any such analysis. The purpose of the present paper is twofold. Part I contains three additional criteria which, in view of results in [1], are not at all satisfied by the analysis of Hempel and Oppenheim. Part II contains a proposal for a revised analysis which satisfies all criteria so far proposed.

I

Before discussing the newly proposed criteria we must mention certain preliminaries. Let us prefix "H-O" to "explanation" and its cognates "explainable" and "explanans" to designate the relations defined by Hempel and Oppenheim. Thus we shall call a singular sentence E, *H-O explainable by T* just in case there is a C such that (T, C) is an *H-O explanans for E*. A certain simplification of the notion of H-O explanation is afforded by the following theorem.

Theorem 1. An ordered couple (T, C) of sentences is an H-O explanans for a singular sentence E if and only if the following conditions are satisfied:

(1) T is a theory,

(2) C is singular and true,

(3) $\{T, C\} \vdash E$, and

(4) there is a class K of basic sentences such that $K \vdash C$ but neither $K \vdash E$ nor $K \vdash \sim T$.

Proof. Examination of Hempel and Oppenheim's definition of "explanans" as given, for example, in [1] reveals that it is sufficient to show that conditions

* Received, September 1960.

[1] This paper was prepared for publication while the author was a Graduate Fellow of the National Science Foundation.

[2] Throughout the paper the notation of [1] will be used. Notions from [2] such as "singular", "theory", etc. which are defined in [1] will not be redefined herein.

(1)-(4) imply that T is not equivalent to any singular sentence. Thus, assume (1)-(4). If T is a derivative theory, the conclusion follows by definition. Therefore, we may also assume,

(5) T is a fundamental theory.

In order to show that T is not equivalent to any singular sentence, suppose to the contrary that S is a singular sentence and

(6) $\vdash T \equiv S$.

Hence, with (5) and Lemma 3 of [1],

$\vdash T \supset L$,

where L is the universal closure of the result of replacing the distinct individual constants of S with distinct variables. Thus, with (6), $\vdash S \supset L$, which by Lemma 1 of [1] yields,

either $\vdash \sim S$ or $\vdash L$.

But the logical provability of $\sim S$ implies, by (6), the logical provability of $\sim T$, which is impossible in view of the truth of T. Hence,

$\vdash L$.

Therefore, by (6) and the fact that $\vdash L \supset S$, $\vdash T$. Thus, by (3),

(7) $\vdash C \supset E$.

This completes the proof, since (7) contradicts (4).

The characterization of H-O explanation given by Theorem 1 will be used without specific reference throughout the rest of the present paper.

It seems natural to require that the relation of explanation be closed under a certain restriction of the relation of logical derivability. Let us formulate this requirement in two parts.

R1. If a singular sentence is explainable by a given theory, then it is explainable by any theory from which the given theory is logically derivable.

R2. Any singular sentence which is logically derivable from singular sentences explainable by a theory is itself explainable by that theory.

Although the requirements R1 and R2 are intended to hold for every interpreted language, R3, below, is of a different character. To arrive at the third requirement, we first note that the existence of a fundamental theory T and a singular sentence E such that E is explainable by T is clearly dependent on both the expressive resources of the language L in which T and E are to be formulated and certain contingent matters concerning the interpretation of L.[3] However, it is reasonable to require of an analysis of explanation that there be some interpreted language containing theories by which

[3] For example, Let L contain only a single one-place predicate F, and let the interpretation of L take for the domain of discourse the class of all living beings and for the interpretation of F the property of being a man. L may also contain any number of individual constants. Then the only fundamental laws which can be formulated in L are logically true and hence have no explanatory power. That no contingently true fundamental laws can be formulated in L is shown by transforming the quantifier free part of any fundamental law of L into conjunctive normal form, distributing the quantifiers over the conjuncts, dropping vacuous quantifiers, and confining the remaining quantifiers within the disjunctions in such a way that no quantifier lies within the scope of another. It can then easily be established that in each conjunct any true disjunct is logically provable.

certain singular sentences are explainable and other singular sentences are not. Thus we are led to the following requirement:

R3. There is an interpreted language L which contains a fundamental theory T and singular sentences E and E' which are true but not logically provable such that E is explainable by T and E' is not explainable by T.[4]

We can now conclude from Theorems 4 and 5 of [1],

Lemma 1. Let L be any interpreted language with infinitely many individual constants. Then if H-O explanation satisfies R3 with respect to L, it satisfies neither R1 nor R2 (with respect to L).

The difficulties in H-O explanation are made more explicit by Theorems 2 and 3.

