Proceedings of the British Academy, **95**, 135–158

COMMENTARY

Williamson on Iterated Attitudes

DOROTHY EDGINGTON

HERE IS ONE WAY of thinking of the phenomenon under scrutiny in Timothy Williamson's impressive paper. In Fregean semantics, the distinction between sense and reference applies within every semantic category. Two expressions may have the same reference yet differ in sense. Two expressions may even have the same reference in every possible situation, yet differ in sense: Frege's first example of the distinction is a mathematical one, two ways of designating the same point (1892, p. 57). Two expressions which differ in sense are not universally intersubstitutable salva veritate. Williamson's study could be seen as a partial investigation of the sense-reference distinction for *sentence operators*: its effect on the iteration of an operator—what happens when an operator is embedded in another occurrence of itself.

Two sentence operators O_1 and O_2 may be extensionally equivalent: for any sentence p, O_1p iff O_2p; they may even be, in one good sense, intensionally equivalent—in all possible worlds O_1p iff O_2p; yet they behave differently on iteration. For instance, although their equivalence ensures O_1O_1p iff O_2O_1p, we may not have that O_1O_1p iff O_2O_2p. Or again, $O_1(O_1p \rightarrow p)$ iff $O_2(O_1p \rightarrow p)$, but perhaps not $O_1(O_1p \rightarrow p)$ iff $O_2(O_2p \rightarrow p)$. As Williamson puts it, the operators may 'satisfy different principles'—even inconsistent principles.

It is not an easy task to reply to a paper whose aim is 'not to advance a thesis, but to explore a phenomenon', and which carries out that exploration with ingenuity and painstaking care. I shall first add some comments about Williamson's initial examples. I shall sometimes

pursue an example a little further than he chooses to do, in order to assess, at the end, the extent to which the philosophically interesting problems he raises have this phenomenon as their source or the key to their solution. I shall then give a selective overview of Williamson's formal development of his topic, examine a closely-related problem about iteration which arises in a semantics for conditionals, and make some concluding remarks on the philosophical significance of this endeavour.

1. Knowing that you know

First consider operators of the form '*a* knows that'. Oswald, a timid, bespectacled, bookish schoolteacher, is widely known to generations of his pupils as Tarzan. Oswald is aware that this name is in use, and has picked up a fair amount of information about Tarzan, not least from the early pages of an autobiography of an ex-pupil, but it does not occur to him that the name refers to himself. Whatever Oswald knows, Tarzan knows: 'Oswald knows that' and 'Tarzan knows that' are coextensive operators. If we follow Saul Kripke (1972) and treat the names as rigid designators, the operators are coextensive in all possible worlds. Oswald has been marking, and now knows that the year's prize goes to Ann. Oswald knows that Oswald knows that Ann wins the prize. Hence, Tarzan knows that Oswald knows that Ann wins the prize. But Oswald (Tarzan) doesn't know that Tarzan knows that Ann wins the prize, for he doesn't know that he is Tarzan.

If the case is a little strained with proper names, it is not at all strained for identities of the form '*a* is the *F*', for it is easy for you to be the *F* without knowing that you are. This can be so even if it is necessary that you are the *F*. *F* could be a uniquely identifying essential property of yours, of the Kripkean sort, which you do not know you have. It could even be analytic, hence knowable a priori, that you are the *F*, but you haven't figured out that this is so. Let 'the *F*' be 'the grandchild of your maternal grandmother the set of whose older siblings and first cousins on the grandchild's mother's side is equinumerous with the set of your maternal grandmother's grandchildren who are older than you'. Whatever you know, the *F* knows. You know (the *F* knows) that you know that it rained in London today. But you don't know (the *F* doesn't know) that the *F* knows this.

Williamson discusses this example in the context of the 'KK

principle': if one knows something, one knows that one knows it. He does not accept the principle, and nor do I; but whether one accepts it or not, the question arises how it should be formulated. Consider

For all p, if t knows that p, then t knows that t knows that p.

Substituting 'I' for t gives, in my mouth, the statement that the KK principle applies to me. But any other uniform substitution for t seems to run the risk of falsifying the principle in a way that is irrelevant to its intended meaning: for $t \neq$ 'I', the principle might fail because the knower doesn't recognize themself as t. We cannot rest content with a formulation of a principle intended to be of general application, which is stateable only in the first person, about oneself. I think this shows that the KK principle is not adequately formulated as involving the iteration of a one-place sentence operator. It attributes, to anyone who knows that p, what David Lewis calls a *de se* attitude, in this case a piece of *de se* knowledge, the self-ascription of the property of knowing that p. Lewis argues that a *de se* attitude is not an attitude to a proposition:

> Consider the case of two gods. . . . [T]hey know every proposition that is true at their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is. They are not exactly alike. One lives on top of the tallest mountain and throws down manna; the other lives on top of the coldest mountain and throws down thunderbolts. Neither one knows whether he lives on the tallest mountain or the coldest mountain; nor whether he throws manna or thunderbolts.
>
> Surely their predicament is possible. (The trouble might perhaps be that they have an equally perfect view of every part of the world, and hence cannot identify the perspectives from which they view it.) But if it is possible to lack knowledge and not lack any propositional knowledge, then the lacked knowledge must not be propositional. If the gods came to know which was which, they would know more than they do. But they wouldn't know more propositions. There are no more to know. Rather, they would self-ascribe more of the properties they possess. . . . Some belief and some knowledge cannot be understood as propositional, but can be understood as self-ascription of properties. (1979a, p. 139)

Lewis's gods are an extrapolation from John Perry's amnesiac in the library who, despite having just read a biography of himself, doesn't know who he is (Perry 1977). Hector-Neri Castañeda (1968) coined the pronoun 'he*', or 'he himself' to force the intended reading of 'he' in, for instance, 'The shortest spy doesn't know that he is the shortest spy'. We do not need to settle here whether Lewis's denial of the propositional

nature of such self-ascribing knowledge is correct, to see that the required generality of the KK principle, combined with the special nature of self-ascription, render the principle unsuitable to be expressed by iteration of a single sentence operator.

2. *Not believing what you don't know*

Williamson's second example illustrates a different form of failure of substitutivity. 'On a cartoon version of the Stoic idea of wisdom, x is wise if and only if x believes only what x knows'. Suppose that Socrates has this remarkable property. Assume that knowing entails believing. Then for any proposition p, Socrates believes that p if and only if Socrates knows that p. Let Socrates know that he is Socrates, so that problems of the kind discussed above do not arise. We cannot everywhere substitute 'Socrates believes that' for 'Socrates knows that', salva veritate. Socrates knows that if he knows that it will rain tomorrow, it will rain tomorrow. But Socrates may not believe that if he believes that it will rain tomorrow, it will rain tomorrow. For Socrates may not know (believe) that he is wise.

