DAWES HICKS LECTURE ON PHILOSOPHY

G. E. MOORE ON THE NATURALISTIC FALLACY

By C. LEWY

Read 11 November 1964

G. E. Moore's literary remains contain very little concerning ethics; but they include an unfinished draft (in manuscript) of what was intended to be a preface to the second edition of *Principia Ethica*. For various reasons it seems to me highly probable that this was written in 1920 or 1921; but in the end Moore abandoned the idea of a second edition, and in 1922 *Principia* was reprinted without any alterations, except for the correction of a few misprints and grammatical mistakes and the inclusion of a prefatory note of seven lines.

Owing to the fact that the draft is unfinished and in parts very fragmentary, the task of preparing it for publication would be a very difficult one, though I may possibly attempt it in the future. What I want to do today is first to give a synopsis, or rather a reconstruction, of what seem to me to be the main points of the unpublished preface (which from now on I shall simply call 'the Preface'), and secondly to discuss independently one particular aspect of the subject.

I

Moore begins by pointing out that there are several senses of the word 'good', and that in *Principia* he was concerned with only one of them. He does not now think, however, that this sense can be called *the* ordinary sense of the word, even if any one sense of it is commoner than any other. But he thinks that the sense in question can be specified by saying that it is *the* sense which has a unique and fundamentally important relation to the conceptions of right and wrong. *What* the relation in question *is*, he proposes, he says, to discuss later; but in fact no such discussion is included in the Preface.

He goes on to ask, however, what are the main things that he wished to say in *Principia* about the concept which is expressed by the word 'good', when the word is used in this sense. The

first thing he wished to say, he continues, is that Good[1] is simple in the sense of being indefinable or unanalysable. Is this proposition true?, he asks. He still thinks it is probably true, but he is not certain, for it seems to him that possibly 'right' is unanalysable, and Good is to be analysed partly in terms of 'right'. But whether Good is analysable or not does not seem to him now nearly as important as it did when he wrote *Principia*. If Good *were* unanalysable, it would follow that it could not be identical with any such property as 'is desired' or 'is a state of pleasure', since these *are* analysable; but it would be a great mistake to suppose that, as he implied in *Principia*, the fact that Good is not identical with any such property *rests* on the contention that Good is unanalysable.

He says that in the passage in *Principia* (§§ 6–14) in which he asserted that Good was unanalysable, he made another assertion which must not be confused with it, though he did so confuse it, namely, the assertion '. . . good is good, and that is the end of the matter' (*Principia*, p. 6). What, he asks, did he mean by this? Clearly, he meant to assert about Good what Bishop Butler, in the passage which Moore quoted on his title-page, asserted to be true of everything, namely, that it is what it is, and not another thing. In other words, he meant to assert that Good is Good, and nothing else whatever.

But this, Moore now says, may mean *either* 'Good is different from everything other than Good' *or* 'Good is different from everything which we express by any word or phrase other than the word "good" '. The first is wholly trivial and unimportant; and that Good is unanalysable cannot possibly follow from it, since the property of being different from every property that is different from it, is a property which must belong to every property without exception, analysable and unanalysable alike. And for the same reason it cannot possibly follow from it that certain particular properties such as 'is a state of pleasure' or 'is desired' are different from Good. For even if Good were identical with, say, 'is desired', Good would still be different from every property which was different from it.

The second assertion, however—that Good is different from everything which we express by any word or phrase other than the word 'good'—is far from being trivial. If it were true, it would really follow that Good was different from any such

[1] As Moore himself does in the Preface, I shall write *Good*, with a capital *G* but without quotes, when I talk about the concept, and not the word. But (again like Moore) I shall not adopt this device in connexion with other concepts.

property as 'is a state of pleasure' or 'is desired'. And also, if it were true, it would afford at least a strong presumption that Good was unanalysable. For 'where a word expresses an *analysable* property, that property is generally also sometimes expressed by a phrase, made up of several words, which point out elements which enter into its analysis, and, in that sense, "contain an analysis" of it'. So that if Good were analysable, it would probably be sometimes expressed by some such complex phrase—a phrase, therefore, different from the mere word 'good'. Indeed Moore thinks that this fact probably partly explains how he was led to identify such obviously different propositions as 'Good is Good, and nothing else whatever' and 'Good is unanalysable'. For we have just seen that if the former proposition be understood as asserting that Good is different from any property expressed by any phrase other than the word 'good', this proposition, if true, would at least afford a strong presumption that Good was unanalysable. And he may have supposed—he continues—that, conversely, from the fact that Good was unanalysable, it would follow that it could not be expressed by any phrase other than 'good'. He may have supposed so owing to his perceiving that if Good were unanalysable, it could not be expressed by any phrase which *contained an analysis* of it, but failing to perceive the distinction between expressing the meaning of a word in other words which *contain an analysis* of it, and expressing its meaning by giving a synonym.

