DAVID E. WARD

IMAGINARY SCENARIOS, BLACK BOXES AND PHILOSOPHICAL METHOD[1]

1. INTRODUCTION

When we do philosophy we are often arguing over judgments, over the application of concepts: e.g., is such and such a case of knowledge? Is such and such immoral? Is such and such properly described as a right or a privilege? In trying to answer these questions people rely on their intuitions – a term which, in this context, means a basic tendency to judge that this particular instance *is* a case of knowledge, or that this particular instance *is* a case of the immoral, etc. When people discuss such questions and find they cannot agree on these basic judgments they often feel forced to simply acknowledge the fact that they disagree and declare themselves to be at an impasse. They may then admit that – at this point in the discussion – all they are really doing is trading intuitions.

One traditional way of avoiding this impasse is by inventing new situations – imaginary scenarios – which provide us with intuitions which favour a certain disputed interpretation of a familiar concept. For example Descartes' evil demon presents an imaginary scenario in which our intuitions about what we can know are coerced in a certain direction.[2] (Dennett calls such stories "intuition pumps" [1992, p. 398]). In this paper I shall investigate the legitimacy of this method of arguing.

2. THE PROBLEM AS IT STANDS IN THE LITERATURE

Parfit, for one, uses this method frequently. He defends it as a method of discovery (1984, p. 200) but he recognizes that the method is not above suspicion. He mentions that both Wittgenstein and Quine regarded it as illegitimate. He quotes Quine as follows:

> The method of science fiction has its uses in philosophy, but ... I wonder whether the limits of the method are properly heeded. To seek what is 'logically required' for sameness of person under unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with. (1984, p. 200)

Wittgenstein puts the point more generally in *Zettel*:

Erkenntnis **43:** 181–198, 1995.

It is as if our concepts involved a scaffolding of facts. That would presumably mean: If you imagine certain facts otherwise, describe them otherwise than the way they are, then, you can no longer imagine the application of certain concepts, because the rules for their application have no analogue in the new circumstances. A law is given for human beings, and a jurisprudent may well be capable of drawing consequences for any case that ordinarily comes his way; thus the law evidently has its use, makes sense. Nevertheless its validity presupposes all sorts of things, and if the being that he is to judge is quite deviant from ordinary human beings, then e.g., the decision whether he has done a deed with evil intent will become not difficult but (simply) impossible. (1967, §350 p. 64e)

Wiggins in *Sameness and Substance* also expresses some reservations about the value of thought experiments:

The possibilities of possible possibilities[3] corresponding to these thought experiments may or may not be inconceivable *modulo* the basic or derived laws of the physical world; but they disrupt the expectations on which individuation depends and they disturb the application of the generalizations about the relation of animal and environment whose instantiation by substances sustains definitely their status as persons. (1980, p. 178, note 34)

In the same vein Wiggins notes that:

For me, or for anyone who is willing to be party to the doctrine of individuation that the naturalistic conception of persons makes possible, it seems immensely important that, at the limit, such thought experiments denature the human subject, and create the prospect that, in place of an animal organism with a clear principle of individuation, we shall find some day that we have an entity whose identity has become a matter not of discovery but of interpretation (or even stipulation). (Ibid)

Dennett also inveighs against imaginary cases which are presented as if the situation which they describe were possible in principle. In the case of the brain-in-a-vat he makes the following point:

In the standard thought experiment, it is obvious that the scientists would have their hands full providing the nerve stumps from all your senses with just the right stimulations to carry off the trickery, but philosophers have assumed for the sake of argument that however technically difficult the task might be, it is 'possible in principle'. One should be leery of these possibilities in principle. It is possible in principle to build a stainless steel ladder to the moon, and to write out in alphabetical order, all intelligible English conversations consisting of less than a thousand words. But neither of these are remotely possible in fact and sometimes an impossibility in fact is theoretically more interesting than a possibility in principle, as we shall see. (1992, p. 4)

Wittgenstein, Quine, Wiggins and Dennett warn against this style of argument but, despite these warnings, there is a lot of it about. Clearly this method of arguing is one of the mainstays of philosophical debate.

3. CRITICIZING THE METHOD: TWO APPROACHES

Quine and Wittgenstein favour the view that the method is faulty because it asks us to apply a concept in a novel setting. They argue that once we

move away from the ordinary uses of concepts we simply do not know what to say about the legitimacy of any novel application.

The other approach asserts that the results we get by testing our intuitions within unfamiliar scenarios are unacceptable because the way in which the scenarios are constructed tends to beg the question. This is what Wiggins is suggesting when he says that the method may yield a situation in which the conclusion we reach becomes "a matter not of discovery but of interpretation (or even stipulation)".