Call someone 'modest' if he does not believe that he is wise; Socrates, we are supposing, is wise and modest. One way of being modest, call it 'super-modest', is believing that you are not wise. Most of us (all of us, I surmise) are super-modest: there are things we profess to believe but not to know. Socrates, if wise, cannot be super-modest. The belief that you are not wise is a self-fulfilling belief. If this belief is true, your beliefs don't coincide with your knowledge. And if it is false, your beliefs don't coincide with your knowledge. So Socrates must neither believe nor disbelieve that he is wise. In particular, we have Socrates neither believing nor disbelieving 'If I believe that it will rain tomorrow, it will rain tomorrow'. The reading of 'if' which concerns Williamson here is the truth-functional reading: his purpose is to demonstrate failure of substitutivity of coextensive operators in $O(Op \rightarrow p)$, where '$\rightarrow$' is truth-functional. So Socrates neither believes nor disbelieves what he could express by

(1) Either I don't believe that it will rain tomorrow, or it will rain tomorrow.

(To disbelieve (1) would saddle Socrates with Moore's paradox: if he disbelieves (1), he believes 'I believe that it will rain, and it won't

rain'.) Now Socrates either believes that it will rain tomorrow, or he does not. If he does, he has the wherewithal to deduce (1); if he does not, and realises he does not, he also has the wherewithal to deduce (1). So Socrates, though wise, must also be rather stupid, not to believe that if he believes it will rain tomorrow, it will rain tomorrow, on the truth-functional reading of that thought. (As Williamson points out, there is no inconsistency in this particular combination of wisdom and stupidity.) The example of failure of substitutivity would be more plausibly read: Socrates knows that necessarily, if he knows that p then p; but he doesn't believe that necessarily, if he believes that p then p. (This reading also fits the thought that he doesn't believe that he is not wise, but merely thinks he might not be wise.) This reading, however, is more complex than the form that Williamson wishes to illustrate. No matter: it began, and ends, a cartoon. Our next topic is more serious, technically and philosophically.

3. *Does Gödel's Theorem apply to me?*

Gödel's Second Incompleteness Theorem states that the consistency of Peano Arithmetic (PA) cannot be proved within PA, if PA is consistent. Since $\neg(0 = 1)$ is provable in PA, we may reformulate 'PA is consistent' as '$0 = 1$ is not provable in PA'. So, by Gödel's second theorem,

> if it is not provable in PA that $0 = 1$, it is not provable in PA that *it is not provable in PA that* $0 = 1$.

We have, once more, the iteration of a sentence operator, 'It is not provable in PA that'.

The italicised phrase above is to be understood thus. Each formula α of the language of PA, L_{PA}, is coded by a numeral, $\ulcorner\alpha\urcorner$, of the language of PA. Provability in PA is coded by a formula of L_{PA}, $\mathrm{Bew}(x)$, with one free variable. ('Bew' is an abbreviation of *beweisbar*, the German for 'provable', and hence pronounced 'bev', not 'bue'.) $\mathrm{Bew}(\ulcorner\alpha\urcorner)$ is true iff there is a proof in PA of the formula with Gödel number α. The Second Incompleteness Theorem says that if it is not provable in PA that $0 = 1$, then it is not provable in PA that $\neg\mathrm{Bew}(\ulcorner 0 = 1\urcorner)$.

Williamson points out that Bew is not unique in being a provability operator for PA representable in PA. (O is a provability operator for PA just in case for each sentence α of L_{PA}, $O\alpha$ is true if and only if α is

provable in PA.) Any two provability operators for PA are coextensive: they are true of the same formulas. But it does not follow that Gödel's second theorem can be carried out using any arbitrary provability operator. This is initially surprising, but we have been prepared for it by the failure of substitutivity of coextensive operators in embedded contexts. Although these operators are coextensive, they may not be provably equivalent in PA.

For example, let 'Bew*($\ulcorner\alpha\urcorner$)' be 'Bew($\ulcorner\alpha\urcorner$) $\wedge$ α'. Bew* is coextensive with Bew. But while it is not provable in PA (if PA is consistent) that $\neg$Bew($\ulcorner 0 = 1\urcorner$), it is provable in PA that $\neg$Bew*($\ulcorner 0 = 1\urcorner$), that is, it is provable in PA that $\neg$(Bew($\ulcorner 0 = 1\urcorner$) $\wedge$ (0 = 1)). For it is provable in PA that $\neg$(0 = 1); and it is trivial to prove $\neg(P \wedge Q)$ from $\neg Q$. Williamson mentions that the existence of 'deviant' provability operators has long been recognized, and led to an investigation of the question: which principles must a provability operator satisfy to be usable in Gödel's proof? George Boolos (1993) discusses questions such as this. Like Williamson, Boolos proceeds by translating principles governing provability operators into modal languages. His Chapter 3 is entitled 'The box as Bew(x)'. As I understand it, Williamson's investigation is a generalization of this idea to different areas.

Williamson suggests that the existence of deviant provability operators casts doubt on attempts to use Gödel's second theorem to show that humans are not Turing machines. Grant for the sake of argument that a sense has been given to 'I am a Turing machine' from which it follows that the set of sentences to which I assent is recursively enumerable. The anti-mechanistic argument he considers goes roughly as follows. Suppose I am a Turing machine. Call the sentences to which I assent 'my system'. My system can be recursively axiomatized and Gödel's Theorem applies to it. Suppose I assent to the consistency of my system. It would appear to follow that my system is thereby inconsistent, for any consistent system to which Gödel's Theorem applies cannot include a statement of its own consistency. Perhaps it is extravagant to suppose that my system is consistent. Nevertheless, it is paradoxical that it should follow from my claim to consistency that I am inconsistent.

The fallacy in such an argument, according to Williamson, is that it ignores the *mode of presentation under which I assent to the consistency of my system* (p. 94). The operators: 'It is provable in the system with such-and-such properties that' and 'It is provable by me that' may be extensionally equivalent. But, we have seen, not any two extensionally equivalent provability operators 'satisfy the same principles'—

embed in the same way—and if Gödel's result is provable by manipulation of one of them, it doesn't follow that his result is provable by manipulation of the other. Suppose my system is the *F* and that I can prove that I cannot prove that 0 = 1. Then the *F* can prove that I cannot prove that 0 = 1. It doesn't follow that the *F* can prove that the *F* cannot prove that 0 = 1.

There are two separable issues here. First, the existence of deviant provability operators, with which the theorem does not go through, does indeed show that we cannot freely substitute the coextensive 'I can prove that' and 'the *F* can prove that' in embedded contexts, without further investigation. Second, there are (as we saw from the first example) special problems about the nature of self-ascription, which may be relevant to this case, independently of the issue of embedded operators.