But the fact that there is this distinction is fatal to the truth of the proposition we are now considering. It may be true that Good is unanalysable, and therefore cannot be expressed by other words which contain an analysis of it: but it is certainly not true that it cannot be expressed by any other words at all. For instance (quite apart from the obvious fact that there are languages other than English), the word 'desirable' is sometimes used as a synonym for 'good'.

Moore therefore concludes that the assertion 'Good is Good, and nothing else whatever' is either merely trivial or else obviously false.

But this is not the end of the matter. For Moore also thinks that the examples which he gave in *Principia* do suggest to most people's minds that what he really meant to assert was that Good was not identical with any property belonging to a *particular class*; and *this* assertion still seems to him both true and important. But what is the class in question? Moore says in effect that he can only describe this class by saying that it is

the class of all those properties which are either natural or metaphysical; and what he really wanted to assert, he says, was that Good was not identical with any natural or metaphysical property.

He admits that in *Principia* he confused natural objects (or events) with a certain kind of property which may belong to them. He actually confused a particular event, which consists in somebody's being pleased, with the property which we ascribe to it when we say that it is 'a state of pleasure'—just as he confused a particular patch of yellow with the property of being yellow. And he also admits that he confused *parts* of natural objects with *properties* of such objects.

For these and other reasons his attempts to define a 'natural property' were, he says, hopelessly confused. The nearest he came to suggesting a correct definition in *Principia* was on p. 40, where he said that to identify Good with any natural property resulted in replacing ethics by one of the natural sciences (including psychology). This now suggests to him the following definitions. A 'natural' property is a property with which it is the business of the natural sciences or of psychology to deal, or which can be completely defined in terms of such. A 'metaphysical' property is a property which stands to some supersensible object in the same relation in which natural properties stand to natural objects.

Moore now points out that the proposition that Good is not identical with any natural or metaphysical property (as now defined)—which is what he really wished to assert in *Principia*—neither implies nor is implied by the proposition that Good is unanalysable. For it might plainly be true, even if Good *were* analysable; and, on the other hand, even if Good were *un*analysable, Good might still be identical with some natural property, since many such properties may be unanalysable. At the same time, he says, if Good is not identical with any natural or metaphysical property, it does follow that, if it is analysable at all, it involves in its analysis *some* unanalysable notion which is not natural or metaphysical. That some unanalysable notion of this sort, he says, is involved in ethics was certainly a part of what he wished to assert when he asserted that Good was unanalysable. Only he did not see that this was a far more important and less doubtful assertion than that Good itself was the unanalysable notion in question.

Of course Moore realizes that his new definitions—and it would perhaps be better to call them 'explanations' rather than

'definitions'—are still not fully satisfactory. It is clear that he intended to return to the topic in a later part of the Preface; but he never in fact came to write it.

There are, however, still some pages of the Preface which are of considerable interest, and of special relevance to our subject. It will have been noticed that so far the expression 'the naturalistic fallacy' has not been introduced, although it is obvious that what Moore meant by it is very closely connected with the propositions we have been considering. But he now explicitly raises the question: What *is* 'the naturalistic fallacy'? And he says that the most important mistake which he made in his discussion of the matter in *Principia* was exactly analogous to the chief of those which he made in his assertions about Good. In the latter case, as we have seen, he confused the three entirely different propositions 'Good is not identical with any property other than itself'; 'Good is not identical with any analysable property'; and 'Good is not identical with any natural or metaphysical property'. In the case of the naturalistic fallacy, he goes on, he similarly confused the three entirely different propositions (1) 'So-and-so is identifying Good with some property other than Good'; (2) 'So-and-so is identifying Good with some *analysable* property'; and (3) 'So-and-so is identifying Good with some *natural or metaphysical* property'.

He points out that he sometimes implies that to say of a man that he is committing the naturalistic fallacy is to say (1) of him; sometimes that it is to say (2) of him; and sometimes that it is to say (3) of him.

