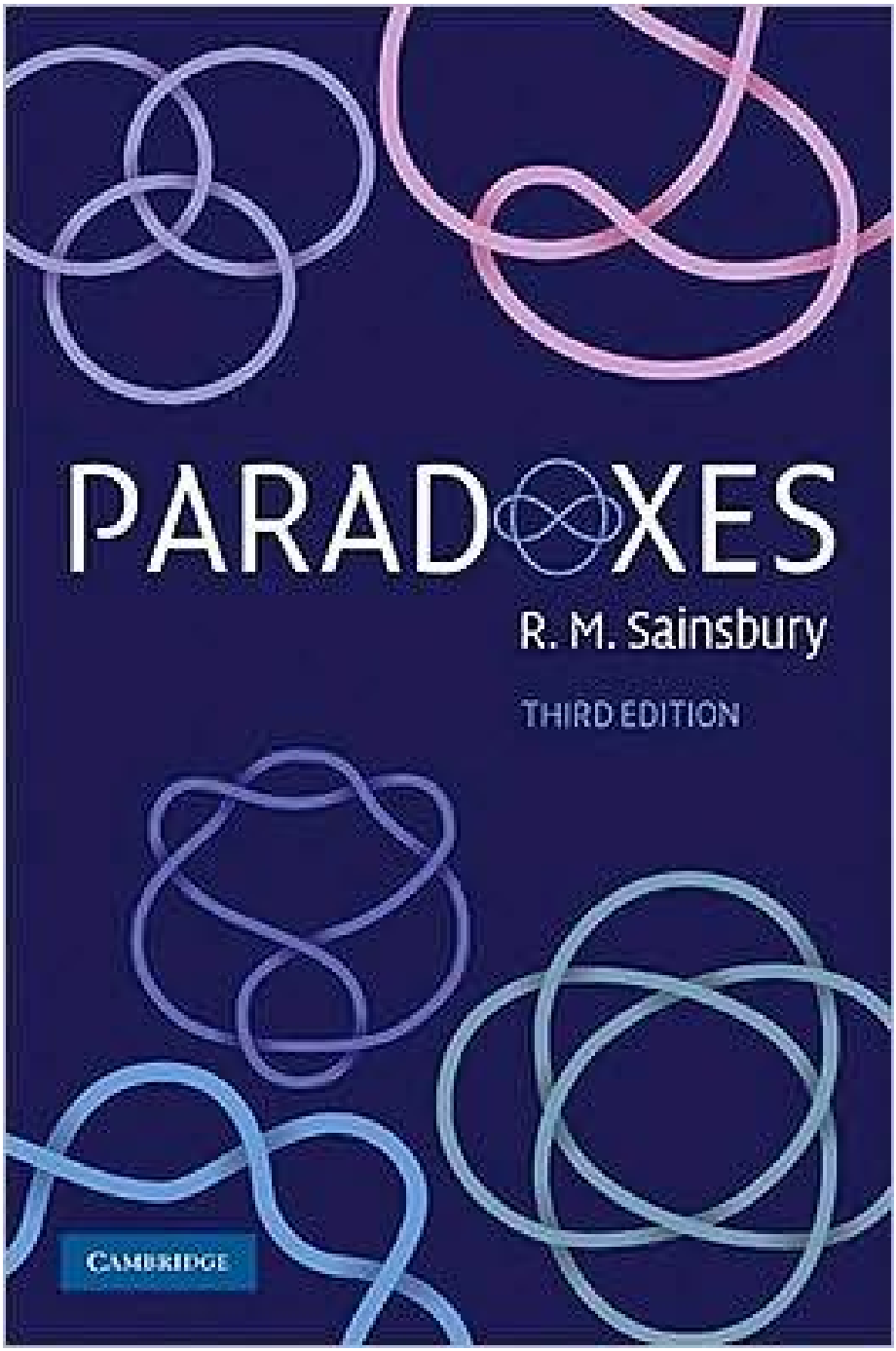
Paradoxes



Paradoxes

THIRD EDITION

R. M. Sainsbury

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521896320

First published in print format 2009

ISBN-13 978-0-521-89632-0 hardback
ISBN-13 978-0-521-72079-3 paperback

Contents

Foreword to third edition

The main change in this edition is the addition of a new chapter on moral paradoxes, which I was inspired to write by reading Smilansky's excellent book *Ten Moral Paradoxes*. To Saul Smilansky I owe thanks for encouragement and comments. The new chapter is numbered 2 and subsequent chapters are renumbered accordingly. I placed the new chapter early in the book in the belief that the discussion is more straightforward than just about any of the other chapters.

I have made some small changes elsewhere, notably to the chapter on vagueness (now chapter 3) and to the suggested reading. Since the second edition appeared in 1995, the internet has transformed many aspects of our life. There are now many websites which help people doing philosophy at every level. The *Stanford Encyclopedia of Philosophy* (plato.stanford.edu) should be the first place to turn if your curiosity has been aroused by this text. Someone with internet access can now do serious research in philosophy, even without the advantage of a university library.

My thanks to Daniel Hill for many useful suggestions for this edition.

Introduction

Paradoxes are fun. In most cases, they are easy to state and immediately provoke one into trying to "solve" them.

One of the hardest paradoxes to handle is also one of the easiest to state: the Liar paradox. One version of it asks you to consider the man who simply says, "What I am now saying is false." Is what he says true or false? The problem is that if he speaks truly, he is truly saying that what he says is false, so he is speaking falsely; but if he is speaking falsely, then, since this is just what he says he is doing, he must be speaking truly. So if what he says is false, it is true; and if it is true, it is false. This paradox is said to have "tormented many ancient logicians and caused the premature death of at least one of them, Philetas of Cos." Fun can go too far.

Paradoxes are serious. Unlike party puzzles and teasers, which are also fun, paradoxes raise serious problems. Historically, they are associated with crises in thought and with revolutionary advances. To grapple with them is not merely to engage in an intellectual game, but is to come to grips with key issues. In this book, I report some famous (and some less famous) paradoxes and indicate how one might respond to them. These responses lead into some rather deep waters.

This is what I understand by a paradox: an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises. Appearances have to deceive, since the acceptable cannot lead by acceptable steps to the unacceptable. So, generally, we have a choice: either the conclusion is not really unacceptable, or else the starting point, or the reasoning, has some non-obvious flaw.

Paradoxes come in degrees, depending on how well appearance camouflages reality. Let us pretend that we can represent how paradoxical something is on a ten-point scale. The weak or shallow end we shall label 1; the cataclysmic end, home of paradoxes that send seismic shudders through a wide region of thought, we shall label 10. Serving as a marker for the point labeled 1 is the so-called Barber paradox: in a certain remote Sicilian village, approached by a long ascent up a precipitous mountain road, the barber shaves all and only those villagers who do not shave

themselves. Who shaves the barber? If he himself does, then he does not (since he shaves *only* those who do not shave themselves); if he does not, then he indeed does (since he shaves *all* those who do not shave themselves). The unacceptable supposition is that there is such a barber – one who shaves himself if and only if he does not. The story may have sounded acceptable: it turned our minds, agreeably enough, to the mountains of inland Sicily. However, once we see what the consequences are, we realize that the story cannot be true: there cannot be such a barber, or such a village. The story is unacceptable. This is not a very deep paradox because the unacceptability is very thinly disguised by the mountains and the remoteness.

At the other end of the scale, the point labeled 10, I shall place the Liar. This placing seems the least that is owed to the memory of Philetas.

The deeper the paradox, the more controversial is the question of how one should respond to it. Almost all the paradoxes I discuss in the ensuing chapters score 6 or higher on the scale, so they are really serious. (Some of those in chapter 2 and in appendix I might be argued to rate a lower score.) This means that there is severe and unresolved disagreement about how one should deal with them. In many cases, though certainly not all (not, for example, in the case of the Liar), I have a definite view; but I must emphasize that, although I naturally think my own view is correct, other and greater men have held views that are diametrically opposed. To get a feel for how controversial some of the issues are, I suggest examining the suggestions for further reading at the ends of chapters.

Some paradoxes collect naturally into groups by subject matter. The paradoxes of Zeno which I discuss form a group because they all deal with space, time, and infinity. The paradoxes of chapter 4 form a group because they bear upon the notion of rational action. Some groupings are controversial. For example, Russell grouped the paradox about classes with the Liar paradox. In the 1920s, Ramsey argued that this grouping disguised a major difference. More recently, it has been argued that Russell was closer to the truth than Ramsey.

I have compared some of the paradoxes treated within a single chapter, but I have made no attempt to portray larger patterns. However, it is arguable that there are such patterns, or even that the many paradoxes are the many signs of one "master cognitive flaw." This last claim has been ingeniously argued by Roy Sorensen (1988).

Questions can be found in boxes throughout the text. I hope that considering these will give pleasure and will prompt the reader to elaborate some of the themes in the text. Asterisked questions are referred to in appendix II, where I have made a point that might be relevant to an answer.

I feel that chapter 6 is the hardest; it might well be left until last. The first and second are probably the easiest. The order of the others is arbitrary. Chapter 7 does not introduce a paradox, but rather examines the assumption, made in the earlier chapters, that all contradictions are unacceptable. I think it would not make much sense to one completely unfamiliar with the topics discussed in chapter 6.

I face a dilemma: I find a book disappointing if the author does not express his own beliefs. What holds him back from stating, and arguing for, the truth as he sees it? I could not bring myself to exercise this restraint. On the other hand, I certainly would not want anyone to believe what I say without first carefully considering the alternatives. So I must offer somewhat paradoxical advice: be very skeptical about the proposed "solutions"; they are, I believe, correct.

Suggested reading

There are now a number of excellent books that deal with a spectrum of paradoxes, in particular Nicholas Rescher (2001) *Paradoxes: Their Roots, Range and Resolution* and Roy Sorensen (2003) *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind*. There is also a surprisingly large amount of material on the web. The following webpage lists a whole range of paradox sites, of very diverse kinds: www.google.com/Top/Society/Philosophy/Philosophy_of_Logic/Paradoxes/.

1 Zeno's paradoxes: space, time, and motion

1.1 Introduction

Zeno the Greek lived in Elea (a town in what is now southern Italy) in the fifth century BC. The paradox for which he is best known today concerns the great warrior Achilles and a previously unknown tortoise. For some reason now lost in the folds of time, a race was arranged between them. Since Achilles could run much faster than the tortoise, the tortoise was given a head start. Zeno's astonishing contribution is a "proof" that Achilles could never catch up with the tortoise no matter how fast he ran and no matter how long the race went on.

The supposed proof goes like this. The first thing Achilles has to do is to get to the place from which the tortoise started. The tortoise, although slow, is unflagging: while Achilles is occupied in making up his handicap, the tortoise advances a little bit further. So the next thing Achilles has to do is to get to the *new* place the tortoise occupies. While he is doing this, the tortoise will have gone on a little bit further still. However small the gap that remains, it will take Achilles some time to cross it, and in that time the tortoise will have created another gap. So however fast Achilles runs, all the tortoise need do in order not to be beaten is keep going – to make *some* progress in the time it takes Achilles to close the previous gap between them.

