David Lewis's Semantics for Deontic Logic

HOLLY S. GOLDMAN:

Numerous attempts to provide semantic interpretations for deontic logic are based on the idea that a given state of affairs *ought* to be the case in this world if and only if it *is* the case in every morally perfect world. In classical versions of this theory, a 'morally perfect world' is simply defined as one in which all the particular obligations obtaining in the actual world are fulfilled.[1] Such theories have two disadvantages. First, as Richard Purtill has pointed out, it is extremely difficult to spell out the requirement that all obtaining obligations be satisfied in a morally perfect world.[2] If a world contains no agents at all, are all our obligations fulfilled in it? If it contains agents who are not counterparts to the agents in this world, are our obligations fulfilled in it? Second, the value of such theories lies solely in their elucidation of the logical form of obligation statements, e.g., their explanation of what makes a set of obligation statements consistent. The application of such theories depends on *prior* possession of the list of true statements of particular obligations ('Jones ought to do A', 'Smith ought to do B', etc.). Since they do not offer an independent method for determining whether or not a given obligation obtains, they fail to provide illuminating reductive or eliminative definitions for statements of particular obligation.

David Lewis has proposed a semantic theory which falls within this general tradition but employs an alternative definition of the morally best worlds.[3] Because of this difference, his theory may escape Purtill's direct objections.[4] In addition, this aspect of his theory makes possible the derivation of particular obligations, not from an initial list of such obligations, but rather from a statement of abstract moral principles together with the general notion of a possible world. Thus it may be interpreted as providing not only an account of the logical form of statements of particular obligations, but also a reductive definition of them in terms of these other two concepts. Promising as this theory appears, I shall argue that it does not succeed, because it neglects the fact that

1 Perhaps the clearest examples of this sort of theory are given by Risto Hilpinen, 'Deontic Logic: An Introduction' and Jaakko Hintikka, 'Some Main Problems of Deontic Logic', in *Deontic Logic: Introductory and Systematic Readings* (New York: Humanities Press, 1971), pp. 1–35 and 59–104. Bengt Hansson in 'An Analysis of Some Deontic Logics' (in the same volume) and Bas van Fraassen in 'The Logic of Conditional Obligation' *Journal of Philosophical Logic*, i (1972), 417–438, apparently employ similar notions of morally perfect worlds.

2 Richard L. Purtill, 'Deontically Perfect Worlds and Prima Facie Obligations', *Philosophia*, iii (October, 1973), 429–438.

3 David Lewis, *Counterfactuals* (Cambridge, Massachusetts: Harvard University Press, 1973), pp. 96–104. Lewis speaks of 'better' and 'best' worlds rather than 'morally perfect' worlds.

4 I will not attempt to document this claim because my later arguments show that many of the difficulties which Purtill finds in the classical theories reappear for new reasons in Lewis's theory.

our obligations flow from abstract moral principles *in conjunction with* contingent features of our world.

According to Lewis, the 'goodness' of a possible world *vis-à-vis* our world is determined by an abstract principle of evaluation which imposes a preference ordering over possible worlds relative to the actual world.

> We may base a system . . . on comparative goodness of worlds. Suppose we have a preference ordering of the worlds, perhaps different from the standpoint of different worlds. As is the custom in deontic logic, I shall say nothing definite about the source and significance of this ordering. Perhaps the worlds are ordered according to their total net content of pleasure, measured by some hedonic calculus; or their content of beauty, truth, and love; or their content of some simple, non-natural quality. Perhaps they are ordered according to the extent that their inhabitants obey the law of God, of Nature, or of man. Perhaps according to how well they measure up to some sort of standards of objective morality, if such there be; perhaps according to someone's personal taste in possible worlds; perhaps according to calm, sympathetic, impartial contemplation of alternative possibilities. It does not matter. We can build in the same way on any of these foundations, or on others.[1]

