WHAT CAN WE LEARN FROM THE PARADOXES?
Part I

J. L. MACKIE
University College,
Oxford

There is a group of paradoxes, some well known, other similiar ones less well known, which includes the Epimenides and other forms of the liar, heterologicality, Russell's class paradox, Richard's paradox, and so on. They might be referred to as paradoxes of self-reference, but that would beg several questions. A number of views about such paradoxes are fairly widely held, though not, fortunately, all by the same people, since they are not all compatible with one another. These are that:

(1) These paradoxes are unimportant and easily dismissed, since they arise from trivial misuses of language, such as ambiguity or lack of meaning.

(2) These paradoxes show that self-reference is logically and/or linguistically improper.

(3) These paradoxes are of less importance for general philosophy and the philosophy of language than for the foundations of mathematics and, in particular, set theory; they compel us to revise our naive concept of a set or class.

(4) Some of these paradoxes are of great importance for the philosophy of language, and compel us to distinguish sharply between an object language, a meta-language, a meta-meta-language, and so on. They show that no consistent language can be semantically closed.

(5) These paradoxes can be solved and/or resolved by valid proofs in formal logic.

(6) These paradoxes show that the habits of thought and speech which we find natural require revision in some way;

the only problem is the technical one of selecting the least inconvenient reform.

(7) These paradoxes divide into two radically different groups, semantic or linguistic on the one hand and syntactic or logico-mathematical on the other.[1]

I think that all these views are mistaken, though some, perhaps, are more mistaken than others.

We can dispose fairly easily of the suggestion that these paradoxes arise from, or can be resolved by the exposure of, meaninglessness or ambiguity. Take a version of the simple liar, the utterance 'What I am now saying is false'. There could indeed be uncertainty about the reference of the phrase 'What I am now saying': it might refer to something else in the speaker's recent contributions to the conversation, or to this utterance itself. But the paradox is founded not on this ambiguity but upon one way of removing it. If this phrase refers to something else in the neighbourhood, there is no paradox; there is a paradox just when this phrase is construed as referring unambiguously to the utterance of which it is a part. And so with all the paradoxes of this group: what we may call the *reasoning within the paradox*, the arguments which lead to a contradictory conclusion, requires a precise and unitary way of construing the linguistic components. Nor can the charge of meaninglessness be sustained. A verificationist might point out that the paradox formulae are not empirically verifiable; but if he allows meaning to formulae which are not empirically testable but analytic, he must concede that, for example, 'What I am

[1] View (1) is met more in conversation than in the literature. Examples of most of the others are found in e.g., for (2) A. Ross, 'Self-Reference and a Puzzle in Constitutional Law', *Mind* LXXVIII (1969), pp. 1-24, of the second part of (3), W. V. Quine *Set Theory and its Logic*, p. 5, and *The Ways of Paradox*, pp. 6-18, for (4) A. Tarski, 'The Semantic Conception of Truth' e.g. in *Readings in Philosophical Analysis*, ed. Feigl and Sellars, pp. 32-84, also W. and M. Kneale, *The Development of Logic*, pp. 665-6, for (5) J. F. Thomson, 'On Some Paradoxes', in *Analytical Philosophy* (1st series) ed. R. J. Butler, pp. 104-119, for (6), W. V. Quine, *The Ways of Paradox*, p. 18, for (7) F. P. Ramsey, *The Foundations of Mathematics*, pp. 20-21, and Quine, *Set Theory and its Logics*, p. 255.

now saying is false' is all too meaningful. It is both analytically true and analytically false. Exactly the same sort of calculation that establishes ordinary analytic truths will show that if this expression is false, it is true, and hence that it *is* true, and also that if it is true, it is false, and hence that it *is* false. In any case, it is more plausible to adopt a constructive theory of meaning, and then the paradox formulae are undeniably meaningful. 'What I am now saying is false', for example, is put together out of words used with their standard meanings in a grammatically correct way, the referring phrase has something to refer to, and the predicate 'false' has an item of the correct category to which to apply. Someone might indeed dispute the last two points, arguing that 'false' is to be predicated not of an utterance but only of a statement or proposition, and that this utterance fails to make a statement. So either 'false' is predicated, with a category-mistake, of an utterance, or the referring phrase attempts to refer to a statement or proposition, and there is none. There may be some force in this as a criticism of the formulation given above, but it can be sidestepped by a reformulation: 'This utterance, standardly construed, says something false'. The only referring phrase now has the utterance to refer to, while 'false' is applied to 'something', which is a variable of the correct category, since it is also the object of the verb 'says'. The objector might continue in either of two ways. First, he might demand a namely-rider.[2] But this demand is not in order.[3] Although if it were true that this utterance, standardly construed, says something false, there would indeed be something which could in principle be individually specified which was the (or a) false thing it said, it is not a requirement for the meaningfulness of an existentially quantified statement that it should be filled out by a namely-rider. Secondly, the objector could say that the

[2] Cf. Y. Bar-Hillel, *Aspects of Language*, pp. 253-7 and 273-285.