Wiggins' point here is that if, through thought experiments, we move too far away from the norm (if we "denature the human subject") then our criterion of identity for this denatured subject will be unclear. In such a case we would not have any intuitions to fall back on when trying to decide whether, for example, the two subjects which remain after a fission process, are identical. Instead we will be forced to decide the matter by stipulating what the criterion of identity is to be in such cases. Thus we may stipulate that if two human beings are in different spatial locations then they cannot be identical. Alternatively, we may stipulate that if two human beings share the same memories and dispositions then they are identical. Thus Wiggins' complaint is that thought experiments do not help us to discover anything about the criterion of human identity: they simply force us to stipulate what shall be the ruling criterion. Such a stipulation effectively begs the question and is thus an unacceptable philosophical technique.

Dennett's criticism of Frank Jackson's Mary focuses on the second approach: viz., that the method begs the question. In Jackson's scenario, Mary is supposed to have all the physical information about colours. However, having all the physical information about colours means just that, all. And when we actually try to think of what having all the physical information would amount to, perhaps (*contra* Jackson) we might suspect that Mary would know all about qualia. When the imaginary scenario ("[being] brilliant ... she acquires, let us suppose, all the physical information ... " [1992, 398–400]) is thus properly taken on board, the imaginary case ceases to persuade us either way. We then appreciate that we actually have no idea what having all the information would be like: maybe we would know what the experience of seeing something red would be like prior to our experience of seeing a red object, maybe not. The argument fails as a means of coercing our intuitions about the status of qualia because – properly considered – the premise of the argument ("she acquires, let us suppose, all the physical information") turns out to be quite useless as a means of grounding our intuitions.

4. PARFIT'S RESPONSE

Parfit responds to this second way of criticizing the method (viz., that the method begs the question) as he attempts to rebut the first way of criticizing the method. (This first way involved the Wittgensteinian charge that the method of presenting imaginary cases is illegitimate. It is illegitimate because we do not know what judgment to make, i.e., we do not know whether the concept in dispute can be used in the imagined set-up or not.)

Parfit's rebuttal thus turns on the following point: he notes that both Wittgenstein and Quine assume that "... when considering such cases we [will have] no reactions" (1984, 200). But Parfit asserts that on the contrary "these cases arouse in most of us strong beliefs" (ibid). In other words, the method of imaginary cases does not beg the question because (and this is the claim we will be investigating in this paper): "Though our beliefs are revealed most clearly when we consider imaginary cases, *these beliefs also cover actual cases* and our own lives" (ibid, my italics). In other words, in Parfit's view, the imaginary scenarios always retain some sort of link with the actual cases in which we are interested and are, therefore, a legitimate means of investigating actual cases. In this paper I will criticize Parfit's contention that "these beliefs also cover actual cases" by revealing the illegitimate character of what I shall call the "black box" explanations that provide the link between imaginary cases and actual cases.

To prepare the ground for this criticism I will first discuss the interrelated notions of *physical, logical* and *conceptual possibilities*, in subsections (4.1)–(4.3).

4.1. *Physical Possibilities*

The basic idea of possibility stems naturally from our epistemic limitations: in all but controlled experiments, we simply do not know the laws of nature well enough to be able to say that some particular outcome will eventuate. Thus the very idea of physical possibilities – of an open future – is parasitic upon our ignorance of the laws of nature as they apply in complex situations.

If we think of the universe unfolding apart from epistemic subjects there may actually be no physical possibilities i.e., different ways in which the universe might develop. Thus although we can talk about different futures (because we lack the relevant knowledge) this talk has no force vis-à-vis the actual status of physical possibilities – these might or might not exist: the universe might constantly be branching off, fulfilling certain possibilities in a random way; it might be branching off and fulfilling all possibilities, or it might not be able to branch off in this fashion at all

(there may be only one 'possible' future). It is our ignorance of this fact that gives us the logical room to talk of possibilities. If we were sure that the laws of nature[4] were deterministic we might still wonder about the future since our knowledge of the initial conditions would never be sufficient for us to predict what is going to happen, but we would no longer speak of possibilities as we now do (as events that actually might or might not happen).

4.2. *Logical Possibilities*

These are simply envisaged states of affairs expressed in non- contradictory propositions. Their role in the philosophical method that trades on imaginary scenarios is to guarantee an imaginary scenario a hearing. If the scenario is not obviously self-contradictory (if it does not display the grammar of a contradiction: "x both is and is not y"[5]) then the situation it envisages is logically possible and if logically possible then . . . . Here the unspoken assumption is that logical possibilities are at least candidates for being physical possibilities, i.e., 'possible possibilities' in Wiggins' phrase.

