VI.—DISCUSSIONS

THE PREDICTION PARADOX

THIS article is divided into two parts; the first deals with the so-called "prediction paradox", the latest attack on which is by Mr. R. Shaw (MIND, 1958, p. 382); the second contains an application of my attempted solution of this paradox to the problem of free will. If my dismissal of earlier 'solutions' of the paradox should appear summary, this is not because I do not admire the skill and esotericism with which the problem has been approached, but because I feel that my approach has the advantage of being so simple that anyone who reads it will agree with the solution.

I

The paradox can be considered in the following form (see Shaw, 1958, Weiss, 1952, etc.): A headmaster announces to his pupils that it is an unbreakable school rule that an examination will be given on an unexpected day of the next term. A clever pupil reasons that it cannot be given on the last day, because then they would know on the eve of the last day that it could only take place on the morrow, in which case it would not be unexpected. On the eve of the penultimate day, they would know that, since the master cannot wait until the last day, he must give it on the penultimate day, so again it would not be unexpected, and thus it cannot be given on the penultimate day. And similarly, for any finite number of days; so it would appear that the rules are self-contradictory, and the examination cannot be given. The master sets it, however, say a week after the beginning of term, and this, or indeed any other time, seems to satisfy the rules of the school.

Mr. Shaw, after stating that, to his mind, Quine evades the paradox rather than resolving it, quite explicitly evades it himself. He actually writes: "What rules do we *choose* for the school?" (p. 383, my italics), and then produces a set of rules which are non-self-referring, and gets no paradox. This is hardly surprising; but there is no need for him to consider $(r + 1)$ rules, of some complexity, for an n-day term; he could simply *choose* the rule "There will be an examination on one of the days of next term" which is both non-self-referential, and leaves the situation such that the pupils will in fact not know on which day it will take place. Another excellent way to avoid paradoxes is to say nothing; if the master does this he may offend the boys but he can be sure that he will not offend logic. People with no rules at all are always safe in so far as they cannot contravene them. Compare "A contradiction cannot occur in nature". But contradictions neither occur nor fail to occur *in nature*, only in linguistic expressions. What this means is that if, in argument, we reach a contradiction, then either we have argued

incorrectly, or one of the premisses either contradicts another premise, in which case one of them is false, or is itself self-contradictory. I show below that the school rule is ambiguous; if it means one thing then the clever boy argued incorrectly, if another, then it has no empirical application because it is self-contradictory, and the apparent instantiation of the rules is not an instantiation at all. What is so puzzling about the paradox is that the rule appears to be both self-contradictory and yet instantiated at the same time; naturally, we can get out of this difficulty by choosing a *different* rule, but this is absolutely nothing to the point. We wish to find out what went wrong with the rules *as stated*; surely they cannot be self-contradictory, since they were instantiated! So it would seem that there must be something wrong with the clever boy's argument. But what? Here, as so often in philosophy, the search for the general is the illusion; we look for *the* answer, instead of different answers depending on the way the rules are taken.

It is unclear to me whether Mr. Shaw thinks that we are bound to get paradoxes if we use self-referential statements, perhaps because we are confusing object-languages and meta-languages, but he says "It is clear that the origin of the paradox lies in the self-referring nature of Rule 2*" (p. 384). But many self-referring sentences are perfectly all right, *e.g.* "This sentence is written in black type", which will be true or false, or "Many people expected a paper on expectation paradoxes to appear in MIND, but no one expected *this* paper". The last sentence refers to this paper and thus to itself, as does this sentence (twice). Moreover it is true. I wish, however, to take Shaw's Rules 1 and 2* and show that no paradox need arise. Although the apparent paradox arises over an ambiguity in the phrase "you will be unable to predict", or "you will not know", I cannot accept Quine's solution, for by his criterion I could never be said to know anything about the future. This only means that whenever I say "I know that X will happen" it is logically possible that X should fail to occur, despite all the predictions of hanging judges, schoolmasters, scientists, angels, or prediction machines. For how could I know about the future better than I do when a judge tells me that he will hang me on the morrow, or when a master tells me that he will give me an examination on the morrow in accordance with an unbreakable school rule? True, the rule, like an insurance policy, does not allow for civil riots or acts of God, but it is not in this sense that it is said to be unbreakable.