But in addition to this, his main mistake, he also made, he says, two further mistakes. First, he sometimes talked (*Principia*, p. 14) as if to commit the naturalistic fallacy was to suppose that in, for example, 'This is good', the word 'is' always expresses identity between the thing called 'this' and Good. And secondly, he confused (A) 'To say that so-and-so is committing the naturalistic fallacy is to say that he is holding, with respect to some property of a certain kind, the *view* that that property is identical with Good', and (B) 'To say that so-and-so is committing the naturalistic fallacy is to say that he is *confusing* some property of a certain kind with Good'. But the operation mentioned in (A) is quite different from that mentioned in (B).

Finally, Moore admits that he feels doubtful whether either of these two operations could properly be called the commission of a fallacy, for the simple reason that to commit a fallacy seems properly to mean to make a certain kind of *inference*; whereas

the mere confusion of two properties, or the holding of a view with regard to them, seems not to be a process of inference at all.

Moore ends this part of the Preface by saying that if he still wished to use the term 'naturalistic fallacy', he would define it as follows: 'So-and-so is committing the naturalistic fallacy' means 'He is *either* confusing Good with a natural or metaphysical property *or* holding it to be identical with such a property *or* making an inference *based* upon such a confusion'. And he would also expressly point out that in so using the term 'fallacy' he was using it in an extended, and perhaps improper, sense.

This concludes my synopsis, or reconstruction, of the Preface, or rather of that part of it which it is possible to reconstruct, for the rest is in a very incomplete state indeed. And it will, I think, have been seen that many of the criticisms made of Moore's treatment of the naturalistic fallacy and related topics in the 1930's and 1940's were fully anticipated by him many years earlier.

II

I now wish to discuss independently one particular aspect of the subject. In the Preface, it will be recalled, Moore says that he still believes it to be true and important to assert that Good is not identical with any natural or metaphysical property. But he neither produces any new arguments for this assertion nor makes any comments on the arguments which he gave in *Principia*. I wish now to examine in some detail two passages in the book which contain such arguments. The first occurs in § 13 (pp. 15–16), and runs as follows:

> The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition be offered, it may be always asked, with significance, of the complex so defined, whether it is itself good. To take, for instance, one of the more plausible, because one of the more complicated of such proposed definitions, it may easily be thought, at first sight, that to be good may mean to be that which we desire to desire. Thus if we apply this definition to a particular instance and say 'When we think that A is good, we are thinking that A is one of the things which we desire to desire,' our proposition may seem quite plausible. But, if we carry the investigation further, and ask ourselves 'Is it good to desire to desire A?' it is apparent, on a little reflection, that this question is itself as intelligible, as the original question 'Is A good?'—that we are, in fact, now asking for exactly the same information about the desire to desire A, for

which we formerly asked with regard to A itself. But it is also apparent that the meaning of this second question cannot be correctly analysed into 'Is the desire to desire A one of the things which we desire to desire?': we have not before our minds anything so complicated as the question 'Do we desire to desire to desire to desire A?' Moreover anyone can easily convince himself by inspection that the predicate of this proposition—'good'—is positively different from the notion of 'desiring to desire' which enters into its subject: 'That we should desire to desire A is good' is *not* merely equivalent to 'That A should be good is good'. It may indeed be true that what we desire to desire is always also good; perhaps, even the converse may be true: but it is very doubtful whether this is the case, and the mere fact that we understand very well what is meant by doubting it, shews clearly that we have two different notions before our minds.

The second passage occurs a little later (p. 38). Moore there says that he will discuss certain theories which claim that only a single kind of thing is good. He thinks that such theories rest on the naturalistic fallacy, and goes on as follows:

> That a thing should be good, it has been thought, *means* that it possesses this single property: and hence (it is thought) only what possesses this property is good. The inference seems very natural; and yet what is meant by it is self-contradictory. For those who make it fail to perceive that their conclusion 'what possesses this property is good' is a significant proposition: that it does not mean either 'what possesses this property, possesses this property' or 'the word "good" denotes that a thing possesses this property'. And yet, if it does *not* mean one or other of these two things, the inference contradicts its own premise.

It will have been noticed that Moore speaks in these passages as if he were showing that Good is not analysable at all; but what I chiefly wish to discuss is the question whether he has shown that Good is not identical with the property of being one of the things which we desire to desire—that is, with the property which he takes as an example in the first passage. Moreover, I cannot hope to say here all that ought to be said about these passages. In particular, I cannot consider all the different arguments which they contain and which are not clearly distinguished from each other. All I can do is to try to reformulate and discuss what seems to me to be the chief of these arguments.