4.3. *Conceptual Possibilities*

These are exemplified in typical imaginary scenarios, e.g., the possibility of a man with two heads; or of a tree that grew to be ten thousand feet tall. The situations described in such scenarios are unlikely as physical possibilities but we nevertheless regard them as not being ruled out altogether (as physical possibilities) on the grounds that they are logical possibilities i.e., not obviously self-contradictory. For this reason we at least seem to understand such unlikely conceptual possibilities.

By contrast we would not understand a Russellian proposition like "Procrastination drinks bells". However, since this sentence does not contain a formal contradiction, it is, apparently, a logical possibility, but, despite this, it does not make any sense. This shows us the clear difference between conceptual possibilities and logical possibilities and it also shows us the order of their dependence. To express a logical possibility, a sentence must not contain a contradiction. But in order to assess whether or not it does contain a contradiction we must first be able to understand the sentence. To understand a sentence is simply to see that it expresses a conceptual possibility. 'Procrastination drinks bells' does not express a conceptual possibility because the rules governing the employment of the individual concepts which occur in this sentence are incompatible with the combination presented. Anyone uttering such a sentence who thought that it made sense – i.e., that it was a conceptual possibility – would have to be con-

strued as making some category revisions, e.g., thinking that bells are a kind of drink, (whisky perhaps), that procrastination is a character from Pilgrim's Progress, etc. Thus logical possibilities cannot be assigned to a proposition unless the sentence that expresses the proposition is itself conceptually possible.[6]

To sum up: a conceptual possibility which is also a logical possibility (i.e., does not contain a contradiction) may or may not be physically possible. But as a conceptual possibility *and* a logical possibility it is a candidate physical possibility. Again, this is what I take Wiggins to mean by a 'possible possibility'.

5. BLACK BOXES

We are now in a position to explain how black boxes are related to imaginary scenarios. To do this we need to consider the relationship between conceptual possibilities and physical possibilities. One way of approaching this relationship is to try to pinpoint the exact stage at which variations on a proposition which is conceptually possible begin to envisage situations which are not physical possibilities and then see what happens.

Take the case of a man with two heads: is that physically possible? Well, yes, surely it is. A freak of nature, but definitely a possibility. Or is it? There have been Siamese twins certainly but a true, two-headed man? Perhaps we feel unsure here but we would not want to rule the matter out of court: to simply declare that such a phenomenon is physically impossible.

What about a man with three heads? Well, if two heads are physically possible, surely three are as well. Now, to make a long story short: what about a man with 365 heads? We can envisage the situation described in a cartoon-like fashion at least. Is the situation described therefore a physical possibility? It certainly seems unlikely. Is this conceptual possibility then simply a logical possibility? And what kind of possibility is envisaged when we declare it to be a logical possibility? Surely all we mean here is that the proposition is not self-contradictory. So is this all that we mean when we say there could be a man with 365 heads viz., that the conceptual possibility is not self-contradictory? Do we not have to give some sort of physical account of the possibility envisaged in such cases if they are to be taken seriously in philosophical debate?

I suggest that we do feel the need to supply such an account and that (typically) we satisfy this need by making up a "black box" story. It is this black box which "magically" turns the conceptual possibility (our imaginary scenario which is in danger of being dismissed as a mere logical possibility) back into a physical possibility (and hence a 'possible possibility' – a real possibility as we say). Suppose, for example, we

imagine the man to be living in the year 3995 and to have 365 cloned heads that he exchanges, one for another, each morning of the year with the help of a skilled surgical team using instruments and techniques far in advance of our present technology.

We like this answer since it provides a black box – advanced surgery – which explains how the envisaged conceptual possibility (demoted briefly to the status of a mere logical possibility) could actually be realized. It thus allows the imaginary scenario to become a physical possibility once more. As such, the imaginary scenario retains the modal status (that of a physical possibility) it needs in order to be able to affect our intuitions and thus serve its philosophical purpose.

A black box explanation is thus offered in order to allow the imaginary scenario to be regarded as a physical possibility and, as such, carry some weight in a philosophical discussion. But do black box explanations really turn imaginary scenarios into physical possibilities?

How do we distinguish between those situations in which we only seem to understand the possibility of the scenario (via the black box) and those when we actually do understand it (cases in which the black box is employed legitimately)? This is the central problem because the usefulness of employing imaginary scenarios in philosophical debate is a function of our belief that they are intelligible and that, therefore, anyone who listens to our scenarios will understand them and have his or her intuitions altered as a consequence. So these imaginary scenarios must "make sense" if they are to convince our interlocuters. I have suggested that a crucial aspect of their "making sense" is that the scenarios which they imagine should be regarded as physical possibilities. In what follows I use a simple example to show the vital role which black boxes play in this process.

6. THE CRUMPLED FENDER

Consider a conceptual possibility which could not – via some more or less plausible black box scenario – be physically instantiated. Would it no longer be regarded as a conceptual possibility? As a consequence, would the proposition which states it come to be regarded as conceptually unintelligible and thus be useless as an intuition pump? To help determine the answer, consider the following situation.