Weiss is almost correct when he says that if the announcement is made on the eve of the penultimate day, then "it is not predictable which of two days will be the day on which the exam will be given",[1] but this is not because "There is as yet no distinct next day or day after on which the exam could be given", nor because, of $f(x \vee \bar{x})$ and $f(x) \vee f(\bar{x})$, "the former is a necessary truth, the latter is true only when one has isolated the x and the non-x, an act which requires

[1] MIND (1952), p. 267.

one to leave the realm of possibility for the realm of time, history, becoming " . . .

Now suppose that a man A, holding six black cards and one red card in his hand,[1] says to a man B " I am now going to lay the cards one at a time face upwards on the table in front of me. You will not be able to predict, before I turn it over, which time I am going to play the red card." B replies, " Well, you cannot leave the red until last, because then I would know, and you cannot leave it until last but one because then, knowing that you could not leave it until last, I would know that you were going to play it last but one, and then you cannot etc. etc. etc. . . . and therefore you cannot lay it at all ! In fact, come to think of it, if you hold 52 cards, all different, and say that I cannot predict which card you will lay, then you have contradicted yourself. For, if you leave the ace of clubs until last, I would know that you were going to lay it last, and so you cannot lay it last. Therefore neither can you lay it one but last, nor 52-n but last, (n = 50, . . . 1), and so you cannot lay it at all. And similarly for all the other cards."

The absurdity of this argument is manifest. When A holds just one black and one red card, and says to B " You cannot predict which card I shall lay ", he means that B cannot predict the first card, not that B will be unable to predict the colour of the second card when he has seen the first. Or if he does mean this, he is just wrong, as B will quickly show him. Similarly, when, on the eve of the penultimate day, the master claims that the boys cannot predict which day the examination will be given, he does not mean that *if* he does not give it on the morrow the boys will *then* be unable to predict that it will be on the last day. Or *if he does mean this, then what he has said is false, and his giving it on the first day does nothing to prove him right*, for it simply fails to satisfy his claim that " the date of the examination is unpredictable even if we wait until the eve of the last day ".

The headmaster may perfectly well claim that the rules just *are* Shaw's self-referring and unbreakable [2] R1 and R2*, *viz.* :

R1 An examination will take place on one day of next term.

R2* The examination will be unexpected in the sense that it will take place on such a day that on the previous evening it will not be possible for the pupils to deduce from Rules 1 and 2* that the examination will take place on the morrow.

What the headmaster surely cannot deny, without contradicting himself, is that R2* entails, or if you like, means, either S1 or S2, *but not both*, where these are :

S1 The examination will be unexpected in the sense that . . . it will not be possible for the pupils to deduce from Rules R1 and S1 that the examination will take place on the morrow, unless it takes place on the last day.

[1] I believe that I owe this example to Mr. S. Anstis.

[2] This cannot mean logically unbreakable ; they must allow for epidemics, wars, or sudden chaotic cancellations of natural laws.

S2 The examination will be unexpected in the sense that . . . it will not be possible for the pupils to deduce from Rules R1 and S2 that the examination will take place on the morrow, even if it takes place on the last day.

If it means S1, which is the more sensible, then the argument put forward by the clever boy is fallacious, because the rule, applied for instance on the eve of the penultimate day, states merely that the boys will not know on which day the examination will take place *unless* the master waits until the last day. If it means S2, then it can have no possible application, must always remain false, for nothing, including setting the examination earlier, would make it true that the boys would be unable to deduce on the eve of the last day that it would occur on the morrow, *if* the master were to wait that long. For R1 and S2 applied together on the eve of the last day give us :

(1) The examination must take place tomorrow.

(2) (The examination will be unexpected in the sense that) it is not possible to deduce from (1) and (2) that it will take place on the morrow.

(1) and (2) clearly contradict each other, as opposed to Quine's solution. The latter's use of 'know' ('predict', 'expect', etc.) such that I could never know anything about the future, is just as implausible as that opposite use whereby a man who guessed the day correctly and then said that he had predicted rightly might be said to have known. (See further discussion of the card game in Part 11.) Shaw is correct in saying that we are only interested in *valid inference*, in this case deductive as opposed to inductive, etc.

Thus if the rule means S2, there can be no instantiation of this rule, only an apparent instantiation, and the boy argued correctly, if he argued that *this* rule could not be instantiated. Of course the master can *say* that the rule means S2, or even S1 and S2 at the same time, and still set the examination, just as I can make the self-referring, self-contradictory remark "I always lie, including now", and then spend the rest of my life telling lies. But in neither case are we instantiating our statements, for nothing could instantiate a self-contradictory statement.