Theorem 2. There is a language L with infinitely many individual constants such that H-O explanation satisfies R3 with respect to L.

Proof. Let L be any language which contains the one-place predicates F and G and infinitely many individual constants including the individual constant a. Assume further that the interpretation of L is such that $(x) \sim Fx$ and Ga are true in L. There certainly are such languages. It is clear that $(x) \sim Fx$ is a fundamental theory and that Ga and $\sim Fa$ are singular sentences which are true but not logically provable. It is also clear that $\sim Fa$ is H-O explainable by $(x) \sim Fx$. In order to show that Ga is not H-O explainable by $(x) \sim Fx$ assume to the contrary that C and K are respectively a singular sentence and a class of basic sentences such that

(1) $\{(x) \sim Fx,\ C\} \vdash Ga$,
(2) $K \vdash C$,
(3) not $K \vdash Ga$,
(4) not $K \vdash \sim (x) \sim Fx$.

From (1) and (2) we have,

$$K \cup \{\sim Ga\} \vdash (\exists x)Fx.$$

Hence, using Hilbert's extension of the first ε-theorem,[5] there are a finite number of individual constants $b_1, \ldots, b_n$ such that

(5) $K \cup \{\sim Ga\} \vdash Fb_1 \vee \ldots \vee Fb_n$.

It now follows that

(6) there is an i $(1 \leqslant i \leqslant n)$ such that $Fb_i \,\varepsilon\, K$.

For assume the contrary. Then, by (3), $K \cup \{\sim Ga\} \cup \{\sim Fb_1, \ldots, \sim Fb_n\}$ is a class of basic sentences which contains no sentence and its negation, and

[4] It is perhaps of interest to note that there are languages which satisfy the simpler requirement of containing a fundamental theory by which some singular sentence is explainable, but which do not satisfy R3. Such a language L can be obtained from the language of the preceding note by revising the interpretation of F so that it becomes a universal predicate. Note that every contingently true singular sentence of L is logically derivable from, and hence explainable by, every contingently true fundamental law of L.

[5] As stated in [3] the theorem asserts that if an existential formula is logically derivable from a class of quantifier free formulas, then a disjunction of its instances is also so derivable.

hence is consistent. But this contradicts (5), thereby establishing (6). This completes the proof, since (6) and (4) are incompatible.

From Lemma 1 and Theorem 2 we infer,

Theorem 3. H-O explanation satisfies neither R1 nor R2.

In particular, as Lemma 1 states, R1 and R2 fail for every language of the kind described in Theorem 2.

II

In their discussion of self-explanation and predictive import, Hempel and Oppenheim point out that if a singular sentence E is to be explainable in a non-trivial sense on the basis of a theory T and a set of initial conditions C, then it must be possible to verify C without simultaneously either verifying E or falsifying T. In their conception, to verify a sentence S is to find a class of true basic sentences from which S is logically derivable. The requirement of possible verification is realized in the Hempel and Oppenheim analysis by a clause which requires the existence of a class K of basic sentences which is compatible with T and $\sim E$ and from which C is logically derivable. A requirement of this sort is certainly necessary to prevent trivial self-explanation, but in the form given above it is not yet strong enough. It must not be merely *possible* for the verifying class K to exist; there must be an *actual* verifying class. That is to say, the basic sentences of the class K must be true. The failure to make this stipulation leads directly to the difficulty embodied in the following theorem. Let us say that E is *possibly H-O explainable* by T if there is a sentence C such that (T, C) satisfies all the requirements for an H-O explanans of E with the possible exception of the requirement that C be true.

Theorem 4. If E is possibly H-O explainable by T, then E is H-O explainable by T.

Proof. Assume the hypothesis. Then there is a set of initial conditions C, such that if C were true, (T, C) would be an H-O explanans for E. To establish the conclusion of the theorem it suffices to show that $(T, C \vee E)$ is an H-O explanans for E. The reader can easily verify this assertion. In doing so he should note that the new explanation is objectionable in just those cases in which C is actually false, and that in these cases at least one element of K must also be false.[6]

There are other natural ways of arriving at the same requirement. For example, let the singular sentence C which represents the initial conditions be in disjunctive normal form. Then in order to avoid trivial self-explanation we might require that the explanation does not depend in an essential way on any disjunct of C from which E is logically derivable. That is to say, the explanation can be carried through where the initial conditions are represented by a sentence in disjunctive normal form none of whose disjuncts logically imply E.

[6] Note that the first counterexample of [1] illustrates the idea used in the proof of this theorem.