[3] Cf. G. Ryle, 'Heterologicality', *Análysis* II (1950-51), pp. 61-69, and P. T. Geach, 'Ryle an Namely-Riders' *Analysis* 21 (1960-61), pp. 64-67.

reasoning within the paradox no longer goes through. If it is not the case that this utterance says something false, it does not follow that it says something true: it might fail to say anything either true or false. This may in the end be an important comment, but what requires stressing at this stage is that if this utterance does fail to say anything either true or false, it is not because it lacks meaning: some other explanation of its failing to say anything true or false will have to be found. In any case, both the objector's possible continuations seem to be blocked by a small further reformulation: 'This utterance, standardly construed, says nothing true'. If *this* utterance failed, for whatever reason, to say anything true, this would seem to ensure that, standardly construed, it did say something true; although equally if it did it could not. So the reasoning within the paradox goes through and the objector's second protest fails. And since there is not even a 'something' in this formula, there is no shadow of a pretext for demanding a namely-rider.

It may be maintained that a consistent language requires hierarchies, that such words as 'true' and 'false' need numerical subscripts, and that when subscripts are inserted in the formulae used to introduce the paradoxes they either cease to be paradoxical (if the subscripts are inserted in one way) or become guilty of type violations (if they are inserted in another) which are equivalent to category-mistakes. That is, the charge of meaninglessness might ride on the back of a doctrine of linguistic hierarchies. But then it is the hierarchical theory that first needs to be established: the misuse of language, if any, is of a subtle and not of a trivial kind.

On the other hand, I would not agree that the whole topic should now be handed over to some class of technicians, say the mathematical foundation-layers or the constructors of formalized languages. These paradoxes, however artificial they may seem and however trivial their subject matter may be, constitute a challenge to the rationality of human thinking in general: they are items about which it is difficult to

say anything comprehensive without ourselves falling into contradiction. If we are unwilling to adopt a general scepticism about reason, we must either take up the challenge ourselves or hope that someone else has done so or will do so on our behalf. But not the technicians. The interest that general philosophy has in solving the paradoxes is different from the interests of either the philosophy of mathematics or the study of formal languages, and what constitutes an adequate solution from these different points of view may differ accordingly. Someone who is constructing a set theory, say, or a set of formal languages is primarily concerned to *exclude* the paradoxes, to ensure at least that none of the known ones will arise within his system and subvert it by committing it to contradictions, and if possible to ensure also that no as yet unknown paradoxes will break out there. But the general philosopher or informal logician wants not to keep the paradoxes out of this or that intensively cultivated area, but to be able to look them calmly in the face when he encounters them in the wildernesses where they are at home. In other words, he wants to show that they are only apparent antinomies, that the issues about which we are tempted into formal contradictions are insubstantial; he wants to understand how our ordinary resources of thought and language allow us to construct paradoxes without being himself committed to endorsing contradictory judgements. Also, his resources for solving them are more limited: unlike the man who is constructing something he is not in a position to ban this or to lay down that; he has to comment on what is there already, and he must hope that his comments will themselves be rationally defensible, not *ad hoc* or arbitrary. Although there may be a wide choice between possible exclusion-devices, it seems unlikely that there will be any real choice between solutions to the wider philosophical problem: it is hard enough to find even one. Of course, it might be that the paradoxes I am grouping together are of two or more radically different kinds, so that different paradoxes

will need different philosophical explanations and resolutions. But I shall try to prove that this is not so.

Various forms of the Theory of Types will exclude some or more or perhaps all of the paradoxes from a system on which they are imposed, but they would provide a philosophical solution only if they had some independent rationale and justification. As *ad hoc* restrictions, imposed just because the paradoxes would arise without them, they would do nothing to *solve* the paradoxes, though they might be a convenient way of excluding them. In fact the very simplest Theory of Types, one which says merely that there are individuals, classes of individuals, classes of classes of individuals, and so on, which can be labelled as items of order 0, 1, 2, and so on, and that an item of order n can have as its members only items of orders less than n, has some intrinsic plausibility, and it will defeat, for example Russell's class paradox. All classes will be non-self-membered, but there will be no class of all the non-self-membered classes, since it would have to be of an order higher than itself. This theory, then, might provide a philosophical solution of Russell's paradox taken on its own. But it does not seem reasonable to suppose that the formally similar but 'semantic' paradoxes are to be explained and resolved in some quite different way, while any extension or ramification of the Theory of Types in order to deal with them in the same way seems utterly implausible. It would have no independent rationale, and so could not solve the philosophical problem.