No one nowadays would dream of accepting the conclusion that Achilles cannot catch the tortoise. (I will not vouch for Zeno's reaction to his paradox: sometimes he is reported as having taken his paradoxical conclusions quite seriously and literally, showing that motion was impossible.) Therefore, there must be something wrong with the argument. Saying exactly *what* is wrong is not easy, and there is no uncontroversial diagnosis. Some have seen the paradox as produced by the assumption that space or time is infinitely divisible, and thus as genuinely proving that space or time is *not* infinitely divisible. Others have seen in the argument nothing more than a display of ignorance of elementary mathematics – an ignorance perhaps excusable in Zeno's time but inexcusable today.

The paradox of Achilles and the tortoise is Zeno's most famous, but there were several others. The Achilles paradox takes for granted that Achilles can start running, and purports to prove that he cannot get as far as we all know he can. This paradox dovetails nicely with one known as the Racetrack, or Dichotomy, which purports to show that nothing can *begin* to move. In order to get anywhere, say to a point one foot ahead of you, you must first get halfway there. To get to the halfway point, you must first get halfway to *that* point. In short, in order to get anywhere, even to begin to move, you must first perform an infinity of other movements. Since this seems impossible, it seems impossible that anything should move at all.

Almost none of Zeno's work survives as such. For the most part, our knowledge of what his arguments were is derived from reports by other philosophers, notably Aristotle. He presents Zeno's arguments very briefly, no doubt in the expectation that they would be familiar to his audience from the oral tradition that was perhaps his own only source. Aristotle's accounts are so compressed that only by guesswork can one reconstruct a detailed argument. The upshot is that there is no universal agreement about what should count as "Zeno's paradoxes," or about exactly what his arguments were. I shall select arguments that I believe to be interesting and important, and which are commonly attributed to Zeno, but I make no claim to be expounding what the real, historical Zeno actually said or thought.

Aristotle is an example of a great thinker who believed that Zeno was to be taken seriously and not dismissed as a mere propounder of childish riddles. By contrast, Charles Peirce wrote of the Achilles paradox: "this ridiculous little catch presents no difficulty at all to a mind adequately trained in mathematics and in logic, but is one of those which is very apt to excite minds of a certain class to an obstinate determination to believe a given proposition" (1935, vol. VI, §177, p. 122). On balance, history has sided with Aristotle, whose view on this point has been shared by thinkers as dissimilar as Hegel and Russell.

I shall discuss three Zenonian paradoxes concerning motion: the Racetrack, the Achilles, and a paradox known as the Arrow. Before doing so, however, it will be useful to consider yet another of Zeno's paradoxes, one that concerns space. Sorting out this paradox provides the groundwork for tackling the paradoxes of motion.

1.2 Space

In ancient times, a frequently discussed perplexity was how something ("one and the same thing") could be both one and many. For example, a book is one but also many (words or pages); likewise, a tree is one but also many (leaves,

branches, molecules, or whatever). Nowadays, this is unlikely to strike anyone as very problematic. When we say that the book or the tree *is* many things, we do not mean that it is identical with many things (which would be absurd), but rather that it is made up of many parts. Furthermore, at least on the face of it, there is nothing especially problematic about this relationship between a whole and the parts which compose it (see question 1.1).

1.1

Appearances may deceive. Let us call some particular tree *T*, and the collection of its parts at a particular moment *P*. Since trees can survive the loss of some of their parts (e.g. their leaves in the fall), *T* can exist when *P* no longer does. Does this mean that *T* is something other than *P* or, more generally, that each thing is distinct from the sum of its parts? Can *P* exist when *T* does not (e.g. if the parts of the tree are dispersed by timber-felling operations)?

Zeno, like his teacher Parmenides, wished to argue that in such cases there are not many things but only one thing. I shall examine one ingredient of this argument. Consider any region of space, for example the region occupied by this book. The region can be thought of as having parts which are themselves spatial, that is, they have some size. This holds however small we make the parts. Hence, the argument runs, no region of space is "infinitely divisible" in the sense of containing an *infinite* number of spatial parts. For each part has a size, and a region composed of an infinite number of parts of this size must be infinite in size.

This argument played the following role in Zeno's attempt to show that it is not the case that there are "many things." He was talking only of objects in space, and he assumed that an object has a part corresponding to every part of the space it fills. He claimed to show that, if you allow that objects have parts at all, you must say that each object is infinitely large, which is absurd. You must therefore deny that objects have parts. From this Zeno went on to argue that *plurality* – the existence of many things – was impossible. I shall not consider this further development, but will instead return to the argument against infinite divisibility upon which it draws (see question 1.2).

1.2

* Given as a premise that no object has parts, how could one attempt to argue that there is no more than one object?

The conclusion may seem surprising. Surely one could convince oneself that any space has infinitely many spatial parts. Suppose we take a rectangle and bisect it vertically to give two further rectangles. Taking the right-hand one, bisect it vertically to give two more new rectangles. Cannot this process of bisection go on indefinitely, at least in theory? If so, any spatial area is made up of infinitely many others.

Wait one moment! Suppose that I am drawing the bisections with a ruler and pencil. However thin the pencil, the time will fairly soon come when, instead of producing fresh rectangles, the new lines will fuse into a smudge. Alternatively, suppose that I am cutting the rectangles from paper with scissors. Again, the time will fairly soon come when my strip of paper will be too small to cut. More scientifically, such a process of physical division must presumably come to an end *sometime*: at the very latest, when the remainder of the object is no wider than an atom (proton, hadron, quark, or whatever).

The proponent of infinite divisibility must claim to have no such physical process in mind, but rather to be presenting a purely intellectual process: for every rectangle we can consider, we can also consider a smaller one having half the width. This is how we conceive any space, regardless of its shape. What we have to discuss, therefore, is whether the earlier argument demonstrates that space cannot be as we tend to conceive it; whether, that is, the earlier argument succeeded in showing that no region could have infinitely many parts.

We all know that there are finite spaces which have spatial parts, but the argument supposedly shows that there are not. Therefore we must reject one of the premises that leads to this absurd conclusion, and the most suitable for rejection, because it is the most controversial, is that space is infinitely divisible. This premise supposedly forces us to say that either the parts of a supposedly infinitely divisible space are finite in size, or they are not. If the latter holds, then they are nothing, and no number of them could together compose a finite space. If the former holds, infinitely many of them together will compose an infinitely large space. Either way, on the supposition that space is infinitely divisible, there are no finite spaces. Since there obviously are finite spaces, the supposition must be rejected.

The notion of infinite divisibility remains ambiguous. On the one hand, to say that any space is infinitely divisible could mean that there is no upper limit to the number of imaginary operations of dividing we could effect. On the other hand, it could mean that the space contains an infinite number of parts. It is not obvious that the latter follows from the former. The latter claim might seem to rely on the idea that the process of imaginary dividings could somehow be "completed." For the moment

let us assume that the thesis of infinite divisibility at stake is the thesis that space contains infinitely many non-overlapping parts, and that each part has some finite size.

The most doubtful part of the argument against the thesis is the claim that a space composed of an infinity of parts, each finite in size, must be infinite. This claim is incorrect, and one way to show it is to appeal to mathematics. Let us represent the imagined successive bisections by the following series:

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{8} \ldots$$

where the first term $\left(\frac{1}{2}\right)$ represents the fact that, after the first bisection, the right-hand rectangle is only half the area of the original rectangle; and similarly for the other terms. Every member of this series is a finite number, just as each of the spatial parts is of finite size. This does not mean that the sum of the series is infinite. On the contrary, mathematics texts have it that this series sums to 1. If we find nothing problematic in the idea that an infinite collection of finite numbers has a finite sum, then by analogy we should be happy with the idea that an infinite collection of finite spatial parts can compose a finite spatial region (see question 1.3).

This argument from mathematics establishes the analogous point about space (namely, that infinitely many parts of finite size may together form a finite whole) only upon the assumption that the analogy is good: that space, in the respect in question, has the properties that numbers have. This is controversial. For example, we have already said that some people take Zeno's paradoxes to show that space is not continuous, although the series of numbers is. Hence we would do well to approach the issue again. We do not have to rely on any mathematical argument to show that a finite whole can be composed of an infinite number of finite parts.

There are two rather similar propositions, one true and one false, and we must be careful not to confuse them.

(1) If, for some finite size, a whole contains infinitely many parts none smaller than this size, then the whole is infinitely large.
(2) If a whole contains infinitely many parts, each of some finite size, then the whole is infinitely large.

Statement (1) is true. To see this, let the minimum size of the parts be δ (say linear or square or cubic inches). Then the size of the whole is $\infty \times \delta$, which is clearly an infinite number. However, (1) does not bear on the case we are considering. To see this, let us revert to our imagined bisections. The idea was that however small the remaining area was, we could

1.3

Someone might object: is it not just a *convention* in mathematics to treat this series as summing to 1? More generally, is it not just a convention to treat the sum of an infinite series as the limit of the partial sums? If this is a mere mathematical convention, how can it tell us anything about space? Readers with mathematical backgrounds might like to comment on the following argument, which purports to show that the fact that the series sums to 1 can be derived from ordinary arithmetical notions, without appeal to any special convention. (*Warning*: mathematicians tell me that what follows is highly suspect!)

The series can be represented as

$$x + x^2 + x^3 + \cdots$$

where $x = ½$. Multiplying this expression by x has the effect of lopping off the first term:

$$x(x + x^2 + x^3 + \cdots) = x^2 + x^3 + x^4 + \cdots$$

Here we apply a generalization of the principle of distribution:

$$a.(b + c) = (a.b) + (a.c).$$

Using this together with a similar generalization of the principle that

$$(1 - a).(b + c) = (b + c) - a.(b + c)$$

we get:

$$(1 - x).(x + x^2 + x^3 + \cdots) = (x + x^2 + x^3 + \cdots) - (x^2 + x^3 + x^4 + \cdots)$$

Thus

$$(1 - x).(x + x^2 + x^3 + x\ldots) = x$$

So, dividing both sides by $(1 - x)$:

$$x + x^2 + x^3 + \cdots = \frac{x}{(1 - x)}$$

So where $x = ½$, the sum of the series is equal to 1.

always imagine it being divided into two. This means that there can be no lower limit on how small the parts are. There can be no size δ such that all the parts are at least this big. For any such size, we can always imagine it being divided into two.

To see that (2) is false, we need to remember that it is essential to the idea of infinite divisibility that the parts get smaller, without limit, as the imagined process of division proceeds. This gives us an almost visual way of understanding how the endless series of rectangles can fit into the original rectangle: by getting progressively smaller.

It would be as wrong to infer "There is a finite size which every part possesses" from "Every part has some finite size or other" as it would be to infer "There is a woman who is loved by every man" from "Every man loves some woman or other." (Readers trained in formal logic will recognize a quantifier-shift fallacy here: one cannot infer an $\exists\forall$ conclusion from the corresponding $\forall\exists$ premise.)