Gödel's first theorem works by the construction of a sentence *G* such that it is provable in PA that $G \leftrightarrow \neg \mathrm{Bew}(\ulcorner G \urcorner)$. Informally, *G* is equivalent to a statement of its own unprovability in PA. It follows that if PA is consistent, neither *G* nor $\neg G$ is provable in PA, and hence *G* is true. (For the second theorem, we represent the consistency of PA by a formula of PA, *Con*; the reasoning used in the proof of the first theorem is expressed and carried out within PA, so we prove within PA that $(Con \rightarrow G)$. But if PA is consistent we cannot prove *G*, by the first theorem; so, if PA is consistent, we cannot prove *Con*, for if we could, we could prove *G*).

Gödel's reasoning makes use of the contrast and interplay between reasoning within a system and reasoning about a system. If a reasoner identifies themself (cognitively speaking) entirely with the system under study, they do not have available the 'outside view' for reasoning about that system. If I try to perform Gödel's reasoning not about a system such as PA, but about my system, nonsense (or rather, a version of the Liar paradox) ensues. Consider the first theorem; substitute 'I' for 'PA'. Suppose I have constructed a *G* such that I can prove that ($G \leftrightarrow$ I can't prove that *G*). From the supposition that I can prove that *G*, it follows that I can prove that I can't prove that *G*; and hence that I can't prove that *G*. The foregoing is a proof that I can't prove that *G*. But then it is a proof that *G*! It seems to follow that I can't construct such a *G*.

J. R. Lucas's anti-mechanistic argument does not fall foul of the possibility of different modes of representation of a system. He insists that the mechanist come up with a definite specification of the alleged machine, to get the argument started. In his first article he presents a challenge:

> It is like a game. The mechanist has first turn. He produces *a—any*, but only a *definite one*—mechanical model of the mind. I point to something that it cannot do, but the mind can. The mechanist is free to modify his example, but each time he does so, I am entitled to look for defects in the revised model. If the mechanist can devise a model that I cannot find fault with, his thesis is established; if he cannot, then it is not proven; and since—as it turns out—he necessarily cannot, it is refuted. (1961, p. 118, his emphasis)

He reiterates this point in a reply to critics:

> The argument is a dialectical one. It is not a direct proof that the mind is something more than a machine, but a schema of disproof for any particular version of mechanism put forward. *If* the mechanist maintains any specific thesis, I show that a contradiction ensues. But only if. It depends on the mechanist making the first move and putting forward his claim for inspection. (1968, pp. 145–6, his emphasis)

That is, Lucas only claims to have an argument that he, Lucas, is not machine M, given a relevant, adequate specification of M.

David Lewis (1979b) replies as follows. In insisting on the dialectical character of the argument, Lucas is insisting that his output depends on his input: what sentence he will produce as disproof depends upon what hypothesis the mechanist puts forward. (He has agreed that he cannot, once and for all, disprove all such hypotheses, but claims that he can disprove any specific one which is presented.) Hence we must distinguish between Lucas's output when he is not being accused of being any particular machine; and his output when accused of being machine M. Let O_L be Lucas's arithmetical output when not accused of being any particular machine, and $O_{L,M}$ be his arithmetical output when accused of being a particular machine M. $O_{L,M}$ is O_L plus a sentence $\phi(M)$, a Gödel-sentence expressing the consistency of M's arithmetical output. But M's output under what conditions? If Lucas is a machine, then he is a machine whose output depends on its input. We must similarly distinguish between O_M, the output of machine M when it is not accused of being a particular machine, and $O_{M,M}$, the output of machine M when it is accused of being machine M. Suppose $O_L = O_M$, and $\phi_1(M)$ expresses the consistency of the system which generates O_M; and, on accused of being M, Lucas produces the sentence $\phi_1(M)$. This we can accept. But $\phi_1(M)$ does not express the consistency of the extended system $O_{M,M}$, which, if Lucas is M, represents the system that characterises his new output, after the accusation. If Lucas, accused of being M, instead produces a

sentence $\phi_2(M)$ expressing the consistency, not of O_M, but of $O_{M,M}$, his new extended system, there is no reason to think $\phi_2(M)$ is true.

Lucas's anti-mechanistic argument, then, while not subject to the fallacy Williamson diagnoses, is nevertheless questionable at best.

4. *Modal bicycles*

Do metaphysical necessity and possibility satisfy the principles of iteration known as S4, $\Box p \rightarrow \Box\Box p$ (equivalently, $\Diamond\Diamond p \rightarrow \Diamond p$), and S5, $\Diamond p \rightarrow \Box\Diamond p$ (equivalently, $\Diamond\Box p \rightarrow \Box p$)? Hugh Chandler (1976) and Nathan Salmon (1989) think not. Salmon's diagnosis of the opinion to the contrary is that we confuse the necessary and the actually necessary, the possible and the actually possible. If Salmon is correct, we have another putative instance of Williamson's phenomenon: coextensive operators which satisfy different principles.

Take an artefact—in Chandler's example, a particular bicycle, B, constructed, as bicycles are, out of many components. Had B been constructed with one different component—one different hand grip, say—it would still have been the same bicycle, B would still have existed. A bicycle constructed out of sufficiently many different components, or in the extreme, all different components, however, would not have been B, but a different bicycle. Now think of a series of possible worlds, starting from the actual world with the actual B, and such that neighbouring worlds have sufficiently similarly constituted bicycles to count as the same bicycle. Let D_1 and D_2 be descriptions of the construction of bicycles at w_1 and w_2 respectively, such that B could have been D_1, B could not have been D_2, but the bicycle in w_1 which is D_1 could have been D_2. We have a contradiction: B could not have been D_2 (when we think of B from the standpoint of the actual world); but B could have been D_2 (when we think of B from the standpoint of w_1). Chandler's and Salmon's solution to the paradox is to relativize possibility to worlds, and hence to deny S4 and S5. From the point of view of the actual world, there is no possible world in which B is D_2. But from the point of view of the actual world, there is a possible world (w_1) at which there is a possible world (w_2) at which B is D_2. The possibly possible outstrips the possible. S4 fails. Something can be possible from the standpoint of one world, but not possible from the standpoint of another: S5 also fails.