The same holds for all kinds of linguistic hierarchies. If someone is setting up, artificially, a system of formal languages, he can if he wishes make it consist of a distinct object language, meta-language, meta-meta-language, and so on. But it would be a piece of pure mythology to pretend to find such a hierarchy within a natural language such as standard English. It is sheer fantasy to suppose that instead of one word 'true' with a pretty simple general meaning

English contains a family of predicates 'true$_1$', 'true$_2$', and so on, with different ranges of possible application.

Of course Tarski[4] does not say this, nor does he say quite that a natural language such as English is inconsistent and needs to be reformed. His view is rather that a natural language has no 'exactly specified structure', and consequently that the question whether it is consistent or not has no exact meaning. But if a language had an exactly specified structure like that which natural languages seem to have, so that it was both what Tarski calls 'semantically closed' and such that the ordinary rules of logic held within it, it would necessarily be inconsistent. The hierarchical distinctions are introduced not as a description of what is already there, but as a requirement that must be satisfied if inconsistency is to be avoided. But since the supposed proof of this is just the paradoxes themselves, these distinctions do nothing to solve the philosophical problem.

Tarski's thesis, that a language which both is semantically closed and contains the ordinary logical rules must be inconsistent, calls for some examination. A language is semantically closed, in his terminology, if it 'contains, in addition to its expressions, also the names of these expressions, as well as semantic terms such as "true" referring to sentences of this language' and if 'all sentences which determine the adequate usage of this term can be asserted in the language'.[5] But what is it for a *language* to be inconsistent? *Prima facie*, what can be inconsistent is a statement or set of statements or theory, while a language is only a vehicle, a medium in which things can be said, not a set of statements. However, what is meant is that the formation rules of the language permit the construction of sentences which its other rules — especially the meaning-rules for such terms as 'true' — will require us to call both true and not true.

[4] 'The Semantic Conception of Truth' in *Readings in Philosophical Analysis*, ed. Feigl and Sellars, pp. 59-60.
[5] *Op. cit.*, p. 59.

Thus interpreted, Tarski's thesis amounts only to the fact that in a language which is semantically closed in the way in which ordinary English, for example, appears to be, such antinomies as the simple liar can be constructed. But why does this matter? Tarski stresses the commonsense point that an inconsistent theory must contain falsehoods, and is therefore unacceptable. But this comment applies to *theories*, not to a *language* which is inconsistent in the sense explained. Still, it would follow that the rules of an inconsistent language would commit those who always obeyed them and who were ready to answer all questions to saying things not all of which are true. But it is worth noting that an 'inconsistent' language can be used without embarrassment by anyone who steers clear of certain questions, in much the same way that a car which would fall to pieces at ninety miles an hour can be safely driven at more modest speeds.

But must we accept Tarski's thesis even in the sense explained? I think not. Antinomies like the liar can be blocked not merely by taking things away from a natural language, but by adding extra items, or by insisting that they are there already. Arthur Prior, for example, has argued that a method proposed by Buridan in the fourteenth century and more recently by Peirce would achieve this.[6] If we assume that every statement asserts its own truth (whatever else it may assert as well) we remove the contradictions by blocking one arm of the reasoning within each paradox, for example, the argument that if the simple liar utterance is false, it is true, and hence that it *is* true. On the hypothesis that the utterance is false, it follows that one part of what it asserts, namely its own falsehood, is true, but another part of what it asserts, its own truth, is on this hypothesis not true. The other arm of the reasoning within the paradox, that if the utterance is true, it is false, and hence that it *is* false, still stands, but it

[6] 'Some Problems of Self-Reference in John Buridan', British Academy Lecture, reprinted in *Studies in Philosophy*, edited by J. N. Findlay, esp. p. 254.

stands without opposition: there is no longer a paradox. The simple liar utterance has become contradictory within itself, asserting its own falsehood and its own truth. We can therefore classify it as simply false, and are relieved of the necessity of making contradictory comments upon it. Thus, Prior says, 'a language *can* contain its own semantics . . . provided that this semantics contains the law that for any sentence x, x means that x is true'.

I am using this point only destructively, as a disproof of Tarski's thesis. I am unwilling to accept it as a philosophical solution because its range of application is too narrow. Neither it nor anything closely analogous to it will, as far as I can see, resolve the paradoxes of Richard and Berry, or the class paradox. Also, all the paradoxes have what we may call truth-teller counterparts: for example, consider the remark 'What I am now saying is true'; if 'autological' is the opposite of 'heterological', consider whether 'autological' is autological or not; whether the class of self-membered classes is a member of itself or not; and so on. None of these commits a commentator to a contradiction; but there is still a puzzle. If the truth-teller utterance is true, then it is true; but equally if it is false, it is false; it might be either, and the question which it is is undecidable. And similarly with all the rest of the truth-teller variants. The Buridan-Peirce-Prior move does nothing to resolve this puzzle. I believe that an adequate philosophical solution would deal at once with all the paradoxes of this group and with their truth-teller variants as well.