The explanation for any tendency to believe that (2) is true lies in a tendency to confuse it with (1). We perhaps tend to think: *at the end of the series* the *last* pair of rectangles formed have some finite size, and all the other infinitely many rectangles are larger. Therefore, taken together they must make up an infinite area. However, there is *no such thing* as the last pair of rectangles to be formed: our infinite series of divisions has no last member. Once we hold clearly in mind that there can be no lower limit on the size of the parts induced by the infinite series of envisaged divisions, there is no inclination to suppose that having infinitely many parts entails being infinitely large.

The upshot is that there is no contradiction in the idea that space is infinitely divisible, in the sense of being composed of infinitely many non-overlapping spatial parts, each of some finite (non-zero) size. This does not establish that space *is* infinitely divisible. Perhaps it is granular, in the way in which, according to quantum theory, energy is. Perhaps there are small spatial regions that have no distinct subregions. The present point, however, is that the Zenonian argument we have discussed gives us no reason at all to believe this granular hypothesis.

This supposed paradox about space may well not strike us as very deep, especially if we have some familiarity with the currently orthodox mathematical treatment of infinity. Still, we must not forget that current orthodoxy was not developed without a struggle, and was achieved several centuries after Zeno had pondered these questions. Zeno and his contemporaries might with good reason have had more trouble with it than we do. The position of a paradox on the ten-point scale mentioned in the introduction can change over time: as we become more sophisticated detectors of mere appearance, a paradox can slide down toward the Barber end of the scale.

Clearing this paradox out of the way will prove to have been an essential preliminary to discussing Zeno's deeper paradoxes, which concern motion.

1.3 The Racetrack

If a runner is to reach the end of the track, he must first complete an infinite number of different journeys: getting to the midpoint, then to the point midway between the midpoint and the end, then to the point midway between this one and the end, and so on. Since it is logically impossible for someone to complete an infinite series of journeys, the runner cannot reach the end of the track. It is irrelevant how far away the end of the track is – it could be just a few inches away – so this argument, if sound, will show that all motion is impossible. Moving to any point will involve an infinite number of journeys, and an infinite number of journeys cannot be completed in a finite time.

Let us call the starting point Z (for Zeno), and the endpoint $Z^\star$. The argument can be analyzed into two premises and a conclusion, as follows:

(1) Going from Z to $Z^\star$ would require one to complete an infinite number of journeys: from Z to the point midway to $Z^\star$, call it Z_1; from Z_1 to the point midway between it and $Z^\star$, call it Z_2; and so on.
(2) It is logically impossible for anyone (or anything) to complete an infinite number of journeys.
Conclusion: It is logically impossible for anyone to go from Z to $Z^\star$. Since these points are arbitrary, *all* motion is impossible.

Apparently acceptable premises, (1) and (2), lead by apparently acceptable reasoning to an apparently unacceptable conclusion.

No one nowadays would for a moment entertain the idea that the conclusion is, despite appearances, acceptable. (I refrain from vouching for Zeno's own response.) Moreover, the reasoning appears impeccable. So for us the question is this: which premise is incorrect, and why?

Let us begin by considering premise (1). The idea is that we can generate an infinite series, let us call it the Z-series, whose terms are

$$Z, Z_1, Z_2, \ldots$$

These terms, it is proposed, can be used to analyze the journey from Z to $Z^\star$, for they are among the points that a runner from Z to $Z^\star$ must pass through en route. However, $Z^\star$ itself is not a term in the series; that is, it is not generated by the operation that generates new terms in the series – halving the distance that remains between the previous term and $Z^\star$.

The word "journey" has, in the context, some misleading implications. Perhaps "journey" connotes an event done with certain intentions, but it is obvious that a runner could form no intention with respect to most of the members of the Z-series, for he would have neither the time, nor the

memory, nor the conceptual apparatus to think about most of them. Furthermore, he may well form no intention with respect to those he *can* think about. Still, if we explicitly set these connotations aside, then (1) seems hard to deny, once the infinite divisibility of space is granted; for then all (1) means is the apparent platitude that motion from Z to $Z^\star$ involves traversing the distances Z to Z_1, Z_1 to Z_2, and so on.

Suspicion focuses on (2). Why should one not be able to complete an infinite number of journeys in a finite time? Is that not precisely what *does* happen when anything moves? Furthermore, is it not something that *could* happen even in other cases? For example, consider a view that Bertrand Russell once affirmed: he argued that we could imagine someone getting more and more skillful in performing a given task, and so completing it more and more quickly. On the first occasion, it might take one minute to do the job, on the second, only a half a minute, and so on, so that, performing the tasks consecutively, the whole series of infinitely many could be performed in the space of two minutes. Russell said, indeed, that this was "medically impossible" but he held that it was *logically* possible: no contradiction was involved. If Russell is right about this, then (2) is the premise we should reject.

However, consider the following argument, in which the word "task" is used in quite a general way, so as to subsume what we have been calling "journeys."

> There are certain reading-lamps that have a button in the base. If the lamp is off and you press the button the lamp goes on, and if the lamp is on and you press the button the lamp goes off.
>
> Suppose now that the lamp is off, and I succeed in pressing the button an infinite number of times, perhaps making one jab in one minute, another jab in the next half-minute, and so on, according to Russell's recipe. After I have completed the whole infinite sequence of jabs, i.e., at the end of two minutes, is the lamp on or off? It seems impossible to answer this question. It cannot be on, because I did not ever turn it on without at once turning it off. It cannot be off, because I did in the first place turn it on, and thereafter I never turned it off without at once turning it on. But the lamp must be either on or off. This is a contradiction. (Thomson 1954; cited in Gale 1968, p. 411)

Let us call the envisaged setup consisting of me, the switch, the lamp, and so on, "Thomson's lamp." The argument purports to show that Thomson's lamp cannot complete an infinite series of switchings in a finite time. It proceeds by *reductio ad absurdum*: we suppose that it *can* complete such a series, and show that this supposition leads to an absurdity – that the lamp is neither on nor off at the supposed end of the series of tasks.

The argument is not valid. The supposition that the infinite series has been completed does not lead to the absurdity that the lamp is neither on nor off. Nothing follows from this supposition about the state of the lamp *after* the infinite series of switchings.

Consider the series of moments T_1, T_2, …, each corresponding to a switching. According to the story, the gaps between the members of this T-series get smaller and smaller, and the rate of switching increases. At T_1 a switching on occurs, at T_2 a switching off occurs, and so on. Call the first moment after the (supposed) completion of the series $T^\star$. It follows from the specification of the infinite series that, for any moment *in the T-series*, if the lamp is on at that moment there is a later moment in the series at which the lamp is off; and vice versa. However, nothing follows from this about whether the lamp is on or off at $T^\star$, for $T^\star$ does *not belong* to the T-series. $T^\star$ is not generated by the operation that generates new members of the T-series from old, being a time half as remote from the old member as its predecessor was from it. The specification of the task speaks only to members of the T-series, and this has no consequences, let alone contradictory consequences, for how things are at $T^\star$, which lies outside the series (see question 1.4).

1.4

Are we entitled to speak of "the first moment after the (supposed) completion of the task"?

The preceding paragraph is not designed to prove that it is logically possible for an infinite series of tasks to be completed. It is designed to show only that Thomson's argument against this possibility fails. In fact, someone might suggest a reason of a different kind for thinking that there is a logical absurdity in the idea of Thomson's lamp.

Consider the lamp's button. We imagine it to move the same distance for each switching. If it has moved infinitely many times, then an infinite distance has been traversed at a finite speed in a finite time. There is a case for saying that this is logically impossible, for there is a case for saying that what we *mean* by average speed is simply distance divided by total time. It follows that if speed and total time are finite, so is distance. If this is allowed, then Thomson was right to say that Thomson's lamp as he described it is a logical impossibility, even though the argument he gave for this conclusion was unsatisfactory.

This objection might be countered by varying the design of the machine. There are at least two possibilities. One is that the machine's

switch be so constructed that if on its first switching it travels through a distance δ, then on the second switching it travels δ/2, on the third δ/4, and so on. Another is that the switch be so constructed that it travels faster and faster on each switching, without limit (see questions 1.5, 1.6).

1.5

Does this mean that it would have to travel infinitely fast in the end?

1.6

* Does this mean that the switch would have to travel faster than the speed of light? If so, does this mean that the machine is *logically* impossible?

It is hard to find positive arguments for the conclusion that this machine is logically possible; but this machine is open neither to Thomson's objection, which was invalid, nor to the objection that it involves an infinite distance being traveled in a finite time. Therefore, until some other objection is forthcoming, we can (provisionally, and with due caution) accept this revised Thomson's lamp as a logical possibility. What is more, if *it* is a possibility, then there is nothing logically impossible about a runner completing an infinite series of journeys (see question 1.7).

1.7

Evaluate the following argument:

> We can all agree that the series of numbers $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots$ sums to 1. What is controversial is whether this fact has any bearing on whether the runner can reach $Z^\star$. We know that it would be absurd to say that energy is infinitely divisible merely because for any number that is used to measure some quantity of energy there is a smaller one. Likewise, Zeno's paradox of the runner shows that motion through space should not be thought of as an endless progression through an infinite series. It is as clear that there is a smallest motion a runner can make as it is that there is a smallest spatial distance that we are capable of measuring.

One does not need to establish outré possibilities, such as that of a Thomson's lamp that can complete an infinite number of tasks, in order to establish that the runner can reach $Z^\star$. The argument is supposed to work the other way: if even the infinite Thomson's lamp is possible, then there can be no problem about the runner.

In the next section, I discuss a rather sophisticated variant of the Racetrack. The discussion may help resolve some of the worries that remain with this paradox.

1.4 The Racetrack again

Premise (1) of the previous section asserted that a necessary condition for moving from Z to $Z^\star$ is moving through the infinite series of intermediate Z-points. In this rerun, I want to consider a different problem. It is that there appear to be persuasive arguments for the following inconsistent conclusions:

(a) Passing through all the Z-points is sufficient for reaching $Z^\star$.
(b) Passing through all the Z-points is *not* sufficient for reaching $Z^\star$.

We cannot accept both (a) and (b). The contradiction might be used to disprove the view that the runner's journey can be analyzed in terms of an infinite series, and this would throw doubt on our earlier premise (1) (p. 11).

Let us look more closely at an argument for (a):

> Suppose someone could have occupied every point in the Z-series without having occupied any point outside it, in particular without having occupied $Z^\star$. Where would he be? Not at any Z-point, for then there would be an unoccupied Z-point to the right. Not, for the same reason, between Z-points. And, *ex hypothesi*, not at any point external to the Z-series. But these possibilities are exhaustive. (Cf. Thomson 1954; cited in Gale 1968, p. 418)

In other words, if you pass through all the Z-points, you *must* get to $Z^\star$. Contrasted with this is a simple argument against sufficiency – an argument for (b):

> $Z^\star$ lies outside the Z-series. It is further to the right than any member of the Z-series. So going through all the members of the Z-series cannot take you as far to the right as $Z^\star$. So reaching $Z^\star$ is not logically entailed by passing through every Z-point.

The new twist to the Racetrack is that we have plausible arguments for both (a) and (b), but these are inconsistent.