The paradox arises from taking the rule to mean both S1 and S2 at the same time. The master might claim that the rules in the rule-book do not state which it is meant to mean; but what is certain is that it cannot mean both without being self-contradictory from the start. In this case, like all self-contradictory statements, it can do no harm, it can hurt no one, unless it misleads them into wrongly thinking for instance that there will be no examination, when what they ought to do is expect one, and thus revise, on *every* night.

II

I would like now to consider an application of the prediction paradox to the problem of free will. Certain philosophers have

claimed that we must have free will, or that we certainly often know that there is a goldfinch in the garden or that someone is in pain, etc. etc., because in using the relevant words to describe these situations we are only claiming that the situations easily referred to when we utter these words, do occur. No one likes to drag a corpse from its coffin and then kick it, so I will say no more about this except to refer to the various articles on the Paradigm Case Argument in ANALYSIS during 1958. The question now arises whether the situations paradigmatically described as ones of knowledge, free will, etc., really are correctly so described, when certain beliefs, such as that all knowledge comes through private data, or that there are causal universal laws, seem to conflict with these descriptions.

I do not wish to discuss here whether and in what sense I could have chosen to have chosen . . . to choose differently in a given situation, and whether this lands us in an infinite regress. Neither do I wish to consider whether certain machine-states are essentially unpredictable because of Gödel's theorem. All I want to do is to take a very simple situation, discuss in what sense it is predictable and whether this conflicts with the belief that we have free will, and then to generalize this case.

I presume that no one would deny that in the case of the card players considered above, there is a sense in which A, when he holds but two cards, can play whichever one he chooses. If I understand them correctly, the ordinary language philosophers, by contrasting this sort of case with those in which A is compelled because someone holds a pistol to his head or forcibly moves his hand, declare that free will exists since these are the sorts of case everyone refers to when they use the words "free will". The question whether, as some "metaphysicians" have maintained, this use is a misuse because incompatible with there being universal causal laws, is straightaway ruled out. I shall not repeat the overwhelming arguments against the ordinary language view, but will immediately attempt a consideration of the metaphysician's claim.

Firstly, the mere idea of determinism *per se* does not seem to conflict with our notion that we have free will. Card-player A, if he can decide freely whether to play the red card or the black, and then play it, does nothing contrary to the laws of nature such as float up out of his chair and hit the ceiling; if he did, he could not play the card which he chose, which would hamper his freedom of choice very badly. "Everything which happens does so in accordance with the laws of nature" tends towards analyticity, not because it is one of those statements which Quine and Waismann claim are semi-analytic—for none are—but because we tend to *make* it express a necessary proposition by not allowing anything to count against it. There are excellent pragmatic reasons for doing this, but if we give it up because we have reason to believe that no laws could be found in certain cases, it does not mean that the proposition which the sentence used to express is now false.

Let us therefore assume that it is true ; we have then granted the metaphysician as much as he could possibly want. Now the feeling that this destroys free will might be expressed by A as follows : " Whether I play the red card or the black, my action was determined by causal laws, and so one who knew enough about me and about the laws concerned could have predicted which one I was going to play first, and there is nothing which I could do to frustrate him." We might immediately reply : " Yes there is ; you could lay the other one, for we have surely admitted that to lay either card would be in accordance with the laws." A might reply : " But in that case, the Ace Predictor would not have predicted correctly, which is contrary to hypothesis. He knows the truth of all statements P_1 about world-state W_1 at time t_1, and he knows the laws of nature L ; these, taken together, entail that the world-state W_2 at t_2 shall be P_2, and one of the statements P_2 will be ' A lays red ' or ' A lays black ', which the Ace Predictor could have foreseen. So all the future is in this way laid out before me, and I must follow the paths which fate has laid down for me."

Now does this only make the tautological claim that ' whatever will be, will be ' [1] or does it claim more ? Suppose that B, when he was arguing against A, had said that he could predict which card A was going to lay first. A might say, ridiculously enough, " Well, which card *am* I going to lay first ? ", and when B says " Black " A always lays red, and *vice versa*. I believe that this is all that A could possibly hope for when he claims that he has free will, *i.e.* that *whatever* the Angel of Fate or the Ace Predictor predicts, then if A gets to know what has been predicted, what is written in the Book of Fate, he can act otherwise, thus destroying the power of the Book.