Another criticism of Tarski's thesis is that merely refraining from the use of semantically closed languages is not sufficient to prevent the appearance of antinomies. Suppose that there were two languages, L_1 and L_2, neither semantically closed, but each serving as the meta-language of the other. And suppose that S_1 is in L_1 and reads 'S_2 is false in L_2', while S_2 is in L_2 and reads 'S_1 is true in L_1'. Then if S_2 is false in L_2, S_1 is true in L_1 — because what it says is so — and hence, since this is what S_2 says, S_2 is true in L_2. But equally if S_2

is true in L_2, 'S_2 is false in L_2' is false in L_1, that is, S_1 is false in L_1, so S_1 is not true in L_1, and therefore S_2 is false in L_2. Thus we can prove that S_2 is both true and not true in L_2, and similarly that S_1 is both true and not true in L_1, and we still have a paradox. This is, of course, merely an indirect variant of the liar expressed in terms of languages. To exclude it, we should need not only the rule that no one language can be semantically closed, but also the rule that no circle of languages can be semantically closed: their relations must be hierarchical and therefore open-ended. But then it is plain that it is not the semantic openness or clasedness of a *language* that matters, but the possibility of a semantic circularity.

It looks as if we should go back to Russell's (or Poincaré's) notion that the basic trouble in all these paradoxes is some kind of vicious circle.[7] But Russell did not suceed in saying exactly what kind of circularity is at fault.

It is sometimes suggested that the fundamental error is *self-reference*.[8] But literal self-reference is not in general vicious, and it is neither sufficient nor necessary for paradox. 'This is an English sentence' and the like are quite in order, while the sentences in an indirect variant of the liar refer not to themselves but to one another. We may try to distinguish genuine (and vicious) from spurious (and harmless) self-reference; but we also need to distinguish partial from total self-reference, and even genuine self-reference, if only partial, seems to be harmless, although in some cases what starts as partial self-reference can become vicious in particular contingent circumstances. I shall say more about this whole topic in Part II; but this much is clear: it is not merely *referring* to itself that makes any item logically defective, but

[7] Whitehead and Russell, *Principia Mathematica* I, pp. 37-38.

[8] E.g. A. Ross, 'On Self-Reference and a Puzzle in Constitutional Law', *Mind* LXXVIII (1969), pp. 1-24. This view is opposed by e.g. K. R. Popper, 'Self-Reference and Meaning', *Mind* LXIII (1954), pp. 162-9 and H. L. A. Hart, 'Self-Refering Laws', in *Festkrift tillägnad Karl Olivecrona.*

perhaps being so constructed that in an important way it *depends* upon itself.

Before trying to work this out more accurately, I shall glance at another approach to the paradoxes, one that relies essentially on a formal logical proof. One exponent of this approach is J. F. Thomson.[9] He starts by proving a 'small theorem': 'Let S be any set and R any relation defined at least on S. Then no element of S has R to all only those S-elements which do not have R to themselves'. I shall call this the *barber theorem*, because its most obvious application is to the barber paradox: No collection of men contains a man who shaves all and only those men in the collection who do not shave themselves.

As Thomson says, this theorem is a plain and simple logical truth, and so is the barber application of it. But further applications include 'No collection of classes contains a class having as members all and only those classes in the collection which do not have themselves as members', and 'No collection of adjectives contains an adjective which is true of all and only those adjectives in the collection which are not true of themselves', and so on. In other words, this plain and simple theorem shows that there is no such class as Russell's paradoxical class, no such adjective as 'heterological' is supposed to be, and so on.

But does this proof solve the paradoxes? Surely not.[10] It disposes of the barber, because we have no reason to suppose that there is, and not much reason to suppose that there might be, such a barber as the story requires. But it does not dispose of Russell's paradox or Grelling's, because we still have on our hands a contradiction between the appropriate interpretation of the barber theorem and the *prima facie* case for saying that since there clearly are non-self-membered

[9] 'On Some Paradoxes' in *Analytical Philosophy* (1st Series), ed. R. J. Butler, pp. 104-119.

[10] Cf. A. A. Fraenkel and Y. Bar-Hillel, *Foundations of Set Theory*, pp. 6-7, and W. V. Quine, *The Ways of Paradox*, pp. 6-14.

classes, there must be a class that contains them all and only them, or for saying that 'not true of itself' or 'not truly applicable to itself' is a clear and meaningful description; there are precise and known rules for its use, and even if there were not, they could be introduced; this is an English adjective (for the distinction between adjectival phrases and adjectives is irrelevant here) of just the sort which the barber theorem says cannot exist, and the coinage 'heterological' is merely shorthand for it.

In effect, all the work remains to be done. We need to demolish the *prima facie* cases for saying that Russell's class exists, that 'heterological' and/or its longhand equivalents exist and mean just what they are intended to mean, and so on. It is here that the vicious circle notion is important.