The following objection to the argument for (a) has been proposed by Paul Benacerraf (1962, p. 774). A possible answer to the question "Where would the runner be after passing through all the Z-points?" is "Nowhere!" Passing through all the Z-points is not sufficient for arriving at $Z^\star$ because one might cease to exist after reaching every Z-point but without reaching $Z^\star$. To lend color to this suggestion, Benacerraf invites us to imagine a genie who "shrinks from the thought" of reaching $Z^\star$ to such an extent that he gets progressively

smaller as his journey progresses. By Z_1 he is half his original size, by Z_2 a quarter of it, and so on. Thus by the time he has passed through every Z-point his size is zero, and "there is not enough left of him" to occupy $Z^\star$.

Even if this is accepted (see question 1.8), it will not resolve our problem. The most that it could achieve is a qualification of (a). What would have to be said to be sufficient for reaching $Z^\star$ is not merely passing through every Z-point, but doing that and *also* (!) continuing to exist. However, the argument against sufficiency, if it is good at all, seems just as good against a correspondingly modified version of (b). Since $Z^\star$ lies outside the Z-series, even passing through every Z-point *and* continuing to exist cannot logically guarantee arriving at $Z^\star$.

1.8

* Can the following objection be met?

> Where is the runner when he goes out of existence? He cannot be at any Z-point since, by hypothesis, there is always a Z-point beyond it, which means that he would not have gone through all the Z-points; but if he goes out of existence at or beyond $Z^\star$, then he reached $Z^\star$, and so the sufficiency claim has not been refuted.

Part of the puzzle here lies, I think, in the exact nature of the correspondence that we are setting up between mathematical series and physical space. We have two different things: on the one hand, a series of mathematical points, the Z-series, and on the other hand, a series of physical points composing the physical racetrack. A mathematical series, like the Z-series, may have no last member. In this case, it is not clear how we are to answer the question "To what physical length does this series of mathematical points correspond?" That this is a genuine question is obscured by the fact that we can properly apply the word "point" both to a mathematical abstraction and to a position in physical space. However, lengths as ordinarily thought of have *two* ends. If a length can be correlated with a mathematical series with only *one* end, like the Z-series, this can only be by stipulation. So if we are to think of part of the racetrack as a length, a two-ended length, corresponding to the mathematically defined Z-series, a one-ended length, we can but stipulate that what corresponds to the physical length is the series from Z to $Z^\star$. Given this, it is obvious that traversing the length corresponding to the Z-series is enough to get the runner to $Z^\star$. On this view, the paradox is resolved by rejecting the argument for (b), and accepting that for (a) – modified, perhaps, by the quibble about the runner continuing to exist.

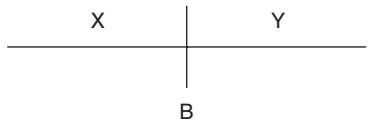


Figure 1.1

This suggestion can be strengthened by the following consideration. Suppose we divide a line into two discrete parts, *X* and *Y*, by drawing a perpendicular that cuts it at a point *B*, as in figure 1.1. The notions of *line, division*, and so on are to be just our ordinary ones, whatever they are, and not some mathematical specification of them. Since *B* is a spatial point, it must be somewhere. So is it in *X* or in *Y* or both? We cannot say that it is in both *X* and *Y*, since by hypothesis these are discrete lines; that is, they have no point in common. However, it would seem that any reason we could have for saying that *B* is in *X* is as good a reason for saying that it is in *Y*. So, if it is in either, then it is in both, which is impossible.

If we try to represent the intuitive idea in the diagram in mathematically precise terms, we have to make a choice. Let us think of lengths in terms of sets of (mathematical) points. If *X* and *Y* are to be discrete (have no points in common), we must choose between assigning *B* to *X* (as its last member, in a left-to-right ordering) and assigning *B* to *Y* (as its first member). If we make the first choice, then *Y* has no first member; if we make the second choice, then *X* has no last member. So far as having an adequate model for physical space goes, there seems to be nothing to determine this choice – it seems that we are free to stipulate.

Suppose we make the first choice, according to which *B* is in *X*. Imagine an archer being asked to shoot an arrow that traverses the whole of a physical space corresponding to *X*, without entering into any of the space corresponding to *Y*. There is no conceptual problem about this instruction: the arrow must be shot from the leftmost point of *X* and land at *B*. Now imagine an archer being asked to shoot an arrow that traverses the whole of a physical space corresponding to *Y*, without entering into any of the space corresponding to *X*. This time there appears to be a conceptual problem. The arrow cannot land at the point in space corresponding to *B* because, by stipulation, *B* has been allocated to *X* and so lies outside *Y*; but nor can the arrow land anywhere in *Y*, since for any point in *Y* there is one between it and *B*. There is no point that is the *first* point to the right of *B*.

What is odd about this contrast – the ease of occupying all of *X* and none of *Y*, the difficulty of occupying all of *Y* and none of *X* – is that *which*

task is problematic depends upon a stipulation. If we had made the other choice, stipulating that B is to belong to Y, the difficulties would have been transposed.

Two real physical tasks, involving physical space, cannot vary in their difficulty according to some stipulation about how B is to be allocated. There is some discrepancy here between the abstract mathematical space-like notions, and our notions of physical space.

If we think of X and Y as genuine lengths, as stretches of physical space, the difficulty we face can be traced to the source already mentioned: lengths – for example, the lengths of racetracks – have *two* ends. If B belongs to X and not Y, then Y lacks a left-hand end: it cannot have B as its end, since B belongs to X and not Y (by hypothesis); but it cannot have any point to the right of B as its left end, for there will always be a Y-point to the left of any point that is to the right of B.

The difficulty comes from the assumption that the point B has partially to *compose* a line to which it belongs, so that to say it belongs to X and Y would be inconsistent with these being non-overlapping lines. For an adequate description of physical space, we need a different notion: one that allows, for example, that two distinct physical lengths, arranged like X and Y, should touch without overlapping. We need the notion of a boundary that does not itself occupy space.

If we ask what region of space – thought of in the way we think of racetracks, as having two ends – corresponds to the points on the Z-series, the only possible answer would appear to be the region from Z to $Z^\star$. This explains why the argument for sufficiency is correct, despite the point noted in the argument against it. $Z^\star$ does not belong to the Z-series, but it does belong to the region of space that corresponds to the Z-series.

In these remarks, I have assumed that we have coherent spatial notions, for example, that of (two-ended) length, and that if some mathematical structure does not fit with these notions, then so much the worse for the view that the structure gives a correct account of our spatial notions. In the circumstances, this pattern of argument is suspect, for it is open to the following Zeno-like response: the *only* way we could hope to arrive at coherent spatial notions is through these mathematical structures. If this way fails – if the mathematical structures do not yield all we want – then we are forced to admit that we were after the impossible, and that there is no way of making sense of our spatial concepts.

The upshot is that a full response to Zeno's Racetrack paradox would require a detailed elaboration and justification of our spatial concepts. This is the task Zeno set us – a task that each generation of philosophers of space and time rightly feels it must undertake anew.

1.5 Achilles and the Tortoise

We can restate this most famous of paradoxes using some Racetrack terminology. The Z-series can be redefined as follows: Z is Achilles' starting point; Z_1 is the tortoise's starting point; Z_2 is the point that the tortoise reaches while Achilles is getting to Z_1; and so on. $Z^\star$ becomes the point at which, we all believe, Achilles will catch the tortoise, and the "proof" is that Achilles, like the runner before him, will never reach $Z^\star$.

We can see this as nothing more, in essentials, than the Racetrack, but with a receding finishing line. The paradoxical claim is this: Achilles can never get to $Z^\star$ because however many points in the Z-series he has occupied, there are still more Z-points ahead before he gets to $Z^\star$. Furthermore, we cannot expect him to complete an infinity of "tasks" (moving through Z-points) in a finite time. An adequate response to the Racetrack will be easily converted into an adequate response to this version of the Achilles.

In such an interpretation of the paradox, the tortoise has barely a walk-on part to play. Let us see if we can do him more justice. One attempt is this:

> The tortoise is always ahead of Achilles if Achilles is at a point in the Z-series. But how is this consistent with the supposition that they reach $Z^\star$ at the same time? If the tortoise is always ahead in the Z-series, must he not emerge from it before Achilles?

This makes for a rather superficial paradox. It is trivial that the tortoise is ahead of Achilles all the time until Achilles has drawn level: he is ahead until $Z^\star$. Given that both of them can travel through all the Z-points, which was disputed in the Racetrack but which is not now challenged, there is no reason why they should not complete this task at the same point in space and time. So I have to report that I can find nothing of substantial interest in this paradox that has not already been discussed in connection with the Racetrack.

1.6 The Arrow

At any instant of time, the flying arrow "occupies a space equal to itself." That is, the arrow at an instant cannot be moving, for motion takes a period of time, and a temporal instant is conceived as a point, not itself having duration. It follows that the arrow is at rest at every instant, and so does not move. What goes for arrows goes for everything: nothing moves.

Aristotle gives a very brief report of this paradoxical argument, and concludes that it shows that "time is not composed of indivisible instants" (Aristotle 1970, Z9. 239b 5). This is one possible response, though one

that would nowadays lack appeal. Classical mechanics purports to make sense not only of velocity at an instant but also of various more sophisticated notions: rate of change of velocity at an instant (i.e. instantaneous acceleration or deceleration), rate of change of acceleration at an instant, and so on.

Another response is to accept that the arrow is at rest at every instant, but deny that it follows that it does not move. What is required for the arrow to move, it may be said, is not that it move-at-an-instant, which is clearly an impossibility (given the semi-technical notion of *instant* in question), but rather that it be at different places at different instants. An instant is not long enough for motion to occur, for motion is a relation between an object, places, and various instants. If a response along these lines can be justified, there is no need to accept Aristotle's conclusion.

Suppose we set out Zeno's argument like this:

(1) At each instant, the arrow does not move.
(2) A stretch of time is composed of instants.
Conclusion: In any stretch of time, the arrow does not move.

Then the response under discussion is that this argument is not valid: the premises are true, but they do not entail the conclusion.

If the first premise is to be acceptable, it must be understood in a rather special way, which provides the key to the paradox. It must be understood as making a claim which does not immediately entail that the arrow is at rest. The question of whether something is moving or at rest "at an instant" is one that essentially involves other instants. An object is at rest at an instant just on condition that it is at the same place at all nearby instants; it is in motion at an instant just on condition that it is in different places at nearby instants. Nothing about the arrow and a single instant alone can fix either that it is moving then or at rest then. In short, the first premise, if acceptable, cannot be understood as saying that at each instant the arrow is at rest.

Once the first premise is properly understood, it is easy to see why the argument is fallacious. The conclusion that the arrow is always at rest says of each instant that the arrow is in the same place at neighboring instants. No such information is contained in the premises. If we think it is implicit in the premises, this is probably because we are failing to distinguish between the claim – interpretable as true – that at each instant the arrow does not move, and the false claim that it is *at rest* at each instant.

If this is correct, then the Arrow paradox is an example of one in which the unacceptable conclusion (nothing moves) comes from an acceptable premise (no motion occurs "during" an instant) by unacceptable reasoning.

Suggested reading

A good starting point is Nick Huggett's *Stanford Encyclopedia* article, available at: plato.stanford.edu/entries/paradox-zeno/.