If you tell me that your car was in an accident, that its right front fender was buckled, but that the paint job is still intact, the sentence sounds conceptually intelligible. (It is not as if you had just said a piece of Russellian nonsense.) It sounds intelligible because – although the scenario goes against my general background knowledge of (in this instance) the behaviour of paint under stress – I can, in such a case, presume that some

unusual conditions might well have been present which prevented the chipping of the paint (perhaps it is a special paint, or the fender is made of rubber, etc.).

Nevertheless, despite the fact that I have treated the sentence as one I can understand for the reasons suggested, I do not as yet actually understand how the situation (crumpled fender/paint job intact) it envisages could exist – given my knowledge of the fragility of ordinary paint. In other words I will want to hear about the unusual or special conditions surrounding this phenomenon which would explain its physical possibility. So I might say to you – seeking to make your proposition conceptually intelligible – "Is it a special paint?". With this question I am saying, in effect: offer me a "black box" explanation (e.g., "the new models have this special paint that doesn't chip when stressed"). so that I can understand how what you say could be true even if I do not understand the physical reasons why this special paint does not chip (i.e., even though I cannot get inside the black box). However, if you answer: "No, it is just ordinary car paint". then I will be baffled. I will then pursue the matter: "Was there then, something else about the prevailing conditions which explains why the paint was not chipped?" If you reply in the negative I will then say that I do not understand how what you say could possibly be true. Indeed at this point it would be normal for you to say: "Well there must be something which would explain it". Then we could both shake our heads in bafflement and let the matter rest. But this agreement that something unknown must have been responsible is important: this something (the ultimate black box) underwrites the conceptual possibility of the sentence describing the crumpled fender and the intact paint job.

I contend that this black box, however vague, is always a crucial element in such an imaginary scenario. Without it, I will not find the scenario conceptually intelligible. If I am denied the black box explanation, such a story will seem (ultimately) to involve a logical contradiction. We can tease out the contradiction as follows: if ordinary paint is stressed, it chips; your fender was stressed; you say that it was painted with ordinary paint; it must therefore be chipped; you say it is not. But then I am baffled: is it ordinary paint or not? It can't be since it didn't chip. Well then it must be special paint but you deny this. If you won't allow the situation to become conceptually intelligible via a black box you will have to admit that what you are saying no longer makes sense. We simply do not understand what is meant when someone says something that amounts to: 'x is both y and not y'. Here the logical form of the sentence tells us that we are blocked from any attempt to envisage the situation. In Wittgenstein's language "we use the perceptible sign of a proposition (spoken or written, etc.) as a

projection of a possible situation. The method of projection is to think of the sense of the proposition" (3.11 *Tractatus*) It is in this sense that we do not understand what is meant by (cannot make sense of) a sentence stating a logical impossibility. (Of course we recognize its logical form as involving a contradiction, and in this sense we can make sense of it, but this is clearly a different sense of sense.)

It might seem obvious that this view is false if we consider the following line of argument. One might feel that one was only in a position to say that some situations are not compatible with the laws of nature due to the fact that one can understand the sentences with which they are described. ('My fender was crumpled, the paint was intact', etc.) Thus the assertion that a sentence – given our laws of nature – is necessarily false is quite different from the assertion that it is without meaning.

Now I acknowledge that there is a difference between the intelligibility of a sentence which is necessarily false – given our laws of nature – and one which has no meaning because the concepts which are involved do not mesh properly – (e.g., 'Procrastination drinks bells'). However, the point I am making depends on the fact that the initial intelligibility of a sentence (which describes a situation which is impossible according to the laws of nature – as in the 'crumpled fender/paint intact' case) is underwritten by the fact that I assume that a black box explanation of some sort will be offered upon request. If no such offer is forthcoming – if my request for an explanation is refused – then, in the end, I will not understand what the person means by his or her sentence: I will not be able to 'project the possibility' involved. (In effect, the concepts being used will not mesh properly and the result will be that I will not understand what the person means by the sentence.)

This lack of understanding will emerge because the rules by which our concepts operate are – ultimately – derived from nature's laws ("your fender is crumpled, you will need a new paint job", makes sense because of a background understanding of the laws of nature: e.g., brittle materials alter under stress). If one straightforwardly insists that a sentence which flouts the rules governing the concepts involved (and thus flouts the rules of nature) still makes sense, then the question must be: "What sort of sense does it make?".