Let us first compare 'heterological' with an imperative paradox like the gallows or Sancho Panza.[11] The lord of the manor's instructions to the guard to hang all and only those travellers who give false reports of what they will do, when applied to the awkward traveller who says he is going to be hanged, amount to telling the guard to hang this man if and only if he does not hang him. We have no difficulty in seeing that these instructions are in this particular case empty —though in other cases they are clear and determinate— because the guard's decision has been made to depend (inversely) on itself. Similarly we may be given the task of filling in a table in accordance with these instructions:[12] 'In the column headed "long", put a tick in any line if and only if the adjective in that line is long; in the column headed "short", put a tick if and only if the adjective in that line is short; and in the column headed "heterological" put a tick if and only if you do not put a tick in that line in the column which has the adjective in that line at is head.'

[11] Cervantes, *Don Quixote*, Pt II Ch 51.
[12] Based on Thomson, *op. cit.*, pp. 111-2.

	long	short	heterological
long			
short			
heterological			

While we have no difficulty in obying these instructions all the way down the columns headed 'long' and 'short' and in the first two lines in the third column, in the third line in the third column they amount to 'Put a tick here if and only if you don't put a tick here'; and this clearly fails as an instruction.

It is worth noting that similar comments apply to the truth-teller variants. About another awkward traveller who said merely 'I am not going to be hanged on that gallows' the guard's instructions become: 'Hang him if and only if you hang him', which still fails as an instruction; it leaves the guard free to act as his own benevolence or malice may dictate. The appropriate instruction if we added a fourth line and a fourth column for 'autological' would similarly amount to this. 'Put a tick here if and only if you put a tick here'.

But these imperative counterparts are only an illustration. 'Heterological' and its longhand equivalents are intended to be descriptions, and what they do or do not apply to should be a matter of fact, not of decision. But for the same reason why the ticking instructions become empty, the corresponding descriptions fail to describe. The rules for the use of 'not truly applicable to itself' take no grip when this phrase is being considered for application to itself. But it was precisely the existence of, or the possibility of introducing, those rules which was the foundation of the *prima facie* case for the view that there is, or may be, just such an adjective as 'heterological' is supposed to be. That case is undermined by showing that in this particular situation these rules fail to apply substantially.

This point is very like one of those made by Ryle.[13] The question 'Is "long" heterological?' can be unpacked, in view of the meaning rules for 'heterological', to give the more explicit question 'Is "long" not long?'. But the question 'Is "heterological" heterological?' resists unpacking. But putting it in this way must not be taken as the laying down of some requirement that all terms of a certain class (perhaps 'semantic' ones) should be finitely unpackable or eliminable. The non-unpackability merely reveals and illustrates the fact that no real issue is being raised, that in the case of 'heterological' itself there is nothing for being heterological to be.

Should we agree with Thomson, then, that 'heterological' is not within its own domain of possible application?[14] Those who use this adjective may intend it to be so. Likewise the standard rules of English would give its longhand equivalents an intended domain of possible application which included all adjectives and adjectival phrases, themselves among them. But these intentions are, we might almost say, providentially frustrated. These words and phrases fail substantially to come within their own domain of possible application. Because of the circularity, when we try to assert or deny any of them of itself we raise no real issue.

This fact undermines the otherwise strong case for the presumption that there is such an adjective as 'heterological' is intended to be? (which includes the assumption that it is contained in the class of adjectives within which it is supposed to make a sharp dichotomy). This, then, removes the contradiction which was still there after the barber theorem had been applied, the conflict between the appropriate interpretation of that theorem and the linguistic case for the presumption.

Three further points can be made against the formal proof approach and in favour of the other, essentially Rylean, treatment. First, the barber theorem does not apply directly

[13] G. Ryle, 'Heterologicality', *Analysis* 11 (1950-51), pp. 61-69.
[14] *Op. cit.*, p. 110.

to the simple liar. An analogous proof could no doubt be constructed to show that there is no such utterance, with its standard English meaning, as 'What I am now saying is false'. But there would then be a crying need for a further explanation of this, for something to undermine the *prima facie* case for supposing that there can be such an utterance. Secondly, there is no truth-teller counterpart of the barber theorem, nothing to show, for example, that there is not such an adjective as 'autological' is intended to be; there may well be a village in which the barber, somewhat superflously, shaves all and only those men who do shave themselves. But while it will be a simple matter of fact whether this barber also shaves himself or not, it cannot be a simple matter of fact, independently decidable, whether 'autological' applies to itself or not. This ought to be decidable on logical grounds, but it isn't. This and all such truth-teller puzzles are almost as embarrassing as their negative counterparts, and the formal proof approach does nothing to resolve them, but the Rylean approach is equally effective here. Thirdly, the formal proof approach itself generates, in what I call Prior's family of paradoxes, a new series of puzzles which the Rylean treatment is needed to resolve.[15]