Salmon (1970) contains the articles by Thomson (1954) and Benacerraf (1962) from which I drew the discussion of infinity machines, as well as many other important articles, including a clear introductory survey by Salmon. It also has an excellent bibliography. For a fine introduction to the philosophy of space and time, including a chapter on Zeno's paradoxes, see Salmon (1980).

For a historical account see Vlastos (1967). For an advanced discussion, see Grünbaum (1967).

The quotation from Peirce, written late in his life, is not representative. In many other places, he discusses Zeno's paradoxes very seriously. However, it is not uncommon for people to see a paradox as trivial once they think they have a definitive solution to it. The cure for this reaction is to try to persuade someone else of one's "solution."

The phrase "medically impossible" comes from Russell (1936, p. 143).

2 Moral paradoxes

2.1 Crime Reduction

Suppose that crimes of a certain category (e.g. car-jacking) are completely eliminated by prescribing an extraordinarily severe penalty (e.g. death). The penalty is so severe that it is 100 percent effective as a deterrent: car-jacking (or whatever crime we consider) never occurs, and so is never punished (so the prescribed severe penalties are never in fact imposed). It seems that we are forced to make conflicting judgments about this imaginary situation:

Good: A crime has been eliminated. There are no bad side-effects: no car-jackers are executed (which might indeed be unjust), for there are no car-jackers.

Bad: A crime has been associated with a punishment of unjust severity. This makes for an unjust society. Even if injustice is a means to a good end (crime reduction) it is still unjust, and should be condemned.

Both views are apparently reasonable; but as they conflict, it appears we cannot hold both.

It may be objected that the imaginary situation is unrealistic, and so need not be taken seriously: we cannot be expected to have consistent judgments about such situations. England tried using capital punishment for petty theft in the seventeenth century, and the deterrent effect was far from complete. Many crimes are committed on impulse, under the influence of alcohol or drugs, out of desperation, or in the false belief that they will go unpunished, and are thus isolated from the threat of penalties. We simply could not eliminate car-jacking by prescribing capital punishment for this crime. Because the situation we are asked to imagine is one that could not obtain, it is not surprising that our responses are inconsistent. To take an extreme case, suppose we were asked for our opinion about a situation that was both good overall and also not good overall. Is it good or not? Clearly we cannot be expected to give a consistent and complete answer.

It is right to say we cannot form consistent and complete views about a logically impossible situation, for example one which is both good overall and not good overall. We should not infer, however, that the same goes for situations which are merely practically, rather than logically, impossible. When the objector says that it would be impossible to eliminate crime by severe penalties, she means only that it would be practically impossible, not that it would be logically impossible. No contradiction is involved in the idea of eliminating car-jacking by these means. We ought to have consistent attitudes toward situations that we know are practically impossible. Suppose we know that it is practically impossible, that is, impossible in practice, to turn lead into gold by purely chemical means. Even so, it would be irrational to think both that it would be, all things considered and overall, a good thing to do, and also that it would be, all things considered and overall, a bad thing to do. Likewise, if we know that the lead-to-gold transmutation cannot be effected, we should not also believe, inconsistently, that it can be effected. We need to be consistent in our attitudes even to states of affairs we know or believe to be practically impossible. So we should have a consistent attitude to the envisaged Crime Reduction scenario, even if we believe that things could never be like that (see question 2.1).

2.1

How would you respond to the following objection?

> So far as the demands of rationality go, there is no good basis for distinguishing between logically impossible situations and practically impossible ones. If a situation is impossible, it cannot obtain, and so a complete set of judgments about it cannot all be true. This is so irrespective of whether the impossibility is logical or practical.

Our objector might amplify her position, drawing on considerations specific to morality. The moral judgments about which we can have any confidence relate to familiar situations: ones in which friends or family lie, or help others, or are cruel, or noble. Familiar situations are ones that actually occur, and so are possible. We cannot expect any confidence in our judgments to extend to completely unrealistic situations. This is not what moral judgments were made for; they cannot be expected to provide reliable views outside their normal sphere of operation.

There is certainly something in this view, but it cannot be accepted as it stands. In deciding what to do, we often have to consider the moral properties of actions we have not performed. If these actions lie beyond our normal sphere of operation, we are not thereby absolved from trying to

form a justified moral view. (Consider how you might act as a US soldier, with no training in interrogation, in charge for the first time of a prisoner you thought had information which could save many of your comrades' lives.) To make the same point from a different direction: fables and fairy stories, and indeed science fiction stories, are often used to suggest moral points of view, applicable to our daily lives, even though the events related are wholly unfamiliar, and perhaps impossible in practice.

The Crime Reduction scenario, even if unrealistic, is not so very unrealistic, and the kinds of considerations in play are of direct practical concern to legislative and judicial bodies. In determining suitable punishments, the presumed deterrent effect has to be taken into account. Whatever reasoning underlies this determination cannot suddenly throw up its hands faced with the claim that some level of punishment will produce complete deterrence. This is a scenario requiring careful thought and reflection, and demanding, as a product, a justifiable view. Hence, even if the Crime Reduction scenario is unrealistic, we cannot be content to endorse both the judgment that it is good, and the judgment that it is bad. To the extent that we have appealing arguments for both judgments, we have a paradox.

Judgments of good and bad are often implicitly relativized: the recent heavy rains were good, in that they replenished our very depleted reserves of water; they were bad, in that they caused flooding, which destroyed crops and houses. There is nothing inconsistent about one and the same thing being good in some respects and bad in others. Is this the situation with our supposedly inconsistent judgments about Crime Reduction? Having a severe penalty is good, in that it reduces crime, and bad, in that it is severe to an unjust degree. No inconsistency; and so no paradox.

In addition to relativized judgments, we are often forced to make "on-balance" judgments, taking into account the possibly competing values of various respects. Action often requires on-balance judgments. For example, we might belong to a legislature considering whether or not to introduce unduly severe penalties; we have to weigh good and bad respects to reach an overall verdict. If we can sustain the conflicting judgments about the Crime Reduction situation as *on-balance* judgments, we have inconsistency. It seems at least arguable that we can do this. A defense of the *on-balance good* judgment might go like this:

> Everyone agrees that a situation in which a certain crime is eliminated, and no harm done, is good at least in that respect. In our case of Crime Reduction, there are no bad respects which need to be weighed against this manifest good. It is true that it would be unjust to carry out the penalty prescribed for the crime. If we had to apply the penalties, then I agree that that would be a bad respect, needing to be weighed against the good respects. While reserving the right to say that the result of

the weighing would still be good on balance, in fact no such weighing is relevant, for in the situation we are asked to envisage, the penalty is never applied. Nothing unjust is done. There is no crime (of the relevant kind), and this is clearly good, and nothing bad or unjust happens.

A defense of the *on-balance bad* judgment might go like this:

Of course it must be agreed that the elimination of crime is a respect in which the Crime Reduction situation is good. But there are plenty of respects in which it is bad. First, it is bad to threaten unjust penalties, even if the threats are so effective that they never have to be implemented. Second, it is bad to have law enforcement officers committed to enforcing an unjust law, even if they never have to make arrests. Third, public awareness of the law would need to be kept at a high level, else it would not have its deterrent effect. In doing this, the authorities would have to soft-pedal the unjust nature of the law, else they could hardly be taken seriously, and this deceptive element would be bad in itself. Finally, given the possibility of wrongful guilty verdicts (especially as a very severe penalty can be used by false accusers for their own ends), the population might live in a state of fear, which would be harmful.

The current dialectic shows something interesting about the nature of paradox, which is likely to be specially conspicuous in moral cases. There is room for disagreement from one person to another on the question whether the crime reduction situation is, on balance, good or bad. There is nothing paradoxical about this. There is moral disagreement about all sorts of things. What would be paradoxical is if a single person felt bound to deliver both of a pair of inconsistent judgments. It seems to me highly unlikely, however, that the arguments given above will seem equally attractive to anyone. Personally, I think the defense of the *on-balance bad* view is decisive, and shows the *on-balance good* view to be wrong. I can imagine that others will disagree. However, their disagreement will consist not in their feeling bound to accept both judgments, but in preferring the *on-balance good* judgment. Disagreement is quite different from paradox.

Suppose someone were to insist that they are equally tempted by both of the inconsistent judgments. Try as they may, they cannot in good conscience reject either. This, too, would show something interesting about paradoxes. What is paradoxical for you might not be so for me. I said I was convinced by the *on-balance bad* view, and that, for me, this simply trumped the arguments for the *on-balance good* judgment. But you may disagree: appealing as you may find the arguments for the one view, you do not find they trump the arguments for the other. The views seem equally well defended. This might be a case, not uncommon in ethics, in which you simply cannot make up your mind. Then it would seem better to say that you confront an ethical difficulty rather than a paradox. A sign would be that you are open to the suggestion that a moral mentor might help you see more clearly, or that some

argument you have not yet considered would resolve the matter. But suppose, by contrast, that you feel confident that all bases have been covered, that there are no further relevant considerations, that no mentoring could help you: then there is a case for saying that you confront a paradox, even if I do not (see question 2.2).

2.2

People often agonize about their moral choices. They feel torn. They can give what would otherwise seem overwhelming reasons in favor of each of two courses of action, while recognizing that they must choose between them. Does every such person confront a moral paradox? Or some but not others? Or is this kind of situation never paradoxical?

Our working definition was that a paradox is an apparently unacceptable conclusion derived by apparently acceptable means from apparently acceptable premises. Applied to Crime Reduction, the apparently unacceptable conclusion is that having a very severe penalty is both on balance good and on balance bad. Paradoxicality is a matter of degree, depending on how cunningly appearance masks reality. Earlier we used the Liar to mark a paradox of the highest degree, and the Barber to mark the lowest. The present discussion suggests, however, that we perhaps ought to relativize degrees of paradoxicality to persons. For, I suggested, what is less paradoxical for me may be more paradoxical for you. You may try to raise the degree of paradoxicality a putative paradox has for me by stressing the excellence of the arguments for the aspect of the paradox I reject. In the present case, you might rehearse the arguments for the *on-balance good* judgment, and challenge me to show where they go wrong. I might try to lower the degree of paradoxicality a putative paradox has for you. In the present case, I might try to persuade you that the considerations adduced in favor of the *on-balance bad* judgment were more weighty than those adduced in favor of the *on-balance good* one. This discussion would involve serious moral issues, and could well be highly constructive.

There is something a little unsettling about this diagnosis of the way in which paradoxicality is relative to persons. The degree of paradoxicality, I said earlier, was the degree to which appearance masks reality. In reality, there cannot be an acceptable argument from acceptable premises for an unacceptable conclusion. There is paradox to the extent that this reality is hidden. It looks as if someone for whom a paradox has a smaller degree of paradoxicality is always wiser than one for whom it has a higher degree, for the former has been less deceived by misleading appearances. Yet this surely cannot be right. I cannot say to my envisaged opponent: the Crime

Reduction scenario is more paradoxical for you than for me, so evidently you have further to go than me in developing your philosophical and moral perceptions.