The point is worth pursuing: when I consider the nonsense sentence, ("Procrastination drinks bells") I call it meaningless, not because I do not understand how to use the individual concepts represented by the words in the sentence. What I do not understand is how these concepts are supposed to fit together in the given instance. As I understand them, the rules governing these concepts do not allow them to fit together in this way: used

in this way they make no sense. To make sense, a sentence must guide us in envisaging some possible situation which would allow us to determine whether the sentence could be true. This is what I did when I made sense of "Procrastination drinks bells" (i.e., thinking that bells is a kind of drink, that Procrastination is a character from *Pilgrim's Progress*, etc.). Thus making sense of a sentence fundamentally involves envisaging possibilities – conceptual possibilities. But envisaging conceptual possibilities ultimately involves reference to physical possibilities.

The reason for this can be explained with the help of Wittgenstein's views on meaning. This is familiar territory so I will be brief: learning a language involves instruction by others. I acquire an understanding of the rules governing the use of various words (and thus learn the meaning of their associated concepts) by gaining knowledge of the criteria which govern their use. These criteria are public criteria. As such they conform to the laws of nature. Thus I learn that nothing can be said to 'drink' unless it is a certain sort of living creature. I am shown examples of horses, children, dogs, etc., drinking and catch on to the rules governing this verb. If someone then says: "Procrastination drinks bells" I do not understand what is meant, and I do not understand because no one could teach me how to use the word 'drink' in this way. Thus it is not a conceptual possibility because the laws of nature rule out the physical possibility of realizing this scenario: there are no public instances of abstract nouns drinking, so I could not be taught how to use the concept 'drink' in this way.

This truth about learning concepts applies universally: I learn all my concepts in public, so to speak. I thus understand a given sentence only if it presents a possibility which could be realized in the public arena. This means that any such realization must conform with the laws of nature. Thus the theory of meaning put forward implicitly in my argument is a verification theory of meaning – derived from the direct implications of Wittgenstein's 'meaning is use' interpretation of the how words acquire meaning.

This argument underpins my approach to the question of the legitimacy of thought experiments. The black box which is so often invoked in imaginary scenarios is crucial to their intelligibility because it provides an implicit promissory note: namely, that an explanation could be provided as to how the concepts used in the scenario might be taught to someone else. But such teaching necessarily assumes a public domain in which the laws governing this domain are independent of the speakers. (If not, there could be no consistency in applying criteria and hence no way to establishing rules for the use of the relevant concepts.) These laws determine what is

possible, and they therefore underwrite the possibility of teaching the use of the relevant concepts.[7]

To conclude this section: the idea that the intelligibility of a conceptual possibility (in the case of imaginary scenario) hinges ultimately on the availability of a black box explanation which would explain its physical possibility, rests on the assumption that our basic idea of possibility is that of a physical possibility. The possibility of conceptual possibilities is thus parasitic upon the notion of physical possibilities. If – in imaginary scenarios – conceptual possibilities lose their black box connection with this realm of physical possibility, they risk becoming conceptually impossible, i.e., unintelligible.

Thus this method of argument – employing an imaginary scenario to support a particular understanding of how a certain concept ought to be applied – is to be recognized as suspect by the offering of a black box explanation which supposedly "explains" the physical possibility of the scenario. (Where no black box is offered – because it is obvious how the conceptual possibility mooted in the scenario could be realized physically – the imaginary scenario is perfectly legitimate and amounts to no more than offering a counter-example.)

7. JUDGING THE LEGITIMACY OF BLACK BOX SCENARIOS

Are any scenarios which employ a black box acceptable? Everything here depends on whether or not the black box which is offered involves a plausible extrapolation from principles that we already understand to some extent. Thus in the case of the crumpled fender example, if – instead of describing the paint simply as 'special' – I say something about the paint containing long chain polymers whose elasticity allows them to stretch under impact, then I have some sense of how the paint can remain undamaged under stress. This crumpled fender incident now becomes a conceptual possibility because I have some sort of line on how the physical possibility of the scenario is to be understood.

In general, distinguishing legitimate black boxes from their illegitimate cousins, turns on a careful assessment of whether we really understand how what is suggested might be possible (again, Wiggins' phrase comes to mind: assessing such scenarios amounts to an assessment of the "possibilities of possible possibilities"). Through considering a variety of examples, one gradually acquires a nose for where the black box lies and whether it is legitimate or not.

To illustrate: suppose we are offered a scenario of the form: "Imagine a computer which is able to pass the Turing test. Would you then regard it as morally wrong to unplug it, or wipe out its memory?". Where is the black

box here? Well, if we take the Turing test to be the test for identifying an agent who is indistinguishable from a human being (when communicating with it, say, over a teletype machine) then, by definition, if the agent passes the Turing test, it deserves the moral regard which we accord to human beings – no problems. The point is, all the work in such a 'test of our intuitions' has been done by the supposition that we understand what it means to say that the computer has passed the test. (This is the black box.) But could it pass? Do we really understand what would be involved for this possibility to be realized? Can we just be told that it has been realized and then consult our intuitions about the moral implications?