David Lewis claims that this is a kind of double-speak. We, in the

actual world, know all the relevant facts about w_2: we have specified them. We allow that B's being D_2 is possible from the standpoint of some possible world, but say w_2 is an impossible world from the standpoint of ours. 'I think this is like saying: there are things such that, ignoring them, there are no such things. Ignoring worlds where such-and-such obnoxious things happen, it is impossible that such things happen. Yes. Small comfort' (1986, p. 248). Lewis, for whom there is no trans-world identity, but counterpart relations based on relevant similarities, explains the phenomenon in terms of the non-transitivity of the counterpart relation. Salmon accuses adherents to S4 and S5 of 'a narrow-minded form of modal ethnocentrism' (1982, p. 239): the view that all the possibilities there are possibilities for us. They confuse the necessary with the actually necessary, the possible with the actually possible.

The introduction of the operator 'actually', @, into modal logic is well motivated. It is an aid to the expression of some modal thoughts. It might seem like a redundant word: actually p if and only if p. But it is not redundant when embedded in modal contexts. Six races were run at a meeting. Each had a close finish. You want to express the thought that it could have been the case that all the horses which (in their respective races) actually finished second, won. You don't mean that there is a possible situation in which: the horses which finished second, won. Nor do you merely mean that for any horse which finished second, there is a possible situation in which it won: that does not entail that there is a possible situation in which they all won. (Take a single race in which all horses finished very close. About that race, you might say, of any horse, that it could have won; but not that they all could have won.) You mean that there is a possible situation in which: all the horses which finished second in the actual situation, won. Whether a horse finished second at w, depends on how things are at w. Whether a horse actually finished second, at w, depends on how things are at the actual world. That is how 'actually' increases the expressive power of modal language. (Compare: 'At some time in the future, all those who are now research students will be professors'. Whether someone is a research student, at some time in the future, depends on how things are then; whether someone is now a research student, at some time in the future, depends on how things are now.) The truth condition for 'Actually p' is: '$@p$' is true at w iff p is true at the actual world.

If Salmon is right, then, $\Box p$ iff $@\Box p$ and $\Diamond p$ iff $@\Diamond p$; $@\Box p \rightarrow @\Box @\Box p$ and equivalently $@\Diamond @\Diamond p \rightarrow @\Diamond p$; but it is not the case that

$\Box p \rightarrow \Box\Box p$ (equivalently, $\Diamond\Diamond p \rightarrow \Diamond p$). Similarly for S5. The coextensive operators embed differently.

Williamson himself, however, has given reasons for rejecting Salmon's response to the paradox (1990, pp. 126–37). First, Williamson argues, the paradox is essentially the same when there is no actual bicycle *B*, but a series of possible worlds each with their possible bicycles, with very small differences between adjacent members, and all these worlds are equally possible from the point of view of the actual world. (We specify these bicycles in terms of the components out of which they would have been constructed: we are in a bicycle factory laden with bicycle parts.) We still have a paradox of identity between possible bicycles, although each is possible from the point of view of the actual world. Second, he notes that similar problems can arise with variation over time. An artist adds another line or two to their drawing; continuing the process, at what point do we no longer have the original drawing? We are not likely to react to this puzzle by renouncing the parallel of S4 for time: if there is a time at which there is a time at which *p*, there is a time at which *p*. The problem we face, in each case, is the Sorites, a vivid highlighting of the problem of vagueness. Its solution should flow from a general solution to the problem of vagueness, rather than a revision of modal logic. Thus, although the formal possibility of Williamson's phenomenon has been demonstrated, it is dubious whether this example provides a good philosophical reason to be interested in it.

5. *A rough guide to the rest*

We have an interpreted language L, whose operators include the truth-functional ones, and whose sentences are evaluated as true or false at indices (e.g. worlds). O_1 and O_2 are sentence operators in L. O_1 and O_2 are *coextensive* if and only if, for every sentence α of L, and every index *i*, $O_1\alpha$ is true at *i* if and only if $O_2\alpha$ is true at *i*. (In some applications, we may be interested in interpretations in which there is just one index.)

Williamson finds it convenient to state *principles* in a modal language, $L_\Box$. A principle is a formula of $L_\Box$. We define an O-translation, a mapping * of formulas of $L_\Box$ to sentences of L which maps $\Box$ to O. O satisfies a principle α if and only if its mapping is true at every index for every O-translation. In his §2 Williamson says

> The choice of $L_{\Box}$ . . . is purely notational: it does not assume the operator O to have anything in common with a necessity operator beyond being a one-place sentential operator. The modal language has the advantage of familiarity; many of its formulas and sets of formulas have names.

There is also a historical precedent for this approach. As mentioned above, provability has been studied in this way (Boolos 1993). There is a more obvious reason for this approach in the case of provability: provability is a kind of necessity. Williamson's ensuing discussion in his §3 shows that it is very tricky to formulate general questions about the phenomenon correctly in this translational mode. I tentatively wonder whether this dog-leg approach is the best one. The question seems clear enough without it. We have defined what it is for two operators in L to be coextensive. Under what conditions do two such operators satisfy different principles in L? That is, when can we have two formulas of L which differ only in that O_2 occurs in the second wherever O_1 occurs in the first, such that one is true at all indices but the other is not?

Williamson shows in his §2 that the principles over which coextensive operators can differ always involve some self-embedding of $\Box$. A *first-degree* principle is a formula in which no occurrence of $\Box$ lies within the scope of another. Coextensive operators satisfy exactly the same first-degree principles.

He also shows that coextensive *truth-functional* operators satisfy the same principles; and the same is true of a wider class, of *extensional* operators. O is extensional just in case for all sentences α and β, and all indices i, if the truth value of α at i is the truth value of β at i, then the truth value of $O\alpha$ at i is the truth value of $O\beta$ at i; that is equivalent to saying that O satisfies the principle $(p \leftrightarrow q) \rightarrow (\Box p \leftrightarrow \Box q)$.

He points out that the phenomenon of coextensiveness without satisfaction of the same principles could be generalized from singulary operators to binary or n-ary ones, although we should no longer be able to investigate the principles by mapping the operators to $\Box$. He mentions conditionals in this connection. I explore the case of conditionals in §6 below.

How widespread is the phenomenon of coextensiveness without satisfaction of the same principles, Williamson asks in his §3. Given the principles that two operators satisfy, when is it consistent to suppose that they are coextensive? And: given two distinct modal systems S_1 and S_2, when is there a language L with an index set I and coextensive operators O_1 and O_2 such that O_1 satisfies all and only the principles of

S_1 and O_2 satisfies all and only the principles of S_2? Much effort is expended to find the most fruitful formulation of this question. He does not answer the question at this level of generality. In his final two sections, special forms of the phenomenon, which arose in the initial examples, are discussed in greater depth.

His fourth section concerns a generalization of the device used to construct the deviant provability operator Bew*($\ulcorner\alpha\urcorner$), which is Bew($\ulcorner\alpha\urcorner$) $\wedge$ α. Suppose that for each index *i*, Oα is true at *i* only if α is true at *i*. (For example, let O be 'Socrates believes that', where Socrates has only true beliefs at all indices.) Define a new operator O^+ such that $O^+\alpha = O\alpha \wedge \alpha$. O and O^+ are coextensive. This, Williamson shows, can be one general source of his phenomenon.