It cannot be right that someone for whom a putative paradox is paradoxical to a lower degree is automatically wiser than someone for whom it has a higher degree. The explanation is this: a person may mistake where reality lies, and how and to what extent it is hidden. I insist that the arguments for the *on-balance bad* judgment in the Crime Reduction paradox simply trump those for the inconsistent alternative, and so the paradox is unmasked. Merely insisting on this does not guarantee that I am wise, for I might be making a mistake. Perhaps I have wrongly evaluated the strengths of the arguments. Perhaps in reality the arguments are equally strong. If so, the person for whom this paradox is more paradoxical is wiser than I am.

What does it mean to say that one set of considerations "trumps" another? (This is what I said the considerations favoring the *on-balance bad* judgment did to those favoring the *on-balance good* judgment.) At a minimum, it needs to mean that the trumped considerations are less than conclusive. If we take the analogy with trumps in card games seriously, we can go further. Trumped considerations cease to count. It is not merely that they are weighed and found not to weigh enough. They simply cease to enter into what needs to be considered.

To say that the considerations favoring the *on-balance bad* judgment trump those favoring the *on-balance good* judgment is obviously highly controversial. Whether the trumping occurs seems not to be a purely logical matter, but in part also a matter of moral sensibility. This suggests that some moral sensibilities might be more paradox-prone than others. I regard one set of considerations as trumping another, so I do not confront a paradox. You regard neither set of considerations as trumping the other, so you confront a paradox. If this difference is a difference in moral view, as opposed to a different perception of the logical facts, we might wonder whether or not it is rational to avoid paradox-prone moralities. Perhaps it is, for surely paradoxes are bad; perhaps it is not, for perhaps this is how the moral landscape lies. What the present example suggests is that there may be *moral* disagreement about the evaluation of reasoning, in particular about whether one set of reasons does or does not trump another.

2.2 Mixed Blessings

Saul Smilanksy describes cases in which people triumph over misfortune, the very misfortune being the goad to the ultimate triumph. Abigail is born with physical defects. In the course of taking swimming exercises to

overcome these, exercises she performs with quite unusual determination, she turns into a champion swimmer. Abraham is born into a poor family, but his early difficulties make him unusually ambitious, and after many years of hardship he succeeds in building up a chain of businesses, and securing wealth and comfort for himself and his family. In both cases, it was the early misfortune which prompted the special efforts which in turn led to unusual success. The eventual success is essentially a product of the early ill fortune. The paradox might be set out like this:

The misfortune is bad: No one can deny that what Abigail and Abraham suffered as children is misfortune. They suffered conditions which we would, if we could, eliminate, and which we would fervently wish would not afflict our own children. So the misfortune is bad.

The misfortune is good: It was only thanks to the misfortune that Abigail and Abraham made such successes of their lives. It is not just that if it had not been for the misfortune, they would never have done the things that produced their eventual happiness (though that is also true). In addition, the very character of the misfortune (the pain of the physical defects or of the impoverished childhood) is causally responsible for the success. Whatever is causally responsible for success is good. So the misfortune is good.

It might seem that the seemingly opposed views are easy enough to reconcile, using the contrast between absolute and relative judgments of value, as discussed in connection with Crime Reduction. Things may be good in some respects, bad in others; this is not a contradiction, as it would be if something were both absolutely or on balance good and absolutely or on balance bad. We may call a situation a "misfortune" because of its bad respects, while recognizing that it also has good respects. There is no reason why a misfortune cannot be for the best on balance, because being a misfortune only requires some significant bad effects, not a net balance of bad over good. All we need is to relativize the seemingly opposed judgments in the displayed paragraphs: the misfortune is bad in some respects (the ones which lead us to call it a misfortune) and good in others (the ones that prompted Abigail and Abraham to develop their special skills). If there is a paradox, it is one of only a modest degree.

Smilansky considers an alternative case, which is relevant to this approach to dissolving the paradox. Someone breaks his leg, is taken to hospital, falls in love with his doctor, marries her and lives happily ever after. In this case, Smilansky says, even though the chance of the patient's meeting that doctor had he not broken his leg was negligible, the good end result is an accidental product of the bad event. There are "*two* interventions of fortune, one bad (breaking a leg) and one good (meeting the

doctor)" (Smilansky 2007, p. 14). By contrast "Cases such as those of Abigail and Abraham, who have been formed by the misfortune, pose the paradox in a deep way." For the case of the broken leg, perhaps Smilansky would accept the defusing strategy; but not for the case of Abigail and Abraham. So we must examine the difference between the cases.

One idea at work here is that the cases differ significantly in how bad the bad thing is, relative to the good thing. The broken leg is trivial compared to the marital bliss. We would gladly and with a clear conscience have wished him to have broken his leg, if this was the only way to get the couple together. In the other case, things are very different. Even though good fortune triumphs, the misfortune was very serious. Most of us would hesitate before wishing the misfortune on someone, even if we had every reason to believe that it would stimulate them to efforts which would eventually produce success and happiness. When the bad and the good are placed on the scale, we may agree that the good weighs more heavily; but if the bad weighs very heavily, even though outweighed, we find it hard to be enthusiastic about the total package of good and bad. In this way of thinking, we push toward accepting the misfortune as bad, and away from accepting it as good.

It is not clear that this is an entirely rational response. What is really an irrelevant issue may produce some mental fog: when we imagine ourselves "wishing" a childhood like Abigail's or Abraham's on someone, we have to recognize that such a childhood may fail to lead to the happy final upshot. In that case, the childhood would be an unmitigated disaster, and we would feel appalled that we had wished it on someone. But that element of chance is not the issue, for in the case in question we are simply to assume that the misfortunate childhood did its causal work in delivering the ultimate success and happiness. Why should we not rejoice at the final preponderance of pleasure over pain?

Another idea at work in Smilansky's remarks is the distinction between essential and accidental good effects. Meeting the future bride was an accidental effect of breaking the leg. Here we have distinct objects for the predication of *bad* and of *good*. *Bad* attaches to the accident of the broken leg, *good* attaches to the accident of meeting the doctor. There is no single thing to which we are at all disposed to ascribe both properties, and so no glimmer of paradox. By contrast, in the case of Abigail and Abraham, there is a single thing, the early misfortune, which is an essential part of the subsequent good: "the misfortune and the good fortune are non-accidentally part of the same life history" (Smilansky 2007, p. 14).

It is fairly plausible to say that, for Abigail and Abraham, their later good fortune could not have existed had the earlier bad fortune not existed: the earlier states played a non-accidental part in producing the later ones; they

are non-accidental causes. We could use this in a principle designed to help clarify what a misfortune is:

M: A *misfortune* is a non-accidental cause of predominantly bad effects.

It would follow that Abigail's and Abraham's early conditions do not count as misfortunes. On the other hand, they plainly are misfortunes. By contrast, the leg break only accidentally caused the marriage, so this resulting good fortune does not detract from the misfortunate nature of the break. In the light of M, we could set out the paradox as follows:

Misfortune: The childhood conditions of Abigail and Abraham were misfortunes. They suffered conditions which we would, if we could, eliminate, and which we would fervently wish would not afflict our own children.

No misfortune: The childhood conditions of Abigail and Abraham were not misfortunes. This is established by principle M, given that the childhood conditions non-accidentally caused predominantly good effects.

We have a contradiction, and at least some inclination to accept each part of it. So we have a paradox of at least some degree. How great that degree is depends on how plausible we find M (or some similar principle that would deliver the "no misfortune" component of the contradiction). I have to say I do not find it very plausible, and, for me, any plausibility it may otherwise have had is immediately abolished by the paradox. The essence of a misfortune is that it essentially involves bad things happening. There is no requirement that it should not also non-accidentally bring about a significant number of good things.

We should consider M in the more general light of properties that are ascribed through an evaluation of an end state. A victory is getting the better of an enemy in a battle. But as we know from King Pyrrhus of Epirus, a victory can be one's undoing; in retrospect it was a step on the road to disaster. With some distortion of the historical facts, let us say the Romans lost more men at the battle of Asculum than the Greeks commanded by Pyrrhus, and they retreated. This was universally acclaimed a victory for the Greeks. The Romans were easily able to replace their dead and wounded with fresh troops, whereas the Greeks were not. The longer-term upshot was the defeat of the Greeks and the Roman dominance of Magna Graecia. Can the same event, in this case the battle at Asculum, be both a (Pyrrhic) victory and a defeat? Yes, if the epithet "victory" is linked to one part of the total process (the short term) and "defeat" to another (the long term) (see question 2.3). There are many similar examples: the triumphant success of her first novel was her undoing (through raising expectations too high, and causing a breakdown of her confidence). Was the success a disaster?

Suitably understood, the answer is Yes. Victory and success are what they are because of the presence of certain favorable outcomes (some advantage in territory or resources; a mission accomplished). It is not required that, in addition, they lack any significant negative outcomes. If that is how things are with misfortune, then M delivers the wrong verdict: it requires not merely that a misfortune should have bad upshots, but that these should outweigh the good ones.

2.3

A toy gun is not a gun. Might it be that, similarly, a Pyrrhic victory is not a victory?

There is thus a case for saying that M is too demanding, and should be rejected. But without M it is hard to see how to construct a genuine paradox.

2.3 Not Being Sorry

If something bad has occurred, then the only morally appropriate attitude to it (assuming one knows the relevant facts) is some kind of sorrow or regret. Yet there seem to be examples of bad things for which it is permissible to have no sorrow or regret. Here is an example from Smilansky (who is speaking of his own case):

> Before you were born your parents gave birth to a seemingly normal daughter, except that she was born with a severe defect in her heart, which led to her death after only a few weeks … You were born afterwards. In time, you learned that, had your sister survived, your even having been conceived would have been precluded. (Smilansky 2007, p. 59)

Smilansky suggests that though it is appropriate to feel sorry *for* the dead sister, "in the overall sense, you are permitted not to be sorry *that* things have happened as they have" (p. 60, my emphasis). The explanation is that "you are … morally permitted to be happy to have been born, even though you know that, for your birth to have occurred, your sister had to have died" (p. 60).

Here is the paradox: you are morally required to feel sorry that your sister died, says common sense morality; but it is morally permissible not to feel sorry, says Smilansky's argument. Let us look more closely at the argument. It involves some kind of *transfer* (as I shall call it): the close causal connection between your being born and your sister having died makes it morally permissible to transfer your feelings about the former

(you are glad you were born) to your feelings about the latter. This would take one to the conclusion that it is permissible to be glad that your sister died. We would then need a further principle, I call it an *exclusion* principle, to the effect that if it is permissible to be glad that your sister died, it is permissible not to be sorry. The argument might then be expressed as follows, using "*d*" for "your sister died," "*b*" for "you are born," S(*p*) and G(*p*) to express being sorry that *p* and being glad that *p*, and PER for "it is morally permissible that," and "¬" for "not."