As he notes, worlds must be ordered from the standpoint of a given world, for a world which is perfect relative to our world may not be perfect relative to some other world. For example, suppose that worlds are ordered according to the extent that their inhabitants obey the law of god. Since different gods, promulgating different codes of law, inhabit different worlds, activities in a world which satisfy the god in world i may not satisfy the god in world j. Moreover, some worlds may not be evaluable at all from the standpoint of a given world. For example, suppose the goodness of a world, from the standpoint of world i, depends on the extent to which its inhabitants obey the laws promulgated by the god who rules *their own* world. Then if there is a world k in which there is no god, it will not be clear whether it is better or worse than other worlds in which there are gods, and hence k may not be evaluable from the standpoint of i.[2] Given these complexities, Lewis offers the following truth definition for statements of unconditional obligation:

> (1) Op is true at world i if and only if
>
> (A) there are worlds evaluable from the standpoint of i, and
>
> (B) p holds at all worlds that are best from the standpoint of i.[3]

1 Lewis, ibid., p. 96. Since this statement is so open-ended, one could interpret it to include a principle of evaluation which ordered worlds according to the extent to which particular obligations obtaining in our world are fulfilled. Use of this principle would of course render Lewis's theory subject to exactly the same complaints I earlier lodged against the classical theories, and I shall therefore ignore it as a possibility.

2 Lewis, ibid., pp. 96–97, 99.

3 Lewis, ibid., pp. 100–101. I have restated his definition in more standard notation. He provides a more general definition which applies even in cases where there are no best worlds, but rather an infinite ascent to better and better worlds.

The difficulty with this definition becomes apparent as soon as it is applied to particular cases, for it fails to attribute obligatoriness to states of affairs we judge to be morally required. We may hold that John ought to save from drowning the small child who falls into the swimming pool directly in front of him. Assume this judgment is based on the abstract principle that each agent ought to fulfil the most stringent duties incumbent upon him. According to (1), it ought to be the case that John saves the child only if he performs this act in all the worlds which are best from the standpoint of our world. Employing the same principle used in the original judgment, we can stipulate that the goodness of a world depends on the extent to which its inhabitants fulfil the most stringent duties incumbent upon them. But with morally preferable worlds so defined, there is no reason to suppose that John saves the child from drowning in *all* the best worlds: there are worlds—at least as good as any others by the above standard, since they involve no violation of duty on anyone's part—in which John fails to perform this act. There are worlds in which the child does not fall into the pool in the first place, and worlds in which he falls in but knows how to swim and does not need to be rescued. In these worlds, John does not—in fact cannot—rescue the child. But (1) implies that if he does not perform this act in *all* the best worlds, then the act is not obligatory, despite the fact that it obviously is.

Lewis's theory goes astray because it ignores the fact that particular obligations flow from abstract principles *together with* contingent features of the world which are not required by the moral principle in question: features such as the child's falling into the pool and his inability to swim. Since these features are not required by the governing moral principle, they do not appear in *all* the morally best worlds—and neither do the actions whose occurrence or obligatoriness depend on them. Deontic logicians have noticed that certain obligations ('contrary-to-duty obligations') depend on contingent features of the world, namely prior violations of duty, and have recognized that such obligations cannot be handled within the classical semantic systems in the same manner as primary obligations.[1] But Lewis, at least, appears to miss the fact that even primary obligations arise from the sort of contingent features of the world I have described, and thus cannot adequately be handled by the definition he proposes.

There are three ways Lewis might attempt to defend his system against this objection. First, he might claim that although his theory fails as a semantic account of the 'ought-to-do', it succeeds as an account of the 'ought-to-be'. Let us loosely explain these two notions by saying that the concept of what ought to be *done* only applies to acts, whereas the concept of what ought to *be* applies to all other states of affairs.[2] This latter is a notion we more commonly express by saying 'It would be good if such-and-such were to happen'. Although Lewis's examples of

1 See discussion and references in Hilpinen, 'Deontic Logic: An Introduction', in Hilpinen, op. cit., Sections VIII and IX.

2 For a discussion of this distinction, see Hector-Neri Castañeda, 'On the Semantics of the Ought-to-Do', *Synthese* xxi (1970), 449–468.

obligatory states of affairs typically involve the performance of actions, some of his remarks suggest he is interested in the notion of the 'ought-to-be':

> 'Obligation' is here used in a special, impersonal sense. What is obligatory (conditionally or unconditionally) is what ought to be the case, whether or not anyone in particular is obligated to see to it. Personal obligations may or may not follow from these impersonal obligations (ibid. p. 100, n.).