This treatment seems to be just what is needed to deal with all the 'semantic' paradoxes and with their truth-teller counterparts. But it is to all appearances too essentially linguistic to cope with the class paradox, that is, to provide what is needed in that region as a supplement to the barber theorem, a way of undermining the *prima facie* case for supposing that there is a class of all and only the non-self-membered classes. The fasionable opinion about this seems to be that this 'case' is merely a prejudice, a consequence of our natural or naive view that there is a class for every property, or, as Quine puts it, a class for every open sentence, and that this naive view must simply be abandoned in the

[15] These will be discussed in Part II of this article.

face of the paradoxes.[16] It is hard to see, however, that this approach differs at all from the simple reliance on the barber theorem, which the same writers commonly condemn as insufficient to solve any paradoxes except the very weakest —the barber paradox itself, for instance, and the crocodile. Nothing has been explained. Why are there classes determined by most properties and most open sentences, but not by a few special ones? How can a property fail to mark off a set of things that have it from all the rest that do not? An explanation is called for; and yet, as I said, the Rylean one looks too linguistic: classes are just there, they do not wait to be defined or constructed. The Poincaré-Russell objection to 'impredicative' definition similarly seems to miss the point. As Quine says, 'we are not to view classes literally as created through being specified... as increasing in number with the passage of time. Poincaré proposed no temporal implementation of class theory. The doctrine of classes is rather that they are there from the start. This being so, there is no evident fallacy in impredicative specification.' And Quine concludes that 'the ban urged by Russell and by Poincaré is not to be hailed as the exposure of some hidden but (once exposed) palpable fallacy that underlay the paradoxes. Rather it is one of various proposals for so restricting the law of comprehension [which is involved in our naive notion of a class]:

$$(\exists y)(x)(x \in y . \equiv Fx)$$

as to thin the universe of classes down to the point of consistency'.[17]

But whatever defects there may have been in the Poincaré-Russell formulation, I think that Quine's conclusion is the reverse of the truth. There *is* a fundamental fallacy to be

[16] See e.g. W. V. Quine, *Set Theory and its Logic*, pp. 3-5, and *Philosophy of Logic*, p. 45.
[17] *Set Theory and its Logic*, pp. 241-3.

exposed, not a need for an *ad hoc* restriction to thin down the universe of classes to consistency.

Contrary to appearances, considerations just like those that affect heterologicality are relevant when we ask whether the class of non-self-membered classes is a member of itself or not. What would it be for it to be, or not to be, a member of itself? For the class of men to be a member of itself would be for it to be a man, which it clearly and simply is not. But for the class of non-self-membered classes to be a member of itself would be for it to be not self-membered, i.e. not a member of itself. And that would be for it to be not non-self-membered, i.e., a member of itself. And that . . . . . and so on *ad infinitum.* The statement that this class is, or is not, a member of itself resists unpacking just as obstinately as the corresponding about 'heterological'. Consequently the apparently concrete question whether it is a member of itself raises no substantial issue; there can be no hard fact either way.

If an alleged class is determined intensionally, by the fact that all and only its members have a certain property, the reality and determinacy of the class depends on the determinacy of the property. The property of being non-self-membered is in general a quite real property, but it is a derivative one, and it is determinate where and only where there is something for it to be derivative from. There is something for the class-of-men's being non-self-membered to be derivative from, the simple fact that it is not a man. But there is nothing analogous for the paradoxical class (or for its counterpart, the class of all self-membered classes). It is because the property here becomes indeterminate that it fails to produce a dichotomy of all classes, including ones determined by it and by its negation, into those which possess it and those which do not.

This local indeterminacy of the key property is what undermines the *prima facie* case for supposing that there must be such classes as the paradoxical one and its counter-

part, and therefore leaves us free to accept the appropriate interpretation of the barber theorem.

We can agree, then, that there is no determinate class of all and only the non-self-membered classes, or of all and only the self-membered classes, though there are classes of each sort. But the reason is not that the former would violate the barber theorem (which the latter would not). Nor — what is practically equivalent — is it because the paradoxes themselves show that we must modify our naive concept of a class. Nor, as we can now see, is it because a class must be of a higher order than its members; though this doctrine has some intrinsic plausibility, its failure to resolve the 'semantic' paradoxes shows that it does not get to the root of the trouble. Nor is it because some axiomatic set theory has been carefully constructed so that the existence of this troublesome class cannot be proved within it; for that would leave the paradoxical class untamed in the wilderness outside that theory. The reason is that the derivative features which one tries to use to determine the supposed classes are not derived and are non-derivable at certain points.

But since it is *this* that undermines the *prima facie* case for the existence of the paradoxical class, we have no reason for giving up the *other* premiss on which that case rests, the assumption that every (determinate) property determines a class. If the property had been all right, the class would have been all right too, as our naive and natural view would require.