(1) PER(G(*b*))
(2) *b* could not have been the case if *d* had not been the case.
(3) If circumstances are related as *d* and *b* in (2), then if PER(G(*b*)), PER (G(*d*)) [Transfer]
(4) PER(G(*d*)) [from 1, 2 and 3]
(5) If PER(G(*d*)) then PER¬ S(*d*) [Exclusion]
(6) PER¬ S(*d*) [from 4 and 5]

The argument is certainly valid, but both principles ((3) – Transfer and (5) – Exclusion) are open to doubt. Perhaps there are circumstances to which one needs to have two responses at once, being both glad and sorry. This may occur if a single circumstance has very different aspects, some calling for sorrow, others for joy. (Many parents respond in this twofold way to an offspring's leaving home to get married.) We will not dwell on this, for we already have something paradoxical at step (4): intuitively it is not permissible to feel glad at the death of an innocent person (especially not if the person is one's sister).

The principle of Transfer is enough to generate this paradoxical result. Is the principle plausible? Not very, in my opinion. The idea behind it is that if it is permissible to be glad about something, it is permissible to be glad about anything that was causally necessary for it. There are common cases in which one deeply regrets having to adopt certain means to some ends. The military speak of civilian deaths inevitably caused by the prosecution of certain forms of warfare as "collateral damage." It is not permissible to be glad that collateral damage occurs, even if it is permissible to be glad that a certain kind of war is waged. This seems to be a clear counterexample to Transfer.

But is it really? Here is a case for the opposed view:

> If the only way of saving the city from the cruel and unjust enemy is through morally justified defensive action that, nevertheless, results in a few unintentional enemy noncombatant casualties, then I regret the necessity of those casualties. But I am not sorry that things happened as they did, for the only alternative was that my city would be overrun by the cruel enemy (etc.). I am glad that things turned out as they did, even though that includes some collateral damage. (Smilansky, personal communication)

We can all agree that the imagined speaker has a right to be glad that the city was saved. Transfer entails that he thereby has a right to be glad about the occurrence of the causally necessary conditions for that upshot, one of which is the death of innocent noncombatants. It is not clear to me that this gladness would be morally acceptable.

There is a paradox about Not Being Sorry only to the extent that Transfer is an apparently acceptable principle. In my view, the appearance of acceptability is slight, so we have a paradox of a fairly modest degree.

The three paradoxes we have discussed so far have a common feature: in every case, there is some kind of causal entanglement. In Crime Reduction, the good end of reducing crime is causally entangled with the bad means of threatening an excessive penalty. In Mixed Blessings, good and bad fortune are entangled. In Not Being Sorry, something to be glad about is causally entangled with something not to be glad about. In each case different threads, with opposite valence, are causally entangled, so that they cannot in practice be separated. As a causal matter, we cannot have the Crime Reduction without the threat of excessive penalty, or the success later in life without the bad childhood circumstances, or the birth of the son without the death of the daughter.

The general idea behind my suggested responses to the three cases is that we can separate in judgment or feeling what cannot be separated in fact. Of causally inseparable threads, we may judge one good and one bad, feel glad about one and not about the other. The proposal, in short, is that we need a more refined object of judgment than simply judgment about "what happened." What happened contained good things and bad; our judgments need to discriminate between these different threads. This can help to dissolve the paradoxes, for instead of conflicting judgments about the same thing, we have judgments about different things; instead of judging that the sister's death merits the attitude of both regret and gladness, we can judge that the death merits unalloyed regret, even though one of its non-accidental effects, the birth of the son, merits unalloyed joy.

There is then often a further question, if the causal threads have opposite valence (one is good, another bad): how can we make an overall judgment, one which relates to the relevant part of reality as a whole: the skein rather than the composing threads? In some cases, the valence of one thread may outweigh or trump the opposed valence of another, and we can reach a stable overall judgment. Thus I suggested that the negative aspects of Crime Reduction trump the positive ones. Even when we reach a stable overall judgment, we should not read it back onto the distinct causal threads. In claiming that, overall, Crime Reduction is bad one does not have to deny that it has causally entangled good aspects; in being pleased by Abigail's and Abraham's eventual success, one does not have to

be pleased at their early misfortunes; in rejoicing in one's own birth, one does not have to rejoice at the causally entangled death of one's sister.

One source of paradox would be the unavailability of a rational and stable overall judgment: influenced now by one thread and now by another, we are tempted to both conflicting overall judgments. This might result from assigning Transfer principles more weight than they deserve: perhaps one is reluctant to abandon one of the overall judgments because one thinks one has to extend it to other entangled threads. One might feel tempted to say that Abigail's childhood cannot count as a misfortune because her life overall is happy and successful, and so what is causally entangled with these good aspects is also good, and cannot constitute misfortune. In response, I would try to weaken the influence of Transfer principles.

If moral reality were somehow inconsistent, so permitting no correct and comprehensive judgments, we would have a radical form of paradox. One aspect of this idea is discussed in the next section.

2.4 Moral dilemmas

Could it be that one morally ought to do something morally bad? Some people think the answer is obviously "No." Indeed, these theorists may say, it is contradictory to suppose that one morally ought to do something morally bad. Others hold that it is not a contradiction but a sad fact of life that one may be morally required to do something morally bad. To the degree that this appears unacceptable, we have a paradox. According to those who think we may be morally required to do something morally bad, the impossibility is mere appearance, and the possibility it obscures needs to be exposed.

Here is a classic example.

> Thirty survivors of a shipwreck are crowded into a lifeboat intended to hold seven. A storm is coming up; the lifeboat has to be lightened if anyone is to survive. The captain reasons that he morally ought to force some individuals to go overboard and drown. He perseveres in this view, even while recognizing that anyone he selects to throw overboard will be an innocent person, and that it is bad to kill an innocent. The captain thinks he is morally obliged to do something morally bad.

In *Sophie's Choice* by William Styron, Sophie is required by a guard in the concentration camp in which she is interned with her two children to select one of them to be killed. If she refuses to choose, both will be killed. By choosing one child for death, Sophie saves the other. The same act is both saving a life and causing a death; it is both morally required and morally bad.

Let us call cases like this moral dilemmas. They are cases in which there is a particular kind of moral conflict: every action available to the agent

involves something morally bad, and at least one of the available options morally ought to be done (see question 2.4). Thus defined, it is controversial whether there really are moral dilemmas. There certainly seem to be: the captain of the lifeboat might do nothing to save the people in his charge, which would be bad because then everyone would die, or he can lighten the boat, which would be good and so ought to be done, by sending innocent people to their death, which would be bad. Sophie might do nothing, and so fail to save a child, which would be bad, or she might save one child, which would be good and so ought to be done, and thereby ensure the death of the other, which would be bad. The question we shall discuss is whether moral dilemmas constitute paradoxes, and in particular whether a moral dilemma involves a contradiction. If it does, we would have to conclude that there cannot be moral dilemmas (assuming that moral reality is consistent). There may seem to be such cases, but closer inspection should reveal that they are not what they seem.

2.4

Quite often a moral dilemma is defined as a situation in which one is obliged to do *A* and to do *B* but one cannot do both. Are there any moral dilemmas in this sense which are not moral dilemmas as defined in the text? Or conversely?

Here is an informal argument for contradiction. We adopt the principle

O: One ought not to do anything morally bad.

In a dilemma, all available options are morally bad and so, by O, one ought not to do any one of them. But, by the definition, there is an option which I ought to take. So this option is one I both ought to take and ought not to take.

This might at first seem like a contradiction, but it is not. It is a case of conflicting obligations. A contradiction would take the form: I ought to do a certain action, and it is not the case that I ought to do it. If we could move from "I ought not to do *A*" to "It is not the case that I ought to do *A*" then we would have a contradiction; but it is far from obvious that this move is correct. It would certainly be incorrect in structurally similar cases. For example, I cannot move from "It's possible that it won't rain today" to "It's not possible that it will rain today." It's also incorrect in the case of desires and preferences, which one might suppose to have a similar logic to obligation. It is commonplace to have conflicting preferences: I would prefer not to go to the dentist (it takes time and is disagreeable), but I would prefer to go (to avoid worse problems from dental neglect).

There could not be anything paradoxical or contradictory about such a set of preferences. The rational person checks to see which one matters most, and acts accordingly. But if we could move from "I'd prefer not to go" to "It's not the case that I'd prefer to go" this commonplace situation would be contradictory, and so impossible.

A little formalism will make it easier to follow the reasoning. Most people agree that in a moral dilemma there is an action, A, which one both ought and ought not to do:

$O(A) \& O(\neg A)$.

("O" abbreviates "ought," "&" abbreviates "and," "¬" abbreviates "not.") To have a contradiction we would need something different:

$O(A) \& \neg O(A)$.

So far, the only way we have contemplated reaching a contradiction is by the principle:

C: if $O(\neg A)$ then $\neg O(A)$ (see question 2.5).

2.5

This is equivalent to "if $O(A)$ then $\neg O(\neg A)$" which might be glossed "the same action cannot be obligatory and forbidden" (McConnell 2006). Does this show that the principle is one that cannot be resisted?

The examples of supposed dilemmas make this principle one which we should not accept without argument. That the captain is obliged not to throw any passengers overboard does not obviously entail that he is not obliged to throw any overboard: his duty to save as many people as possible seems precisely to oblige him to throw some overboard. However, there are other ways in which a contradiction might be reached.

The key feature of a dilemma is that $O(A) \& O(\neg A)$. It is natural to think that obligations "agglomerate": if one is obliged to do each of two things, A and B, then one is obliged to do them both. More formally:

AGG: if $O(A)$ and $O(B)$ then $O(A \& B)$.

Applying this to a dilemma yields $O(A \& \neg A)$. It has also often been held that one can be obliged to do only what one can do, that "ought implies can":

CAN: if $O(A)$ then $C(A)$.

("C" abbreviates "can.") Applying (CAN) to $O(A \& \neg A)$ yields:

$C(A \& \neg A)$.

But this is obviously false: one cannot do both of two inconsistent actions (we are assuming that the context is held constant). Indeed, its negation is an obvious truth:

$\neg C(A \& \neg A)$.

Now we have a contradiction, and thus a paradox: dilemmas seem possible, but they are contradictory, and so impossible.

The principles which lead to this result are open to challenge. Indeed, they are in tension with one another, if the reality of dilemmas is granted. Given CAN, a dilemma situation is precisely one in which obligations do not agglomerate. And given AGG, a dilemma situation is precisely one in which one cannot do everything one ought. But are there any reasons independent of moral dilemmas for denying CAN or AGG?

Against CAN, it could be suggested that it is possible to be under an obligation one cannot fulfill. For example, a drunken driver ought to negotiate the curve accurately, even if his condition makes it impossible for him to accomplish this. An addict ought to stop taking the substance, even if he cannot give it up. The attraction of "ought implies can" (the argument might continue) is that in many cases in which the ability is lacking, we excuse failure to act in accord with the obligation. But this does not mean that the obligation is absent or somehow nullified. Maybe the agent is excused; but this would make no sense if there were no obligation for him to be excused from.

Against AGG, it might be argued that there is a maximum total degree of obligation that can be imposed on a single agent. If actions *A* and *B* are each very demanding, it seems coherent to hold that each is obligatory whereas their conjunction is not. For example, perhaps I ought to give large amounts of money to charity. Perhaps I also ought to give large amounts of money to provide de luxe care for my aging mother. But I cannot do both.