The Angel of Fate might then say, " Of course that wasn't the *real* Book of Fate ; the real one foresaw what you would do when I showed you this one, and in fact predicts everything about your life correctly ". But this is surely only worrying to A if he can be shown the real book of predictions, and he then finds that he cannot help acting in accordance with it because he is somehow constrained or impelled. If the angel replies that the book is kept in heaven and that only God is allowed to look at it, this is no more impressive than if B says, after A has laid the red card, that although he said " Black " he really thought all along that A would lay red.

To make the game fair to both A and B, B should write his prediction on a piece of paper ; this would then be compared with the card after A has laid it. B may, by guessing, always be right ; we should then be mystified, but why should we be any more disturbed if B says that he asks the Angel of Fate, or that he pushes lots of data about neurons, etc., into an enormous computing machine ? It is still the case that A, if shown the prediction, could act differently. If two people B and C are both making predictions about which card A will lay, and if they agree always to differ, then one of them must

[1] See G. Ryle, *Dilemmas*, chap. 11.

be right every time. And if A lays n cards, or performs n actions with m possible alternatives each time, then of m^n possible predictors all predicting differently, one of them logically must be right. That one will be right is no more disturbing than that B should write his first prediction on two pieces of paper, and then after the event produce the one which corresponds with the card that A has lain. In this 'guessing' sense of "predict", it is always false that a person is unable to predict what another will do. Thus when the master says that the date of the examination is unpredictable he does not mean that the boys will be unable to say each evening "Exam tomorrow" and one day be right; indeed he wishes them to say it every night, so that they revise every night.

This sense of 'know' or 'predict' is one diametrically opposed to that used by Quine; we are more interested in a more rational, intermediate sense, where it means "validly infer on deductive or inductive grounds". But even if an Ace Predictor can do this, it still is the case that if A sees part of the machine-tape referring to the future, he can decide to act in discordance with it, if the specifications are precise enough. For instance, suppose that the tape says "A dies 28 Jan., 1962, London"; A may unfortunately die when the machine states, just by coincidence, but he can cheat the machine by taking a train to Liverpool and refusing to move. We must surely admit that such actions are possible, *i.e.* that if someone makes a prediction about our behaviour and then tells us about it, we are very often able to act otherwise. And if some statement about determinism is even analytically true, then this only makes it analytic that the behaviour immediately after viewing a predicting-tape is explicable in terms of a law of nature which probably would not even have been instantiated if the predicting-tape had not been seen.

Thus the Book of Fate kept in heaven, which only God is allowed to see, is as unworrying as the boy's prediction of the examination or the devil's prediction about the cards, whether these are based on natural laws or inspired guesswork. Suppose that the predicting angel gave me a book and on each evening provided me with a key to turn another page, whereon I always found a minute description of all that I had done that day. The angel declares that the story of my life is printed in the book, and that when the book ends, so will my life. At first I might suspect that he is cheating, as B did, when he wrote on two pieces of paper. For he might be noticing everything that I do and then transmitting it daily onto the as yet blank pages by means of invisible angelic high-frequency waves. But as the pages of the book get fewer, I might get more worried; eventually when very few pages are left, I try to break the lock in order to see ahead and escape my doom. If I cannot break the lock, the performance of the Angel of Fate is only startling in that he can make strong locks and use invisible rays, and possibly that he makes one correct prediction, *viz.* the date of my death. If I do open the lock, and see that the moving finger has got there before me, I can take evasive

action and thus cheat the moving finger. It is surely irrelevant whether I or the moving finger always act in accordance with the laws of nature or not.

Thus it would seem that neither determinism nor indeterminism is incompatible with free will, and that the apparent conflict arises through not examining what it is for a thing to be predictable, that is, through not examining particular-case uses of the word "predict" —more and less complex. Once again we tend to rely on a vague, general gesture in the direction of what it is for such and such to be the case, as though we could understand the meaning of the word apart from its particular case-by-case uses, ordinary and extra-ordinary. If we examine particular concrete examples, real or imaginary, such as that of the card-players, and each time ask ourselves in what sense a move is predictable, and in what sense this is, or might be, incompatible with choosing freely, we will see, I believe, that the worry about the Ace Predictor should be laid to rest alongside that of the Arch Deceiver. For each moment of our lives it is as though we play another card, determined partly by those we have played so far, partly by what we have been dealt, and partly by free choice from those we still hold.

ARDON LYON

University of Cambridge