If Lewis does intend to provide a semantic account for the notion of the 'ought-to-be', then the foregoing counterexample does not show the account to be incorrect. However, it is not difficult to construct counterexamples showing that even on this interpretation his definition is subject to the same sort of difficulty. Suppose we ascertain whether or not states of affairs ought to occur according to their contribution to human happiness. Then we might judge that the volcano ought not to erupt, since this would destroy the village below and cause great distress to the villagers. But it is not true in *all* the best worlds, judged by this principle, that the volcano does not erupt. There are worlds as happy as any other in which the village is located on the opposite side of the volcano and would not be destroyed by its erupting—in fact would benefit from the increased tourist trade an eruption would generate. Thus on Lewis's definition it is false that the volcano ought not to erupt, whereas clearly it should not. His theory fares no better as an account of the 'ought-to-be' than it does as an account of the 'ought-to-do', again because it fails to accommodate the role of contingent features of the actual world (such as the village's location) in determining what states of affairs are obligatory.

The second way Lewis might attempt to defend his system is by claiming that although (for example) there is no unconditional obligation in the previous case for John to save the child, there is a *conditional* obligation, and it is this conditional obligation we mean to assert when we say 'John ought to save the child'. The following passage strongly suggests Lewis might take this line:

> There is a natural way to construe 'It ought to be that [Jesse confesses and gives back the loot]' so that it does become true when Jesse robs the bank. It can be taken as tacitly conditional, meaning something like 'Given those actual circumstances that now cannot be helped, it ought to be that [Jesse confesses and gives back the loot]' (ibid., p. 102, n.).

Thus he might suggest that 'John ought to save the child' should be understood as 'Given that the child falls into the pool, cannot swim, etc., then it ought to be that John saves the child'. But this tactic is unsatisfactory. Since virtually every obligation (not just those like Jesse's which depend on prior violations of duty) depends on contingent facts about the world, this strategy involves claiming that virtually every statement of unconditional obligation is a tacit statement of conditional obligation. Because we are conscious of the distinction between conditional and

unconditional obligation, and careful to make it in everyday conversation, this is implausible, and should not be accepted merely on the grounds that it is necessary to save a certain theory.[1] Moreover, if the alleged conditional obligation is interpreted according to Lewis's account of conditional obligation, it is not possible to detach from it, together with the truth of its antecedent, an *unconditional* obligation (i.e., 'John ought to save the child'). Thus we have nothing *but* the conditional obligation. But unless their consequents are detachable, conditional obligations have no import for action. Nevertheless what we need and mean to assert in this case is a direct prescription for action—a direct prescription which cannot be captured by any merely conditional obligation. For these reasons interpreting the statement 'John ought to save the child' as a conditional obligation is not a satisfactory way for Lewis to meet the criticism outlined above.

My initial argument shows that Lewis's account goes astray in defining obligation by reference to morally preferable worlds which may not be sufficiently similar to the actual world. If the preferable worlds were required to duplicate relevant contingent features of the actual world, then (for example) they would all have to include the child's falling into the pool and his inability to swim, and thus include John's saving the child. But Lewis might claim, as a third method of defending his account, that his notion of 'evaluability from the world *i*' includes constraints on the morally preferable worlds which enable him to counter my objection.[2] Specifically he might argue that this notion must be understood in such a way that a world *j* is evaluable from the standpoint of a world *i* only if *j* is sufficiently similar to *i*. Unfortunately, this expedient would not solve the problem. Any world *i* contains many acts, any one of which, say act *A*, must be regarded as part of the fixed conditions a sufficiently similar world *j* must duplicate when the moral status of some other act is being assessed, but which must not be included among these conditions when the moral status of *A* itself is being assessed. Thus if two drivers Green and White are set on a collision course, we judge that Green ought to swerve north partly because White will in fact swerve south. To determine how Green should act, we must restrict our attention to worlds resembling ours with respect to White's swerving south. But when we ask what *White* ought to do, we cannot restrict our attention to worlds in which he swerves south, since part of what we want to know is whether his swerving south is better than his not swerving south. Thus the notion of 'evaluable from world *i*' would have to be relativized to the act or state of affairs whose moral status is at issue. Incorporating this relativization into Lewis's scheme would