His final section returns to S4 and S5,

$$4 \quad \Box p \rightarrow \Box\Box p$$
$$5 \quad \neg\Box p \rightarrow \Box\neg\Box p.$$

If $\Box$ is interpreted epistemically, these are appropriately called 'principles of introspection': if you know, you know that you know; if you don't know, you know that you don't know; similarly for belief. Take an operator O which satisfies neither 4 nor 5. Is there an operator coextensive with O which does satisfy 4 and 5? Recall the dispute about metaphysical necessity: according to Salmon, his opponents confuse the necessary, which does not satisfy 4 and 5, with the actually necessary, which does satisfy 4 and 5, despite these two operators being coextensive. Formally, we can generalize this idea. We introduce an operator @, formally like 'actually', such that @O satisfies 4 and 5 even if O does not. (Williamson uses the plain 'actually' symbol @ for the new operator, but I need to distinguish the two notationally.) O and @O will be coextensive, but satisfy different principles. In his final paragraph, Williamson presents this as a worry. Suppose you have a theory of propositional attitudes, for instance a functionalist theory, which gives a method for fixing the extension of '*x* believes that *p*', '*x* desires that *p*' etc, at all possible worlds. This will be insufficient to determine how the attitudes embed: which principles of iteration they satisfy. 'An account of the operators "*x* believes that" and "*x* desires that" must be based on considerations fine-grained enough to discriminate between them and operators merely coextensive [in all possible worlds] with them'.

Return to the semantics of 'actually': @*p* is true at *w* if and only if *p* is true at the actual world. So if *p* is true at the actual world, @*p* is true

at every world, and hence $\Box @p$. (This is the pathological case of a necessary truth which is knowable only a posteriori. $\Box$ becomes uninterestingly undiscriminating, when applied to propositions prefaced by 'actually'. There is a more interesting necessity operator in languages with 'actually', the 'fixedly actual', the things that remain actually true no matter which world is taken as the actual world. See Davies and Humberstone (1980). Despite the curiosity of $\Box @$, a purpose is served by having @ in a modal language, as we saw in §4 above.)

Williamson specifies the semantics for @ thus. Take a language L_O and introduce @, extending L_O to $L_{O@}$. If α is true at i in L_O, $@\alpha$ is some specific tautology $\top$ at i in $L_{O@}$, say 'If it is raining it is raining'; and if α is false at i in L_O, $@\alpha$ is the negation of $\top$ at i in the extended language. Thus any proposition $@\alpha$ is true at all indices or none, just as with @.

Let O be 'John knows that'. Suppose O does not satisfy 4: it is not true at all indices that $O\alpha \rightarrow OO\alpha$. Now suppose $O\alpha$ is true at i. Then, by the truth condition for @, $@O\alpha$ is $\top$ at i. At i, Jones knows that $\top$ (i.e. he knows that if it is raining it is raining). So $O@O\alpha$ is true at i. Hence $@O@O\alpha$ is also $\top$, and hence true at i. Thus we have proved the S4 principle for the operator '@ John knows that'. Further, at any index i, for any proposition α, 'John knows that α' is true at i if and only if '@ John knows that α' is true at i. These operators are coextensive at all indices, yet satisfy different principles.

Now @ is a mighty strange operator, as Williamson points out. We must not confuse it with the modal 'actually' of which it is a merely formal analogue. Whatever John knows, he actually knows, and vice versa; and if S4 fails for 'John knows that', it fails for 'John actually knows that'. John knows that p if and only if John knows that actually p. But for any true p such that John doesn't know that p, John knows that $@p$ (for if p is true, $@p$ is $\top$, and John knows that $\top$). And for any false p such that John doesn't know that $\neg p$, John knows that $@\neg p$ (for if p is false, $\neg p$ is true, $@\neg p$ is $\top$; and John knows that $\top$). 'There is no obvious way of rendering @ in English', says Williamson (p. 118). This is putting it mildly. We have a formal device with no conceivable use. Let theorists of propositional attitudes, functionalists or others, claim to have a theory capable of fixing the extension of the attitudes in all possible situations. Williamson objects that this is insufficient, for the attitudes might still satisfy different principles of iteration, and demonstrates this with the @ trick. The theorists can rightly respond that if that is all they have to worry about, they are home free: they know that

natural language and thought could have no use for this device. More generally: no principles about iterated attitudes are logical truths. Special cases of (for instance) believing that you believe, if you believe, or desiring that you desire, if you desire, need to be vindicated as such, in terms of one's theory of attitudes. The formal possibility of inadequacy raised by Williamson's phenomenon seems easily closed off. Which principles of iteration do propositional attitudes satisfy as a matter of logic (broadly construed)? None.

6. *Iterating conditionals*

Are there interesting examples of this phenomenon for binary operators? Something close to it arises in a genuine dispute about iteration of conditionals. Vann McGee (1989) proposed a semantics which disagrees with Robert Stalnaker's (1968) only on iterated conditionals.

Stalnaker gave a possible-world semantics for conditionals, using the notion of a selection function, f, which selects, for any world w and proposition A, a world w'. f satisfies certain formal conditions which make it suitable to be construed informally as yielding the 'nearest' world to w in which A is true. Using Stalnaker's notation for his conditional connective, $A>B$ is true at w iff B is true at $f(w, A)$ (the nearest world to w in which A is true). According to Stalnaker, this provides the form of a semantic framework for both indicative and counterfactual conditionals. Lewis's theory of counterfactual conditionals (1973) is similar in the respects which concern us here.

Now 'If A, then if B, C' and 'If A and B, then C' are, intuitively, equivalent. Let A be 'it will rain or snow tomorrow' ($R \vee S$), let B be 'it won't rain tomorrow' ($\neg R$), and let C be 'it will snow tomorrow' (S). Consider:

(1) If it rains or snows tomorrow, and it doesn't rain tomorrow, it will snow tomorrow.

(1) is indisputable. Consider:

(2) If it rains or snows tomorrow, then if it doesn't rain, it will snow.

On a natural reading, (2) is equivalent to (1). But on Stalnaker's semantics (2) can be false and the following (3) can be true, even though (1) is true:

(3) If it rains or snows tomorrow, then if it doesn't rain, it won't snow.