The following theorem shows the equivalence of the above two proposals and also gives a third more perspicuous formulation.

Theorem 5. Assume that T is a theory and E is a true singular sentence. Then the following conditions are equivalent.

(i) There is a singular sentence C'' such that E is logically derivable from the set $\{T, C''\}$, and there is a class K of *true* basic sentences such that C'' is logically derivable from K, and neither E nor $\sim T$ is logically derivable from K.

(ii) There is a sentence C' in disjunctive normal form such that E is not logically derivable from any disjunct of C' and (T, C') is an H-O explanans for E.

(iii) There is a conjunction C of *true* basic sentences such that E is logically derivable from the set $\{T, C\}$ and E is not logically derivable from C alone.

Proof. Assume the hypothesis. The three equivalences will be established by showing that (i) implies (iii), (iii) implies (ii), and (ii) implies (i).

To show that (*i*) implies (iii) assume (i). Since $K \vdash C''$, the Deduction Theorem assures us that there is a conjunction C^* of elements of K such that

(1) $\vdash C^* \supset C''$.

Now C^* is a conjunction of true basic sentences from which E is not logically derivable. For if it were, E would be logically derivable from K contrary to (i). From (i) we also have $\{T, C''\} \vdash E$, which with (1) implies,

$\{T, C^*\} \vdash E$.

Thus C^* satisfies (iii).

To show that (iii) implies (ii) assume (iii). Note that C is in disjunctive normal form and that E is not derivable from any disjunct of C, that is, from C itself. Let K^* be the class of basic conjuncts of C. Then $K^* \vdash C$ but E is not logically derivable from K^*. Further, $\sim T$ is not logically derivable from K^* since all elements of K^* are true, whereas $\sim T$ is false. Hence (T, C) is an H-O explanans for E. Thus C satisfies (ii).

To show that (ii) implies (i) assume (ii). Since C' must be true it has at least one true disjunct, C^*. Let K^* be the class of basic conjuncts of C^*. Now $K^* \vdash C^*$, but E is not logically derivable from K^* since E is not logically derivable from C^*. Further, $\sim T$ is not logically derivable from K^* since, as in the preceding argument, all elements of K^* are true but $\sim T$ is false. To show that C^* and K^* satisfy (i) it remains only to show that $\{T, C^*\} \vdash E$. But this follows from (ii), which implies that $\{T, C'\} \vdash E$, and the fact that $\vdash C^* \supset C'$.

In view of Theorem 5 we might consider revising the analysis of Hempel and Oppenheim to the following:

(T, C) is an explanans for the singular sentence E if and only if the following conditions are satisfied:

(1) T is a theory,

(2) C is a conjunction of true basic sentences,

(3) $\{T, C\} \vdash E$, and

(4) not $\{C\} \vdash E$.

This revision leads to a concept of explanation which satisfies R1. However, before settling on a final analysis let us also consider R2.

We must first point out that on careful examination R2 is not as plausible as R1. Consider, for example, the singular sentence *Ga*. Now suppose that *Fa* and $(x)(Fx \supset Gx)$ are true, then *Ga* is explainable by the theory $(x)(Fx \supset Gx)$. But what shall we say of the singular sentence *Fa* v *Ga*? Although it is a logical consequence of *Ga*, any attempt to explain it by means of the theory in question seems to involve the use of the initial condition *Fa* which alone implies the sentence to be explained. One may at this point simply claim that *Fa* v *Ga* should not be considered explainable on the basis of the given theory and thus reject R2. The author finds his intuition unclear in the above particular case, and thus prefers to preserve R2 in view of two considerations. First, it seems to have great general plausibility, and secondly, it appears to be the most natural way to protect an analysis of explanation against trivializations of the kind embodied in Theorem 5 of [1].

The example of the above paragraph suggests that the notion of explanation be introduced in a stepwise fashion. A certain restricted class of singular sentences will be said to be *directly explainable* by a theory under certain conditions. Then we will call an arbitrary singular sentence *explainable* by a theory if the sentence is a logical consequence of sentences which are directly explainable by the given theory. The class of sentences directly explainable by a theory must logically imply just those singular sentences intuitively explainable by the theory. Thus we wish to restrict this class so that not every sentence will be derivable from it. This can be done by taking means to prohibit a kind of trivial explanation closely related to what Hempel and Oppenheim call partial self-explanation.[7]