1 The suggestion would be less implausible if, as Lewis appears to suggest, the antecedent of the condition could be understood as the same in every case—something on the order of 'Given those actual circumstances which now cannot be helped'. However, as I argue in the next paragraph, this is not possible, since the circumstances which are relevant to the moral status of one action are not necessarily relevant to the moral status of other actions which would occur *at the same time*. Thus the antecedent would have to be different in each case.

2 I owe this suggestion to Louis Loeb.

represent a substantial change in his theory, and it remains to be seen whether or not it could be successfully worked out.

It appears that Lewis's theory of unconditional obligation cannot be defended against the objection I have raised. Reflection on the objection suggests his definition of *conditional* obligation is subject to the same difficulties. We assert conditional obligations when we say such things as 'It ought to be the case that if Jesse robs the bank, then he returns the loot', or, more naturally, 'If Jesse robs the bank, then he ought to return the loot'. Lewis believes that such a statement is true just in case the best worlds in which Jesse robs the bank (which are not, of course, among the best of all possible worlds) are worlds which include his returning the loot. This suggestion is expressed in the following definition:

(2) $O[q/p]$ is true at world i if and only if
  (A) there are no evaluable worlds in which p is true, or
  (B) some p & q world is better, from the standpoint of i, than any p & *not-q* world.[1]

Once again, application to actual cases shows that this definition fails to ascribe truth to statements of conditional obligation which are obviously true, because it fails to take adequate account of the affect of contingent features of the actual world on such obligations. Suppose I promise to return a borrowed book tomorrow. Clearly, if I do not return the book tomorrow, I ought to apologize. But now consider possible worlds in which I do not return the book tomorrow. It is better, other things being equal, not to break a promise. Thus the *best* worlds in which I fail to return the book tomorrow are surely worlds in which I do not thereby break a promise: either because I am released from that promise before tomorrow, or because I never made such a promise in the first place. But in either of these worlds, nothing is gained by my apologizing for not returning the book. So it appears that a world in which I do not return the book and do not apologize (e.g., a world in which I have been released from my promise) is at least as good as any world in which I do not return the book but do apologize. The statement 'If I do not return the book tomorrow, I ought to apologize', turns out to be false according to definition (2). Nevertheless this statement is true, because in the actual world I will not be released from the promise.

One might attempt to defend definition (2) by arguing that the conditional sentence 'If I do not return the book tomorrow, then I ought to apologize', is merely a shorthand way of expressing what the speaker really means, which is more adequately expressed by the following statement: 'If I have promised to return a book tomorrow, and I am not released from that promise, and I do not return the book, then I ought to apologize'. The only version of the original statement which is completely immune to the sort of argument I have just advanced would require an antecedent describing *all* the features of the actual world which affect the suitability of my apologizing. But no speaker of ordinary

1 Lewis, ibid., p. 100. Again, I have slightly changed his notation.

English who utters the original sentence intends it to express a statement containing such an expanded antecedent—an antecedent which may refer to indefinitely many facts, few of which would be known to him. Thus this defence of definition (2) fails.

Lewis's semantic definitions of unconditional and conditional obligation avoid some of the problems afflicting classical accounts, and moreover provide a promising reductive definition of obligation statements, as well as an account of their logical form. However, his theory is incorrect, because it fails to accommodate the fact that particular obligations flow from contingent facts about the actual world as well as from abstract moral principles.[1]

UNIVERSITY OF MICHIGAN

1 I am grateful to John G. Bennett, Brian Chellas, Alvin Goldman, and Louis Loeb for their helpful comments on earlier versions of this paper. Final revisions on it were made during my tenure as an American Association of University Women Fellow.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS ABERDEEN