To evaluate (1), we go to the nearest world in which $(R \vee S) \,\&\, \neg R$, and we find that in that world, inevitably, S. To evaluate (2), we go to the nearest world in which $(R \vee S)$, and evaluate $(\neg R > S)$ there. Suppose that in the actual world it neither rains nor snows, and in all close worlds in which it rains or snows, it rains but doesn't snow. Let the nearest world to the actual world in which it rains or snows be w. At w, it rains. In the nearest world to w in which it doesn't rain, it doesn't snow either. In such circumstances, (3) is true and (2) is false on Stalnaker's semantics, while (1) is true. This is obviously wrong. The point is also persuasive for counterfactuals: 'If it had rained or snowed, and not rained, it would have snowed' is agreed to be trivially true. 'If it had rained or snowed, then if it hadn't rained it would have snowed' sounds equivalent, but may well be false on Stalnaker's and Lewis's semantics.

McGee's semantics for indicative conditionals coincides with Stalnaker's except on this point: he preserves the equivalence of 'If A & B then C' and 'If A, then if B then C'. McGee and Stalnaker agree in counting a conditional vacuously true when its antecedent is impossible. McGee sets aside conditionals with conditional antecedents: these 'are used and understood by English speakers, but they occur sufficiently rarely that it is hard to gather enough examples to get a fix on what is going on with them' (p. 486). Hence we shall compare McGee's and Stalnaker's semantics as applied to sentences without conditional antecedents. Write $\Rightarrow$ for McGee's conditional. Capital letters A, B, etc, range over non-conditional sentences. Greek letters range over all sentences.

McGee defines a *Stalnaker model* as an ordered triple <W, I, f >, where W is a set of worlds, I is an interpretation which assigns to each pair $<w, A>$, where w is a world and A is an atomic sentence, a truth value either T or F, and f is a selection function. (This is a simplified version of a Stalnaker model, which is all he needs to define truth at the actual world. We shall later generalize it to define truth at a world w.)

McGee defines, by recursion, 'true under the hypothesis that A'. This notion is motivated by his fundamental conception of indicative conditionals:

> Reasoning with conditionals is hypothetical reasoning. We decide whether to accept a conditional by assuming the antecedent as an hypothesis and, having done so, seeing whether we are inclined to accept or reject the consequent. If

> we think of a possible world as an epistemically possible state of affairs, we can use possible worlds semantics to give a formal model of hypothetical reasoning. In assuming A as an hypothesis, we agree that, until the hypothesis has been discharged, we shall treat as epistemically impossible any world in which A is false. If, having assumed A as an hypothesis, we have occasion to assess the truth value of a conditional $B \Rightarrow \phi$, we do so by assuming B as a further hypothesis, so that we only treat as epistemically possible worlds in which A and B are both true. (1989, p. 514)

Here are his recursive clauses (pp. 514–5):

> 1. If ϕ is atomic, ϕ is true under the hypothesis that A iff I $<f(A),\phi>$ is true;
> 2. $(\phi \,\&\, \psi)$ is true under the hypothesis that A iff ϕ and ψ are both true under the hypothesis that A;
> 3. $(\phi \vee \psi)$ is true under the hypothesis that A iff ϕ is true under the hypothesis that A or ψ is true under the hypothesis that A;
> 4. $\neg\phi$ is true under the hypothesis that A iff ϕ is not true under the hypothesis that A;
> 5. $B \Rightarrow \phi$ is true under the hypothesis that A iff ϕ is true under the hypothesis that $A \,\&\, B$. (This is the crucial clause.)
>
> Finally, ϕ is true iff ϕ is true under the hypothesis that $\top$ (where $\top$ is a tautology).

For sentences over which each conditional is defined, $A>B$ is true if and only if $A \Rightarrow B$ is true. (Recall that A and B are factual propositions, i.e. do not contain $\Rightarrow$; conditionals in antecedents have been set aside; and McGee has not defined what it is for $B>C$ to be true under the hypothesis A.) But they behave differently on iteration: $A \Rightarrow (B \Rightarrow C)$ is equivalent to $(A \,\&\, B) \Rightarrow C$, which is equivalent to $(A \,\&\, B) > C$; the latter is not equivalent to $A > (B > C)$, as we saw above. We can generalize McGee's semantics to give truth conditions at any world w. Then, at any w, $A>B$ is true at w iff $A \Rightarrow B$ is true at w; but $A > (B > C)$ is not equivalent to $A \Rightarrow (B \Rightarrow C)$.

We do not yet have the Williamson phenomenon in full generality. The phenomenon applies to pairs of operators each of which ranges over sentences containing the other. Now $\Rightarrow$ has been proposed as a correction to $>$: they are very similar, but differ at a crucial point. There is no obvious interest in sentences containing both conditionals. Still, let us try to address the question. Consider

(a) $A > (B > C)$
(b) $A \Rightarrow (B > C)$
(c) $A \Rightarrow (B \Rightarrow C)$
(d) $A > (B \Rightarrow C)$.

We know that (a) and (c) are not equivalent. If we can give a semantics which conservatively extends both Stalnaker's and McGee's for sentences like (b) and (d) which contain both conditionals, the Williamson phenomenon applies only if (a) is equivalent to (b), and (c) is equivalent to (d): if $>$ and $\Rightarrow$ are fully coextensive, any two sentences which differ only in that one contains an occurrence of $>$ of widest scope where the other contains $\Rightarrow$, should be equivalent.

Such a semantics has not been constructed, and the construction is probably of only technical interest: as mentioned above, it is unlikely that we have a use for sentences containing both conditionals. To evaluate (b) we need to specify what it is for $B>C$ to be true under the hypothesis that A. To evaluate (d) we need to specify what it is for C to be true under the hypothesis that B, at a world, $f(A)$, the nearest A-world.

Let us generalize McGee's semantics to yield truth under the hypothesis that A at a world w, and add a sixth clause specifying what it is for $B>C$ to be true under A at w. (Here and henceforth, I abbreviate 'under the hypothesis that' to 'under'). I shall drop mention of an interpretation, taking this to be fixed. I assume that every atomic sentence has a truth value at every world.

1. If ϕ is atomic, ϕ is true under A at w iff ϕ is true at $f(w, A)$.
2. $(\phi \,\&\, \psi)$ is true under A at w iff both ϕ and ψ are true under A at w.
3. $(\phi \vee \psi)$ is true under A at w iff ϕ is true under A at w or ψ is true under A at w.
4. $\neg\phi$ is true under A at w iff ϕ is not true under A at w.
5. $(B \Rightarrow \phi)$ is true under A at w iff ϕ is true under $A \,\&\, B$ at w.
6. $(B > \phi)$ is true under A at w iff ϕ is true under B at $f(w, A)$, i.e. the nearest A-world to w.

Finally, ϕ is true at w iff ϕ is true under $\top$ at w.