To actually be persuaded by such a scenario, would we not have to test out this intuition by conducting the Turing test ourselves? And here we find that we are at a stand: we have to accept that we do not know from our experience what it would be like for a computer to pass the Turing test as far as we are concerned. Thus to determine whether it is a possibility for the computer to pass the test we must determine whether it would be possible for us to pass this judgment on it. And it is just not clear whether it would be possible to make such a judgment. (For example, is there any time limit on the conversation before you have to render a verdict? Might not the very next question that you would have asked have been answered in a way that revealed the non-human nature of the interlocutor?) This uncertainty about what would count as passing the test is what constitutes the irony of the Turing test: it is a test which (perhaps?) cannot be put to the test. Here the black box was subtly concealed in the notion of *passing* the Turing test. (In the case of Dennett's criticism of Jackson's Mary, it is concealed, as we saw, in the notion of acquiring *all* the relevant knowledge: see above, p. 183.)

There does not seem to be a general rule for assessing the legitimacy of the black box. However, this assessment always turns on asking yourself whether you really understand how the suggested scenario could be realized. (It is only this understanding that will allow the story to actually influence your intuitions with regard to the topic which the imaginary scenario involves.) Sometimes this process involves understanding how the scenario could be realized physically (via some plausible extrapolation of principles with which we are familiar – an extrapolation which does not outrun our understanding (as the brain in the vat outruns 'advanced surgery' in Dennett's exposition of its implications: see above pp. 2–3). Sometimes it involves paying attention to what a term like 'passing' actually means as in the case of the Turing test (or what the word 'deception' means in the case of Descartes' demon). In such cases it involves recognizing that we, in fact, lack the appropriate criteria for employing the term in the imaginary scenario which we are offered.

8. THE EXCEPTION THAT PROVES THE RULE: THE CASE OF HUME

Hume does not bother with "black box" explanations to back up his imaginary scenarios when he does philosophy. For example, when he lays the foundations for his argument concerning causality, Hume simply stipulates that what can be separated by the imagination can be separated in existence. To state a conceptual possibility – in Hume's language, to think of or conceive of or imagine a certain state of affairs – is all that you need to do to justify its physical possibility. For example, I can imagine a leaf fluttering to the ground without thinking of the gust of wind (or anything else) which tore it loose from its branch. Therefore it is physically possible for the leaf to flutter to the ground without anything happening to it beforehand. And, according to Hume, I do not need to provide any explanation – black box or otherwise – of how such a thing could actually happen.

Hume insists that when we exclude causes we really do exclude them. He stipulates this as follows: "... it is not contradictory or absurd to separate the idea of a cause from that of a beginning of existence" (*Treatise of Human Nature*, p. 80).

He supports this stipulation with another, namely, his statement that anything we can think of separately could exist separately. For a thing to exist it need only be thought of: 'Whatever the mind clearly conceives includes the idea of its possible existence" (*Treatise*, p. 32). and "The idea of the existence of an object is the same as the idea of the object" (*Treatise*, p. 66).

Since a typical explanation focuses on why or how something came to be and since such explanations are not required in Hume's system, conceptual possibilities (which are "not contradictory or absurd" – i.e., not logically impossible or nonsense in Russell's sense) are promoted to the status of physical possibilities forthwith: if you can say it (or think it or conceive it or imagine it), the situation it envisages could exist.

Now despite the fame of Hume's analysis of causality, you must swallow pretty hard in order to accept the stipulations which underpin his argument. To avoid this, philosophers before and after Hume have always paid some deference to the black box when they themselves move from conceptual possibilities to physical possibilities. They sense that it is necessary to justify the claim regarding the conceptual possibility of some situation by offering some explanation of how the envisaged state of affairs could be brought about physically.[8] We should note here that *physical* black box explanations can be leap-frogged using theological or psychological black boxes. In such cases the "mechanism" (e.g., a demon with the powers of a god; a telepath; a clairvoyant) which allows conceptual possibilities to become real possibilities is not a physical black box but it serves the

same purpose. It allows us to understand how the suggested conceptual possibility could be a real possibility as opposed to simply a logical possibility.

9. WHAT IS WRONG WITH THIS "ANYTHING GOES" METHOD OF DOING PHILOSOPHY?

It is sometimes defended as a way of testing our intuitions. Wiggins makes this suggestion when he notes that Parfit thinks that thought experiments which stretch/challenge our intuitions are "in some obscure way good for us" (1980, p. 178). But does this method really test our intuitions? We need to look at a few cases to find out.