Assume that *E*, the singular sentence to be explained, is in conjunctive normal form. If in an attempt to directly explain *E* we were to take as initial conditions a sentence which by itself logically implied some conjunct of *E* we would consider the result in part a sort of trivial self-explanation. We are thus led to the stipulation that *E* must be in conjunctive normal form and that no conjunct of *E* must be derivable from *C*. We could arrive at the same stipulation by consideration of the proof of Theorem 5 of [1]. One further simplification can be made. Only single conjuncts of *E* need be considered for direct explanation, since a conjunction is logically derivable from the class of its conjuncts. Note that in this form the stipulation is automatically satisfied by the requirement that *E* not be logically derivable from *C*. Hence the

[7] In their discussion they consider a case essentially equivalent to that of the preceding paragraph. They note that a complete prohibition on partial self-explanation would have the result that only singular sentences logically derivable from a theory would be explainable by that theory. Since they do not consider the present R2, nor, apparently, are they aware of the consequences embodied in Theorem 5 of [1]; they conclude that there is no advantage to be gained by attempting some restriction of partial self-explanation. The stepwise procedure of the present paper is an attempt to avoid both the result pointed out by Hempel and Oppenheim and the consequences embodied in Theorem 5 of [1].

essence of this second revision of Hempel and Oppenheim's proposal is contained in the stepwise procedure of factoring a sentence to be explained into its directly explainable components.

For the explanation of singular sentences the concept of S-explanation defined below is proposed as a revision of H-O explanation.

E is *directly S-explainable* by T if and only if there is a sentence C such that the following conditions are satisfied:

(1) T is a theory,
(2) C is a conjunction of true basic sentences,
(3) E is a disjunction of basic sentences,
(4) E is logically derivable from the set $\{T, C\}$, and
(5) E is not logically derivable from $\{C\}$.

E is *S-explainable* by T if and only if E is a singular sentence which is logically derivable from the set of sentences which are directly S-explainable by T.

Of the following three theorems, Theorems 6 and 7 are easily derivable from the preceding definitions, and Theorem 8 has a proof analogous to that of Theorem 2.

Theorem 6. S-explanation satisfies R1.

Theorem 7. S-explanation satisfies R2.

Theorem 8. There is a language L which contains infinitely many individual constants and such that S-explanation satisfies R3 with respect to L.

It is clear that none of the arguments of Theorems 1-5 of [1] can be brought against S-explanation in view of the relationship, exploited in Lemma 1 of the present paper, between these arguments and the requirements R1, R2, R3.

One final theorem will be proved on the replacement, for purposes of explanation, of theories by laws. We must assume, for that part of the following lemma and theorem which refers to derivative theories, that the language in which the explanations take place contains at least one predicate which occurs in neither the theory to be replaced nor the singular sentence to be explained. This predicate may be of any degree. The proof of the following lemma uses the ideas of the proofs of theorems 6 and 7 of [1], although the present proof is simpler.

Lemma 2. Assume that T is a fundamental (derivative) theory. Then if E is directly S-explainable by T, there is a fundamental (derivative) law which is logically derivable from T and by which E is directly S-explainable.

Theorem 9. Assume that T is a fundamental (derivative) theory. Then if E is S-explainable by T, there is a fundamental (derivative) law which is logically derivable from T and by which E is S-explainable.

Proof. According to the hypothesis of the theorem there is a class K of singular sentences which are directly S-explainable by T and such that $K \vdash E$. We can find a finite class K^* such that $K^* \subset K$ and $K^* \vdash E$. Lemma 2 asserts that for each element k of K^* there is a fundamental (derivative) law l_k

which is logically derivable from T and by which k is directly S-explainable. Let L be the conjunction of the laws l_k for $k \varepsilon K^*$. We can construct a sentence L^*, logically equivalent to L but in which all quantifiers stand at the front, such that L^* is a fundamental (derivative) law. Since each conjunct of L is logically derivable from T, L^* is logically derivable from T. Now each element of K^* is S-explainable by a law which is logically derivable from L^*. Hence, by Theorem 6, we conclude that each element of K^* is S-explainable by L^*. Theorem 7 now assures that E, which is logically derivable from K^*, is S-explainable by L^*.

REFERENCES

[1] R. Eberle, D. Kaplan, R. Montague, 'Hempel and Oppenheim on Explanation', in the present issue of this journal.
[2] C. G. Hempel, P. Oppenheim, 'Studies in the Logic of Explanation', *Philosophy of Science*, *15* (1948), pp. 135-175.
[3] D. Hilbert, P. Bernays, *Grundlagen der Mathematik*, *2* (Berlin, 1939), p. 32.