I think I can do this most clearly with the help of an analogy. Let us suppose that we are concerned, not with Good, but with the concept of being a brother. Suppose that someone asserts that to be a brother is to be a male sibling—or, to use the terminology that Moore himself often used in later life—that the concept of

being a brother is identical with the concept of being a male sibling. Now what follows from this proposition? So far as I can see, one thing which certainly follows from it is that the proposition 'John is a brother' is identical with the proposition 'John is a male sibling'. Similarly, in Moore's case, if to be good is to be one of the things which we desire to desire, it follows that any proposition of the form '*x* is good' is identical with the corresponding proposition of the form '*x* is one of the things which we desire to desire'. It follows, for instance, that the proposition 'A is good' (and we must now assume that 'A' is a name or description of a thing or state of things) is identical with the proposition 'A is one of the things which we desire to desire'.

Consequently, Moore could have argued against the identification of Good with the property of being one of the things which we desire to desire, by pointing out that even if at first it may seem plausible to suppose that these two propositions are identical, yet further reflection makes it apparent that they are *not* identical.

But this is not what he does. He obviously thought that he had a more complicated but more convincing argument. For what he asks us to consider are not the two propositions I have just mentioned, but the completely different propositions 'It is good to desire to desire A' and 'The desire to desire A is one of the things which we desire to desire'. And he says that it is apparent on reflection that *these* propositions are not identical.

Let me put the matter in terms of questions rather than propositions. Moore could have argued that the question (1) 'Is A good?' is quite different from the question (2) 'Is A one of the things which we desire to desire?'. Yet if to be good is to be one of the things which we desire to desire, these questions are identical. But what he in fact says is that the question (3) 'Is it good to desire to desire A?' is quite different from the question (4) 'Is the desire to desire A one of the things which we desire to desire?'.

But though the latter questions are more complicated than the former, they are no better. For on the view he is discussing, just as (1) and (2) are identical, so are (3) and (4). And it is no plainer that (3) and (4) are *not* identical than it is that (1) and (2) are not identical. Similarly, on the view in question, the proposition (3A) 'It is good that we desire to desire A' *is* identical with the proposition (3B) 'It is good that A is good' (and each of them is identical with the proposition 'We desire to desire to desire to desire A'). And again, it is no plainer that (3A) and (3B)

are *not* identical than it is that 'A is good' and 'A is one of the things which we desire to desire' are not identical.

Did Moore, then, have at the back of his mind some other questions, even more complicated? I think that the second passage which I have quoted makes it fairly clear that he did, and that they were (5) 'Is A, which is one of things which we desire to desire, good?', and (6) 'Is A, which is one of the things which we desire to desire, one of the things which we desire to desire?'. And I think that he confused (5) with (3), and (6) with (4).

Unfortunately, each of these last two questions—(5) and (6)—is capable of at least two totally different interpretations. Question (5) may mean *either* 'Is it the case that A is good if and only if it is one of the things which we desire to desire?'— where the expression 'if and only if' is used truth-functionally[1]; *or* 'Is it the case that to say that A is good is the same thing as to say that A is one of the things which we desire to desire?'. More generally, the question of which (5) is merely a particular example may mean *either* 'Is it the case that a thing is good if and only if it is one of the things which we desire to desire?' (where the expression 'if and only if' is used truth-functionally); *or* 'Is it the case that to be good is to be one of the things which we desire to desire?'. An affirmative answer to the *first* question would be given by the proposition 'It *is* the case that a thing is good if and only if it is one of the things which we desire to desire', which is logically equivalent to the proposition (α) 'A thing is good if and only if it is one of the things which we desire to desire'. An affirmative answer to the *second* question would be given by the proposition 'It *is* the case that to be good is to be one of the things which we desire to desire', which is logically equivalent to the proposition (β) 'To be good is to be one of the things which we desire to desire'.

Similarly, the question of which (6) is merely a particular example may mean *either* 'Is it the case that a thing is one of the things which we desire to desire if and only if it is one of the things which we desire to desire?' (where 'if and only if' is used truth-functionally); *or* 'Is it the case that to be one of the things which we desire to desire is to be one of the things which we desire to desire?'. An affirmative answer to the *first* question would be given by the proposition 'It *is* the case that a thing is

[1] That is to say, in such a way that the question can also be expressed by asking 'Are the two propositions "A is good" and "A is one of the things which we desire to desire" either *both* true or *both* false?'.

one of the things which we desire to desire if and only if it is one of the things which we desire to desire', which is logically equivalent to (γ) 'A thing is one of the things which we desire to desire if and only if it is one of the things which we desire to desire'. On the other hand, an affirmative answer to the *second* question would be given by the proposition 'It *is* the case that to be one of the things which we desire to desire is to be one of the things which we desire to desire', which is logically equivalent to the proposition (δ) 'To be one of the things which we desire to desire is to be one of the things which we desire to desire'.