Someone may say that I *have* given up or modified the naive view of classes or sets. That depends on just how naive it was. I hesitate to speak for anyone else in such matters, but it seems unlikely that anyone who admitted that the description 'green', say, was a bit fuzzy at the edges would be quite so naive as to suppose *at the same time* that there was a fully determinate class of all and only the green things. Of course, there could be a fuzzy-at-the-edges class of green things, and in many fields we are prepared to work with

fuzzy classes.[18] But in logical and mathematical set theory we want non-fuzzy classes, and it is natural to expect that only determinate properties can be relied on to mark them out. Quine, of course, prudently — or perhaps imprudently — likes to steer clear of properties and attributes, and to deal with open sentences; we have fewer naive convictions about open sentences than about properties, but I do not see why anyone should have been so naive as to suppose that every open sentence determines a class. Consider the open sentence (type or token) 'x is beside this'. Would anyone suppose that this determined a class — and one different from that determined by the open sentence 'x is not beside this' — if the reference of the word 'this' were not tied down? Surely the natural assumption is merely that every *determinate* property or *determinate* description carries a (determinate) class with it, and that assumption has not been impugned.

Of course, the kinds of indeterminacy that may be found in 'green' and in 'this' are different from that in 'self-membered'. Study of the paradoxes brings to light unexpected sorts of fuzziness. But this does not mean that we have to modify our natural view of the relation between properties and classes, but only that we have to apply to new cases the rules already implicit in our use of that relation.

Someone may object that I have surreptitiously made use of the sort of type distinction I am pretending to do without. I accept, e.g., being a man as something definite that needs no further unpacking, but I do not similarly accept being false, or being heterological, or being 'ordinary' — Thomson's innocent-looking shorthand for 'non-self-membered'. These I accept as something definite only where they can be unpacked. Higher order properties occur only where they arise from some first-order properties or states of affairs. It is true that such a distinction is implicit in my treatment. But *this* type-distinction, if it is so described, is one which

[18] E g., the working class, the middle class. This point was drawn to my attention by Mr. G. J. Warnock.

has an independent rationale, and is not introduced simply in order to resolve the paradoxes. And this distinction is quite different from the setting up of an infinite hierarchy of types with rules that restrict the possibilities of class-membership or, what is worse, of the application of predicates to subjects.

I would argue that essentially the same approach will solve such paradoxes as those of Richard and Berry, as well as those in 'Prior's family', some of which I shall discuss in Part II of this article.

Since all these paradoxes have a common source, and since a single approach will provide a philosophical solution for them all, the distinction between the 'semantic' or 'linguistic' ones and the purely logico-mathematical ones is superficial.

Type rules and language hierarchies have no general authority; they are merely devices which someone may or may not adopt in a constructed system. They effectively prevent the sort of self-dependence which is used in all these paradoxes, but they prevent much more besides. Many things that would violate hierarchical principles are in themselves innocent; they become victims of guilt by association. It is widely recognized that these hierarchical devices are clumsy and inconvenient; but what is more, they are philosophically misleading, suggesting impropieties where there are none. It is just wrong to say that 'true', for example, is systematically ambiguous.[19]

The true moral to be drawn from the paradoxes is that we have to take care not to be fooled either by words or by symbols. In deciding what are real contradictions, and perhaps in other tasks as well, we need to add as a check or qualification upon formal and mechanical calculations the informal self-conscious reflection, 'Is there a real issue here or not?' But we can live with merely apparent contradictions, and we must learn to do so.

However, the principles I have suggested still need to be

[19] Whitehead and Russell, *principia Mathematica* I, p. 41.

tested in some further tricky cases, and there are difficulties and objections to be met.

RESUMEN

Hay un grupo de paradojas, algunas bien conocidas, otras similares pero menos conocidas, que incluye la de Epiménides, heterogicidad, la paradoja de Russell, la de Richard y algunas más, las cuales podrían ser llamadas paradojas de auto-referencia; pero ésto sería evadir varias cuestiones. Son ampliamente sostenidos distintos enfoques con respecto a esas paradojas los cuales parecen ser distintos enfoques con respecto a esas paradojas los cuales parecen ser incompatibles entre sí. Yo pienso que todos estos enfoques están equivocados, aunque algunos, quizá, lo están en mayor medida que otros.