As this semantics is a straight generalization of McGee's, it clearly agrees with his on sentences not containing $>$. I shall show that it also agrees with Stalnaker's semantics for sentences not containing $\Rightarrow$ (and so conservatively extends both semantics). First, a lemma:

(*) For any sentence ϕ which does not contain $\Rightarrow$, ϕ is true under A at w iff ϕ is true at $f(w, A)$.

Proof: (*) holds when ϕ is atomic, by 1. And if (*) holds for ϕ and ψ, it holds for $\phi \,\&\, \psi$, $\phi \vee \psi$ and $\neg\phi$, by clauses 2–4. For instance, $\neg\phi$ is true

under A at w iff ϕ is not true under A at w (by 4), iff ϕ is not true at $f(w, A)$ (by the assumption that (*) holds for ϕ), iff $\neg\phi$ is true at $f(w, A)$. It remains to show that (*) holds when ϕ is of the form $B > \psi$. By 6, $B > \psi$ is true under A at w iff ψ is true under B at $f(w, A)$; and by 6 again, $B > \psi$ is true under $\top$ (i.e. true) at $f(w, A)$ iff ψ is true under B at $f(f(w, A),\top)$, i.e. at $f(w, A)$, for the nearest world to $f(w, A)$ in which a tautology $\top$ is true is $f(w, A)$ itself. The right-hand sides of these two biconditionals being identical, the left-hand sides are equivalent, i.e. $B > \psi$ is true under A at w iff $B > \psi$ is true at $f(w, A)$.

This completes our proof of (*). Now 6 implies that $A > \phi$ is true at w iff ϕ is true under A at w, which together with (*), implies, for any sentence ϕ which does not contain $\Rightarrow$, that $A > \phi$ is true at w iff ϕ is true at $f(w, A)$. This is Stalnaker's truth condition, and completes the proof that this is a conservative extension of Stalnaker's semantics.

Are $>$ and $\Rightarrow$ fully coextensive on this enriched semantics? Do we have, for any sentence ϕ, $A > \phi$ iff $A \Rightarrow \phi$? Yes. By 6, $(A > \phi)$ is true (under $\top$) at w iff ϕ is true under A at $f(w, \top)$, i.e. at w. By 5, $(A \Rightarrow \phi)$ is true (under $\top$) at w iff ϕ is true under $\top$ & A at w, i.e. under A at w. The right-hand sides of these biconditionals are identical. So $A > \phi$ is true at w iff $A \Rightarrow \phi$ is true at w. So (a) is equivalent to (b), and (c) to (d). But $>$ and $\Rightarrow$ don't iterate the same way: $A > (B > C)$ is not equivalent to $A \Rightarrow (B \Rightarrow C)$. The latter is true at w iff C is true at $f(w, A \,\&\, B)$; the former is true at w iff C is true at $f(f(w, B), A)$. These are different. So this is a genuine example of the Williamson phenomenon.

Another manifestation of the difference between the two conditionals on iteration is that $A \Rightarrow (B \Rightarrow C)$ is equivalent to $B \Rightarrow (A \Rightarrow C)$, but $A > (B > C)$ is not equivalent to $B > (A > C)$. Adapting an example of McGee's (1985), consider a time shortly before Reagan's first election. Reagan is favourite; a fellow-Republican, Anderson, is a complete outsider; Carter is well behind Reagan. Uncontroversially,

(1) If a Republican wins and Reagan does not win, Anderson will win.

According to McGee, we should also accept:

(2) If a Republican wins, then if Reagan does not win Anderson will win;

and

(3) If Reagan does not win, then if a Republican wins, Anderson will win.

(2) and (3) are false on Stalnaker's truth-conditions, and moreover, not equivalent. Instead, the following are true:

(2′) If a Republican wins, then if Reagan does not win Carter will win.

(Go to the nearest world in which a Republican wins—the actual world, as it happens. In the nearest world to it in which Reagan does not win, Carter wins.)

(3′) If Reagan does not win, then if a Republican wins, Reagan will win.

(Go to the nearest world in which Reagan does not win. In it, Carter wins. In the nearest world to it in which a Republican wins, Reagan wins.) These examples strengthen the suspicion that we do not have a natural use for the thoughts expressed by the iteration of Stalnaker's truth condition, and iterated conditionals are deciphered McGee's way.

I have said that sentences containing both conditionals are unlikely to be interesting. I would be wrong if (i) there are two distinct kinds of conditionals, helpfully or unhelpfully labelled 'indicative' and 'counterfactual', and (ii) the former iterate McGee's way, the latter Stalnaker's way. Even if (i) is true, I think (ii) is false. After the election in question, (3) becomes a counterfactual, and

> If Reagan had not won, then if a Republican had won, Anderson would have won

is acceptable while the following is not:

> If Reagan had not won, then if a Republican had won, Reagan would have won.

Despite their coextensiveness at all worlds, the most striking case of these two conditional connectives 'satisfying different principles' is this: modus ponens fails for conditionals with conditional consequents, if we iterate McGee's way. You can consistently accept (2) 'If a Republican wins, then if Reagan doesn't win, Anderson will win'; and accept its antecedent 'A Republican will win'; yet reject its consequent 'If Reagan doesn't win, Anderson will win' (McGee 1985). The lesson, in my view, is this. Modus ponens is unassailable when the

antecedent and consequent of the conditional premiss are genuine propositions, with truth conditions. Conditionals are not genuine propositions with truth conditions, and a conditional appearing as the consequent of another conditional does not behave as a proposition would as the consequent of a conditional. Sentences with embedded conditionals need to be translated, the best we can, to sentences without embedded conditionals; 'If A, then if B then C' is 'really' of the form 'If A & B then C'; and the inference (which is invalid) from this and A to the conclusion 'If B then C' is not an instance of modus ponens. (I discuss this and other problems about embedded conditionals in Edgington 1995, pp. 280–4.)

I shall consider finally a simpler example of a problem about iteration in the philosophy of conditionals. Frank Jackson (1987) argues that the indicative conditional 'If A, B' is true iff $A \supset B$: its truth conditions are the simple truth-functional ones. But there is more to meaning than truth conditions. For example, 'A and B' and 'A but B' have the same truth conditions, but 'but' and 'and' are not synonyms. They translate differently into other languages. Dictionaries inform us that 'but' signals a contrast between the sentences it conjoins. Similarly, on Jackson's view, 'If A, B' does not mean the same as $A \supset B$, i.e. 'Either not A, or B'. So, while we know how $\supset$ embeds, that leaves it an open question how 'if' embeds, if it embeds at all. Compositionality requires that the meaning of a complex sentence be a function of the meanings of its parts. But if the truth condition doesn't exhaust the meaning of a part, that leaves open how the further ingredient of the meaning will affect embeddings. Consider 'but'. The contrast which makes 'A but B' appropriate may be destroyed when this sentence is embedded in 'C, and A but B': 'The argument is valid but its premise is false'; 'The argument is a *reductio*, and the argument is valid but its premise is false'. The contrast is lost on embedding.