For the sake of simplicity, I will now again speak in terms of propositions rather than questions. The main point I now wish to make is that there is a fundamental difference between (α) and (γ) on the one hand, and (β) and (δ) on the other. For the truth-value (that is, the truth or falsity) of (α) would not be altered if we substituted for any expression which occurs in the sentence which I have used to express (α), another expression with the same extension (that is, another expression which applies to exactly the same things); and the same is true of (γ). But this is not true either of (β) or of (δ). In current logical terminology, whilst the sentences which I have used to express (α) and (γ) are *extensional*, those I have used to express (β) and (δ) are *not* extensional.

It is clear that at the time Moore wrote *Principia* (1903), he did not see this distinction; and he therefore failed to distinguish (α) from (β), and (γ) from (δ). But (α) *is* quite different from (β), and (γ) *is* quite different from (δ). Consequently, we get two different interpretations of Moore's argument.

First, we can interpret him as arguing that to be good is not the same as to be one of the things which we desire to desire, because, if it were, then (β) would be identical with (δ); and maintaining, further, that it is apparent on reflection that (β) is *not* identical with (δ). If interpreted in this way, the argument seems to me to be completely invalid. For in the same kind of way it would be possible to show with regard to any concept whatever that it is unanalysable—in other words, that it is simple. For instance, we could show that to be a brother is not the same thing as to be a male sibling, because, if it were, then the proposition 'To be a brother is to be a male sibling' would be identical with the proposition 'To be a male sibling is to be a male sibling'. Yet it is clear on reflection that these propositions are *not* identical.

In other words, Moore's argument, in this interpretation,

would be a particular instance of what he himself later in life called the 'Paradox of Analysis'. He was never fully satisfied with any solution of it, and said different things about it at different times. But I have no doubt at all, on the basis of a large number of discussions which I have had with him on the subject over a period of many years, that his considered view was that whatever may be the *complete* solution, it was essential to hold that (in the example I have just given) to be a brother *is* to be a male sibling, and that yet the proposition 'To be a brother is to be a male sibling' is *not* identical with the proposition 'To be a male sibling is to be a male sibling'. And he therefore held that from 'To be a brother is to be a male sibling', the identity of these propositions does *not* follow. I think that this is right; and if so, then his *Principia* argument, in the interpretation I am now considering, is clearly invalid.

We must now, however, discuss my second interpretation. Here we should interpret Moore as arguing that to be good is not the same as to be one of the things which we desire to desire, because, if it were, then (α) would be identical with (γ); and maintaining, further, that it is apparent on reflection that (α) is *not* identical with (γ). Now *this* argument seems to me to be perfectly valid. For, although I once succeeded in so confusing myself as to deny it, I now think it undeniable that *if* to be good is to be one of the things which we desire to desire, then (α) *is* identical with (γ). Yet it is absolutely clear that (α) is *not* identical with (γ). And that (α) is not identical with (γ) follows from something which is also absolutely clear, namely, that it is logically possible to doubt (α) *without* doubting (γ); and each of these things follows from something which is also absolutely clear, namely, that whilst (γ) is a necessary proposition, (α) is a contingent proposition.

Moreover, it is *not* possible to use this kind of argument to show with regard to any concept whatever, that it is unanalysable. Indeed, if to be a brother is to be a male sibling, then the proposition 'A creature is a brother if and only if it is a male sibling' is identical with the proposition 'A creature is a male sibling if and only if it is a male sibling' (where in both sentences 'if and only if' is used truth-functionally). But *these* propositions *are* identical.

Of course, Moore's argument, in the present interpretation, may be said to be 'begging the question'. For a person who holds that to be good is to be one of the things which we desire to desire, may admit that if this is so, then (α) is identical with (γ);

and he may then go on to assert that (α) *is* identical with (γ). This is true: but I think we can all see that a person who asserted *this*, would be mistaken.

It seems to me obvious that any theory which identifies Good with a concept which is not itself at least partly ethical, can be refuted in an analogous way. I think therefore that for all his mistakes, Moore can fairly be said to have found a means of refuting any such theory.