No estoy de acuerdo con la idea de que el problema de las paradojas deba ser manejado por algún tipo de técnico especializado, digamos los fundamentadores y legisladores de la matemática o los constructores de lenguajes formalizados. Estas paradojas, con toda la artificialidad que pueda parecer que tienen y lo trivial de su asunto, plantean un problema filosófico de central importancia, pues constituyen un desafío a la racionalidad del pensamiento humano en general: son cuestiones respecto a las cuales es difícil decir cualquier cosa comprensible sin caer en contradicción. Si no estamos dispuestos a adoptar un escepticismo general acerca de la razón, tenemos que hacer una de dos cosas, o aceptar el desafío nosotros mismos, o estar esperanzados en que alguien más lo haya hecho o lo haga en nuestro lugar sin que sea necesario que este alguien sea un técnico. El interés que la filosofía en general tiene en resolver las paradojas es diferente del interés que los filósofos de la matemática o del estudio de lenguajes formales puedan tener; y lo que constituya una solución desde estos diferentes puntos de vista, por consiguiente, puede diferir. El filósofo general o lógico informal no quiere excluir las paradojas de esta o aquella área específica de estudio, sino ser capaz de mirarlas calmadamente enfrentándose a ellas cuando las encuentre en cualquier sitio donde hallen su asiento. En otras palabras, quiere mostrar que son sólo antinomias aparentes, esto es, que las cuestiones que parecen orillarnos a contradicciones formales son en realidad insustanciales. Dicho filósofo quiere entender cómo es que nuestros recursos ordinarios de pensamiento y lenguaje nos permiten construir paradojas sin comprometerse él mismo a endosar juicios contradictorios. Sus recursos para resolver estas paradojas son limitados: a diferencia del hombre que está construyendo un sistema formal él no está en una situación que

le permita prescribir ésto o establecer aquéllo; sólo tiene que comentar sobre lo que ya está ahí, y deberá esperar que sus juicios sean racionalmente defendibles y no *ad hoc* o arbitrarios. Parece haber varias posibilidades de elección entre distintos mecanismos para excluir paradojas.

Teoría de los tipos y jerarquización de lenguaje

En primer lugar estaría el mecanismo propuesto por B. Russell, la Teoría de los Tipos. Esta pretendida solución, restringiendo la ley de comprehensión (la ley que dice que toda propiedad determina una clase), modifica la noción intuitiva de clase y formula requisitos para que una propiedad defina o no la clase de todos los elementos que satisfagan tal propiedad. Pero esta manera de eliminar las paradojas, en primer lugar no las excluiría a todas, por ejemplo la de heterologicidad, y además negaría principios que parecen mucho más obvios que su negación, o sea, este es un mecanismo completamente *ad hoc* y sin racionalidad independiente.

Otra manera como se ha tratado de eliminar las paradojas es establecer una jerarquización de lenguajes (Tarski). Esto resulta inaceptable porque las distinciones jerárquicas se introducen para evitar la inconsistencia, pero como la prueba de ésto son las paradojas mismas, el problem filosófico no queda resuelto por tales distinciones. Por otra parte, esta jerarquización, el mismo Tarski lo afirma, sólo es posible respecto de lenguajes formales y es en el lenguaje ordinario donde el problema filosófico se plantea. Tarski no sostiene que un lenguaje natural, como el inglés, sea inconsistente y que por ello tenga que se reformado. Para él un lenguaje natural no tiene 'una estructura claramente especificada', por lo cual, no tiene un sentido exacto el plantearse si es o no consistente. Pero en el caso de que un lenguaje natural fuese 'semánticamente cerrado', o sea que tuvieran vigencia las reglas ordinarias de la lógica dentro de él, pienso que sería necesariamente inconsistente.

Ni la teoría de los tipos ni ningún tipo de jerarquización propuesta como solución a las paradojas es satisfactoria, ya sea porque deja un número de paradojas por solucionar, o porque tal solución está construida *ad hoc* sin gozar de una racionalidad independiente.

El tratamiento de las paradojas que yo propongo hace uso de una distinción de tipos, si así se quiere describir, pero es tal que tiene racionalidad independiente, no está introducida para resolver las paradojas simplemente. Hacer de esta distinción es completamente distinto a formular una jerarquía infinita de tipos con reglas que restringen las posibilidades de membrecía a clase o, lo que es peor, de

aplicación de predicados a sujetos. Este tratamiento vendrá explícito en la Parte II de este artículo.

Ya que todas estas paradojas parecen tener una fuente común, y por lo mismo parece ser que un solo tratamiento dará la solución filosófica a todas ellas, la distinción entre paradojas 'semánticas' o 'lingüísticas' y las puramente 'lógico-matemáticas' me parece superficial.

Las reglas que gobiernan los tipos y la jerarquización de lenguajes no tienen una autoridad general; son simples mecanismos que alguien puede adoptar o no en la construcción de un sistema. Estos mecanismos previenen de manera efectiva el tipo de auto-dependencia que se usa en todas estas paradojas, pero previenen muchas más cosas aún. Muchas cuestiones que violarían principios jerárquicos son en sí mismas inocentes; se hacen víctimas de culpa por asociación. Está ampliamente reconocido que estos mecanismos jerárquicos son incómodos e inconvenientes; pero lo que es más, son filosóficamente errados, sugieren impropiedades donde no hay ninguna.

La verdadera moraleja extraída de las paradojas es que debemos tener cuidado de no dejarnos engañar por palabras o por símbolos. Al decidir lo que son contradicciones reales, y quizá también en otras tareas, necesitamos añadir sobre los cálculos formales y mecánicos, la reflexión informal auto-conciente: '¿hay realmente un problema aquí o no?'