Jackson gives the extra ingredient of the meaning of 'if', over and above its truth condition, as a condition under which the conditional is *assertable*. This tells us nothing about how conditionals embed in contexts in which the conditional is not asserted, he insists (1987, p. 239). He makes *ad hoc* observations about how we understand embedded conditionals, when we do so, which are independent of his account of the meaning of 'if' when not embedded.

The extensional equivalence of 'if' and $\supset$, on Jackson's theory, is merely extensional equivalence when applied to sentences which do not contain 'if'. Hence, it is weaker than extensional equivalence in

Williamson's sense. It shows that there is a wider class of problems about embeddings; and also, that problems about embeddings can arise even for operators which are 'basically truth functional': truth functional over sentences which do not themselves contain the operator in question.

7. *Concluding observations*

Williamson's paper is innovative, stunningly ingenious and a technical *tour de force*. All his detailed work is (so far as I can see) beyond criticism. An entirely uncritical reply to a philosophical paper, however, would be almost an oxymoron: I shall end by mentioning some unease.

First, I feel I lack a unifying, illuminating, explanatory account of when and why the phenomenon arises. It is as if a new species of animal has been discovered, at a variety of locations and times. We have discovered some of its properties. But we have not yet been provided with an understanding of its underlying nature, or a satisfying account of why it appears where and when it does.

Secondly, we were introduced to a phenomenon 'whose neglect', according to Williamson, 'causes philosophical problems' (p. 85). But it seems to me that the philosophical problems which have arisen are largely independent of the phenomenon. Let us recapitulate. The first problem concerned *de se* knowledge, self-ascription of properties, what is reported in the natural reading of 'he' in the coreferential reading of 'John knows that he is clever'. It is a problem that arises whether we have iterated operators or not. The KK principle, I argued, should not be formulated as the iteration of a one-place operator. The second problem about wise Socrates was merely a toy model. The third was about Gödel's Theorem. It is undoubtedly important and interesting that there are deviant provability operators which cannot be used to prove the theorem. When we turn to the significance of Gödel's theorems for the philosophy of mind, while it is plausible that there is something important and interesting here about the difference between the first-person perspective and the 'outside view', this is independent of iterated operators. And Lucas's anti-mechanistic argument is carefully stated to be immune from the problem of the 'mode of presentation under which I assent to my system'. It has been rebutted, I think, on other grounds. The next philosophical issue was whether the existence of a modal Sorites gave us a reason to reject principles of iteration for

metaphysical necessity. Williamson himself has argued elsewhere, and I agree, that this is not the best solution to the modal Sorites, so we do not have a good philosophical reason to hold that 'It is necessary that' and 'It is actually necessary that' satisfy different principles of iteration, although this is a formal possibility. In his final section Williamson uses a formal analogue of 'actually' to show that, when we have epistemic operators which do not satisfy S4 and S5, there may be coextensive ones which do satisfy these principles. But the formal analogue of 'actually' is artificial in the extreme. It is not a serious threat to a theory of propositional attitudes, e.g. a functionalist one, that this bastard operator exists. More generally, it is not a threat to our theory of propositional attitudes that there exists the formal possibility of operators coextensive in all possible situations, which satisfy different principles of iteration: propositional attitudes do not satisfy *any* principles of iteration as a matter of logic. I then showed that there are problems about iteration of conditionals. These problems are not naturally addressed in the setting of Williamson's phenomenon. It may be that there is some substantial philosophical problem waiting to be illuminated by this machinery, but I do not think we yet know whether this is so.[1]

REFERENCES

Boolos, George 1993: *The Logic of Provability*. Cambridge: Cambridge University Press.

Castañeda, Hector-Neri 1968: 'On the Logic of Attributions of Self-Knowledge to Others'. *Journal of Philosophy* 65, pp. 439–456.

Chandler, Hugh 1976: 'Plantinga and the Contingently Possible'. *Analysis* 36, pp. 106–109.

Davies, Martin K. and Humberstone, Lloyd 1980: 'Two Notions of Necessity'. *Philosophical Studies* 38, pp. 1–30.

Edgington, Dorothy 1995: 'On Conditionals'. *Mind* 104, pp. 235–329.

Frege, Gottlob 1892: 'Uber Sinn und Bedeutung', translated as 'On Sense and Reference' in *Translations from the Philosophical Writings of Gottlob Frege*, edited by Peter Geach and Max Black. Oxford: Basil Blackwell, 1966. (Page references to this volume.)

Harper, W. L., Stalnaker, R. and Pearce, G. eds. 1981: *Ifs*. Dordrecht: Reidel.

Jackson, Frank 1987: *Conditionals*. Oxford: Blackwell.

Kripke, Saul 1972: *Naming and Necessity*. Oxford: Blackwell.

[1] I am grateful to Keith Hossack and Timothy Williamson for comments.

Lewis, David 1973: *Counterfactuals*. Oxford: Blackwell.
—— 1979a: 'Attitudes *De Dicto* and *De Se*'. *Philosophical Review* 88, pp. 513–543. Reprinted in his 1983, pp. 133–156. Page references to this volume.
—— 1979b: 'Lucas Against Mechanism II'. *Canadian Journal of Philosophy* 9, pp. 373–376.
—— 1983: *Philosophical Papers* vol. I. Oxford: Oxford University Press.
—— 1986: *On the Plurality of Worlds*. Oxford: Blackwell.
Lucas, J. R. 1961: 'Minds, Machines and Gödel'. *Philosophy* 36, pp. 112–127.
—— 1968: 'Satan Stultified: A Rejoinder to Paul Benacerraf'. *The Monist* 52, pp. 145–158.
McGee, Vann 1985: 'A Counterexample to Modus Ponens'. *Journal of Philosophy* 82, pp. 462–471.
—— 1989: 'Conditional Probabilities and Compounds of Conditionals'. *Philosophical Review* 98, pp. 485–542.
Perry, John 1977: 'Frege on Demonstratives'. *Philosophical Review* 86, pp. 474–497.
Salmon, Nathan 1982: *Reference and Essence*. Oxford: Blackwell.
—— 1989: 'The Logic of What Might Have Been'. *Philosophical Review* 98, pp. 3–34.
Stalnaker, Robert 1968: 'A Theory of Conditionals', in Rescher, N. ed. *American Philosophical Quarterly* Monograph No. 2, 1968, reprinted in Harper, Stalnaker and Pearce 1981, pp. 41–55.
Williamson, Timothy 1990: *Identity and Discrimination*. Oxford: Blackwell.