Descartes' "evil demon" case is familiar to most readers. Here the intuition we are testing is: does the ordinary case of say, my experience of myself as sitting at my desk tapping out these words on the keyboard, constitute an instance of knowledge? It is suggested to us that we may be being deceived about such ordinary cases of knowing by a malignant demon with the powers of a god. This is the black box that promotes the suggested conceptual possibility to the status of a real possibility and thereby legitimates the doubt which is being fostered. Now: are our intuitions about the application of the concept "knowledge" thereby altered?

Everything seems to depend on our willingness to accept the black box explanation whose hidden premise – "To a god all things are possible" – is true by definition. The definition is hard to dispute: it is, after all, simply a definition. To dispute its force (as a means of converting a conceptual possibility into a real possibility) you must refuse to go along with idea that such a god could exist. For if you accept that the god could exist then all things are possible unto that being and I may, therefore, be being deceived about the most ordinary cases of knowledge. Hence my intuitions about what counts as a case of knowledge will be – on this hypothesis – shaken. The interesting question then becomes: Does such a hypothetical shaking of my intuitions actually shake them?

Does the actual shaking not depend on the plausibility of the hypothesis? But by definition there is not a shred of evidence that the demon exists. What then do I mean when I say that such a being could exist? The answer I have proposed is that I do not actually understand what it means to say these things: where conceptual possibilities cannot be understood as physical possibilities (through projections which follow basic or derived laws of physics) I only seem to understand such possibilities – and the midwife of this seeming to understand is the mention of a black box.

10. SUMMARY

At the beginning I mentioned Wiggins' interesting characterization of the problem of imaginary scenarios, viz., that they involved us in the task of working out the "possibilities of possible possibilities". I have construed this phrase in the following way: to say that something is a possible possibility is to say that the situation envisaged could be understood to be physically possible – if the black box we are using had the powers which it is stipulated to have.

Possible possibilities are usually easily accepted since we seem to understand the possibility mentioned due to the presence of the black box (*sotto voce*: "We might well be being deceived now about the most ordinary things". "*How*?" "Suppose a demon with the powers of a god". "*Ooooh*!"). But to finally destroy this sense of seeming to understand (fostered by the black box explanation) we need to apply the Wittgensteinian thumb-screws. (My presumption here is that some variation of the following process could be applied to any black box explanation and thus render unintelligible the possibility of the imaginary scenario in question. Dennett's treatment of Jackson's qualia argument is an example of such a variation.)

11. TORTURE TIME

What are the criteria for picking out cases of demoniac deceptions?

Sorry there are none, by definition.

So what makes you think that such a deception might be going on at this very moment?

Nothing, it just could be that's all.

In what sense could it be, in what sense is this possibility possible?

In the sense that I thought up this possibility via the notion of a demon, a god. Such a being can do anything.

Is there any way to tell whether it is doing some deceiving right now?

No. It might be or it might not. You just can't tell.

So this talk of a demon doing the dirty is just talk then, pure supposition?

Yes, but that is all that is needed to foster doubt.

What must be present in order to doubt something?

The idea that things might be different from what they seem to be.

Where does this idea come from?

From being fooled.

How do you find out that you have been fooled?

Over time you discover that what you took to be the case is not actually the case, as in the instance of a mirage.

What would it be like to discover that you had been fooled by the demon?

Perhaps suddenly finding yourself on some darkling plain with the rest of the population listening to a demoniac voice shouting "I fooled you all".

But what if your neighbour then turned to you and said: "I wonder if this current situation is still a part of the demon's deception or do you think this is finally the real thing?"

I must admit I wouldn't know what to say. There would be no point in saying: "We will have to wait and see whether this is just one more deception". That just leads to an infinite regress. We have no way of distinguishing realities from deceptions on the evil demon hypothesis.

How then could you think that you might be being fooled by him now?

Well I don't have any reason. It is not as if there is some obvious discrepancy in my experience that I could point to, thereby raising the suspicion that I am being deceived. But I can nevertheless imagine my being fooled by simply adopting the point of view of the demon: it, after all, would know whether I was being fooled or not; it can distinguish between the real world and any deception it imposes on me.

How does the demon make this distinction?

Well, the demon could simply stipulate which of the worlds that I experience is to be counted as the real one.

So the demon does not need to rely on some method of distinguishing between a real world and one which is deceptive, nothing analagous, e.g., to our reliance on the constancy and coherence of the real world which serves as a criterion for its reality and serves to distinguish it from the temporary deceptions which are revealed as such by the relatively brief periods in which they fool us into taking them for reality?

That's right. The demon has the powers of a god – by definition. What it says, goes: whatever set of experiences is nominated by the demon as the real world is the real world.

So, from the demon's point of view, that little hallucination that you detected the other day could have been the real world and all of this ordinary day-to-day experience that we think of as the experience of reality could be the deception?

Yes, in other words, the demon's criterion for distinguishing what is real from what is deceptive bears no relation whatsoever to our criterion.

So, in the light of this fact, why do you insist on saying that we might be being deceived by the demon, when, from your own point of view, you have admitted that you have no way of telling whether you are being deceived and from the demon's point of view the notion of deception bears no relation to our own notion of being deceived?

I don't know how I manage to continue to entertain the notion that I am being deceived by a demon, just by saying the words in a loose way I suppose; in Wittgenstein's lovely phrase "by letting language go on a holiday". That is how such possible possibilities seem to make sense.

12. THE MORAL

Beware the spinner of imaginary scenarios; he has a question to beg.

NOTES

[1] My thanks to the anonymous referees for a number of very helpful suggestions.

[2] Some other familiar examples are: Locke's soul of a prince in a cobbler's body; Parfit's teleportation scenarios; Hume's imagining an effect existing without a cause (discussed below); supposing the universe to have doubled in size overnight; brains in vats; brain interchanges; carbon copy replications or duplications of human beings; perfect copies of works of art; worlds where water is not H_2O but rather XYZ; Swampman; and Frank Jackson's Mary (discussed below).

[3] This is a phrase which I will discuss at several points in the paper.

[4] A fourth sort of possibility is worth mentioning at this point: philosophers often say of the laws of nature that they could change. This is what we might term a metaphysical possibility. However, within this new set of laws the notion of a possibility would still be confined to the idea of a variation in the unfolding of events which those new laws of nature would permit. So the basic idea of possibility is still tied to physical possibilities, whatever the physics might be. Needless to say, the sense of possibility fostered by a fanciful physics is unlikely to change our intuitions on any substantive matter (see Section 9).

[5] Sometimes the contradiction will be implicit (i.e., not apparent in its grammar) and in this case the sentence may seem to pass muster as a logical possibility when it is not. I consider such a case below in the section entitled 'The Crumpled Fender'.

[6] I said above (of the proposition: 'Procrastination drinks bells'): "since it does not contain a formal contradiction, it is – apparently – a logical possibility, but despite this, it does not make any sense". *Apparently* is the key word here for, strictly speaking, we cannot assess the logical possibility of a proposition unless we understand it, i.e., unless it is conceptually possible. So 'procrastination drinks bells', is simply a non-starter in terms of its status as expressing a logical possibility.

[7] This is why the arbitrary nature of the demon's distinction between deception and reality (which is mentioned in the concluding dialogue (under the sub-heading 'Torture Time') is, in effect, another black box, since the demon could not teach anyone his use of deception – there is no public domain between us – and therefore we cannot understand his use of this term except as a mere stipulation.

[8] In an amazing passage (amazing to me in the context of this paper) which I came across

recently, Kant spells out in some detail the importance of restricting the notion of possibility to physical possibilities:

"But if we should seek to frame quite new concepts of substances, forces, reciprocal actions, from the material which perception presents to us, without experience itself yielding the example of their connection, we should be occupying ourselves with mere fancies, of whose possibility there is no criterion since we have neither borrowed these concepts [directly] from experience, nor have taken experience as our instructress in their formation. Such fictitious concepts, unlike the categories, can acquire the character of possibility not in an a priori fashion, as conditions upon which all experience depends, but only a posteriori, as being concepts which are given through experience itself. And consequently, their possibility must be known a posteriori and empirically, or it cannot be known at all. A substance which would be permanently present in space, but without filling it (like that mode of existence intermediate between matter and thinking being which some would seek to introduce), or a special ultimate mental power of *intuitively* anticipating the future (and not merely inferring it), or lastly a power of standing in community of thought with other men, however distant they may be – are concepts the possibility of which is altogether groundless, as they cannot be based on experience and its known laws; and without such confirmation they are arbitrary combinations of thought, which although free from contradiction, can make no claim to objective reality, and none, therefore, as to the possibility of an object such as we here profess to think. As regards reality, we obviously cannot think it *in concreto*, without calling experience to our aid. For reality is bound up with sensation, the matter of experience, not with that form of relation in regard to which we can, if we so choose, resort to a playful inventiveness". (*The Critique of Pure Reason*, A222–223)

REFERENCES

Dennett, D. C.: 1992, *Consciousness Explained*, Allen Lane, London.
Parfit, D.: 1984, *Reasons and Persons*, Oxford University Press, Oxford.
Wiggins, D.:1980, *Sameness and Substance*, Basil Blackwell, Oxford.
Wittgenstein, L.: 1967, *Zettel*, Basil Blackwell, Oxford.
Wittgenstein, L.: 1961, *Tractatus Logico-Philosophicus*, translation by Pears and McGuiness, Routledge & Kegan Paul, London.

Manuscript submitted March 25, 1994
Final version received May 15, 1995

Department of Philosophy
University of Otago
Dunedin
New Zealand