Moral philosophers debate whether moral dilemmas are genuinely possible. My own view is that they are. They seem to me merely a reflection of the fact that sometimes we have to choose the lesser evil. The lesser evil is still an evil, even though chosen after the most careful and high-minded moral deliberation. The best thing to do may still be a bad thing.

In some cases of competing obligations, one obligation may abolish or annul another. The Dutch people who protected Anne Frank and her family from the Gestapo had to lie to do so. Perhaps they felt conflicting obligations: they ought not to lie, but they also ought not to betray the Jews in the attic, who would be sent to the death camps if their presence became

known to the Gestapo. In this case, it seems that the obligation to protect the Jews simply annuls any general obligation to tell the truth. Indeed, the case could reasonably be taken as showing that there is no *universal* obligation to tell the truth: you ought not to tell it in Anne Frank-type situations. It was not a dilemma, because the obligation to save lives simply annulled any obligation to tell the truth. This contrasts sharply with the situation of the boat's captain. We cannot say that the obligation not to kill the innocent was somehow annulled in his situation. All we can say, in my opinion, is that the obligation was one that had to be ignored, if the captain was to act for the best. To do the best thing, he had to do something bad. There would be something wrong with the captain if he did not think that what he had to do was simply terrible.

One source of resistance to the view that there are genuine dilemmas is the belief that a person of good conscience can never be in a position in which she has to do something bad. If she morally *has* to do something, it cannot (in the circumstances) be bad. Otherwise people could fall from moral perfection for reasons entirely beyond their control; and this could not be so. The underlying thought is, I think, a kind of wishful thinking about the moral life: the belief that, given all one's faculties, being morally perfect is a goal that no outside circumstances could prevent one from attaining. Sadly, good and evil sometimes come inextricably entwined, and unless we are lucky we may find ourselves having to do something which is bad (bad even in those very circumstances) in order to act for the best.

If this is the right view, there are important practical consequences. Moralists are often happy to end their discussion with a conclusion of the form "Actions of type A are morally bad, and so ought not to be performed." A conspicuous example is the debate over abortion. Suppose a pregnancy poses a genuine moral dilemma. Even the person who believes that abortion is morally bad can consistently believe that it is morally the best thing to do under the circumstances. The conclusion that it is bad can be accepted by all parties to the debate about how it is morally best to act. Mere belief in its badness is not enough to guarantee that those acting from the most carefully considered moral scruples will not properly regard an abortion as the morally best course in certain circumstances. If the pregnancy is a dilemma, it may be that something morally bad has to be done for the morally best thing to be done.

Moral dilemmas (as defined here) are cases in which obligations conflict: there is some action which both ought to be done and ought not to be done. If this entailed that there is some action of which it is true both that it ought to be done and that it is not the case that it ought to be done, moral dilemmas would be contradictory and so impossible. Our paradox would

then be that they *seem* to be possible, while being impossible. My suggestion is that moral dilemmas are not really impossible, and so the paradoxical contradiction is avoided.

Suggested reading

The first three paradoxes are drawn from Smilansky (2007), who discusses them in more detail. His work, and the references he gives, should be the first resource for further exploration. He suggests that morality may give rise to "existential paradox," revealing that "a segment of moral reality … is absurd" (2007: 4–5). Paradoxes of this kind, if there are any, are not to be "dissolved" or disposed of. We might have to accept that "the very idea of a wholly consistent and coherent moral view is impossible" (2007: 127).

Derek Parfit (1984) proposed a number of moral paradoxes, particularly about population and bringing people into existence. For paradoxes pertaining to nuclear deterrence see Kavka (1987).

The topic of moral dilemmas has generated a very large literature, and the remarks here skim over many interesting issues. A good place to begin is McConnell's *Stanford Encyclopedia* article (2006), which also has a good bibliography: plato.stanford.edu/entries/moral-dilemmas/. Gowans (1987) provides a useful collection of articles. Among the classic articles, I would single out Williams (1966) and Marcus (1980). One among many examples of moralists who end their discussion by concluding that some kind of action is wrong, apparently thinking they have thereby resolved the question of what is to be done, is Gensler (1986), arguing against abortion.

3 Vagueness: the paradox of the heap

3.1 Sorites paradoxes: preliminaries

Suppose two people differ in height by one-tenth of an inch (0.1″). We are inclined to believe that either both or neither are tall. If one is 6′ 6″ and the other is 0.1″ shorter than this, then both are tall. If one is 4′ 6″ and the other is 0.1″ taller, then neither is tall. This apparently obvious and uncontroversial supposition appears to lead to the paradoxical conclusion that everyone is tall. Consider a series of heights starting with 6′ 6″ and descending by steps of 0.1″. A person of 6′ 6″ is tall. By our supposition, so must be a person of 6′ 5.9″. However, if a person of this height is tall, so must a person one-tenth of an inch smaller; and so on, without limit, until we find ourselves forced to say, absurdly, that a 4′ 6″ person is tall (see question 3.1), indeed, that everyone is tall (see question 3.2).

3.1

How should one respond to the objection that someone 4′ 6″ tall may be tall for a pygmy?

3.2

How might one use similar reasoning on behalf of the conclusion that no one is tall?

In ancient times, a similar paradox was told in terms of a heap, and a Greek word for "heap" – *soros* – has given rise to the use of the word "sorites" for all paradoxes of this kind. Suppose you have a heap of sand. If you take away one grain of sand, what remains is still a heap: removing a single grain cannot turn a heap into something that is not a heap. If two collections of grains of sand differ in number by just one grain, then both or neither are heaps. This apparently obvious and uncontroversial

supposition appears to lead to the paradoxical conclusion that all collections of grains of sand, even one-membered collections, are heaps.

Suppose you are looking at a spectrum of colors painted on a wall through a device with a split window which divides the section of the spectrum you can see into two equal adjacent areas. Suppose the spectrum is so broad and the windows so narrow that the colors in the two visible windows are indistinguishable, no matter where in the spectrum the device is positioned. The device is first placed at the red end of the wall, and then moved gradually rightward to the blue end. It is moved in such a way that the area that was visible in the right-hand window in the previous position is now visible in the left. At the beginning you will unhesitatingly judge that both areas are red. At each point, the newly visible area will appear indistinguishable from an area that you have already judged to be red and that is still visible. One feels bound by the principle that if two colored patches are indistinguishable in color, then both or neither are red; yet clearly there must come a time when neither of the visible areas is red. This looks like a contradiction: on the one hand, no two adjacent areas differ in color and the first is certainly red; on the other hand, the first area differs in color from some subsequent area (see question 3.3).

3.3

How would one construct a parallel argument with the paradoxical conclusion that no men are bald?

What do these paradoxical arguments have in common? In each case, the key word is *vague*: "tall," "heap," "red." A vague word admits of borderline cases – cases in which we do not know whether to apply the word or not, even though we have all the kinds of information that we would normally regard as sufficient to settle the matter. We may see how tall a man is, or even know his height to a millimeter, yet we may be unable to decide whether he counts as tall or not. We may see a collection of grains of sand, and even know exactly how many grains the collection contains, yet not know whether it should be called a heap or not. We may see a color under the most favorable conditions imaginable, yet not know whether it should be called red or orange. Our ignorance does not manifest failure to understand our language, and we may not be ignorant of any of the matters that one would expect to resolve the problematic issue: how tall someone is, how big a collection of grains is, how something looks.

Vagueness gives rise to *borderline cases*, ones in which we do not know what to say, despite having all the information that would normally fix the correct verdict. But what is vagueness itself? Is it a property of language?

Or of the world? Or of ourselves? On these questions there have been many opinions, illustrated in these claims:

1. Vagueness is *absence of fact*. When it is vague whether someone is tall, *there is no fact of the matter* whether or not he is tall. The reason we do not know what to say in borderline cases is that *there is nothing to know*.
2. Vagueness is *absence of definite truth*: a person is borderline for "tall" just on condition it is neither definitely true nor definitely not true that she is tall.
3. Vagueness is *absence of a sharp boundary*. E.g. in a series of men of closely similar but steadily diminishing height there is no last (definitely) tall man, and no first (definitely) non-tall man.
4. Vagueness is *incompleteness of meaning*. A vague expression is a bit like a partial function in math. Words whose meaning is fully specific and complete are not vague.
5. Vagueness is *indecision*: "The reason it's vague where the outback begins is not that there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word 'outback.' Vagueness is semantic indecision" (Lewis 1986, p. 212).
6. Vagueness is *a feature of the world*: some things, like mountains, are vague, because it is vague what their spatial extent is; others, like properties, are vague because it is vague what things they apply to.
7. Vagueness is *ignorance*: there are sharp boundaries (facts of the matter, definite truth or falsehood, etc.), but we cannot know where they fall.

These claims are probably not all independent (see question 3.4), and some pairs are inconsistent. For example, if vagueness is a feature of the meaning of expressions, it seems unlikely it could be a feature of the world (see questions 3.5, 3.6).

3.4

Choose a pair of views which seem close, and see if you can argue that if one is correct, so is the other, and vice versa. If this fails, perhaps you can suggest a scenario on which we have vagueness on one but not the other of the views.

3.5

Provide an example of an inconsistent pair of views, other than the pair (4) and (6).

3.6

How could one who holds that vagueness is not a feature of reality (but only of our descriptions of reality) respond to the following argument?

> Mountains are part of reality, but they are vague. They have no sharp boundaries: it is vague where the mountain ends and the plain begins. So it is easy to see that vagueness is a feature of reality, and not just of our thought and talk.

On any reasonable view, vagueness needs to be distinguished from relativity and from ambiguity. Consider the property of *being above average in height*. Assuming that there is no problem about assigning numbers to people as measures of their height, this is not a vague property. A person is above average in height just on condition that the number that measures his or her height is greater than the number that measures the average height, and this is a completely precise condition. However, being greater than the average height, though precise, is *relative* to a given population. Being above average in height for a Swede involves being taller than being above average in height for an Eskimo, since the average height of Swedes is greater than the average height of Eskimos.

As the example shows, relativity does not entail vagueness, since it can hold of precise expressions, such as "is above average in height." Many vague predicates are also relative, but their relativity must not be confused with their vagueness. For example, "tall," unlike "above average in height," is vague and also relative. You can see that the vagueness does not entail the relativity by seeing that you could eliminate the relativity but leave the vagueness. If instead of "tall" we were to say "tall for a Swede," we would have eliminated some relativity, but the vagueness would remain (see question 3.7). Most people think that the difference of 0.1″ cannot make the difference between being tall for a Swede and not. This means that the argument of the opening paragraph will work as well for "tall for a Swede" as it did for "tall" (see question 3.8).

3.7

Does the restriction to Swedes eliminate all relativity?

3.8

Is "heap" relative in the way that "tall" is?

Vagueness must also be distinguished from *ambiguity*. Consider the word "bank." It can mean the edge of a river or a financial institution. As a result, we do not know how to answer the question "Did he go to the bank this morning?" So far, there is a parallel with vagueness. We do not know how to answer the question "Is he bald?" if the person in question is a borderline case. However, there is a difference. In the case of ambiguity, a single sentence can be used to say, or ask, more than one thing. Before communication can proceed, the audience needs to determine *which* thing is being said or asked. In the bank example, there is no one answer because there is no one question. With vagueness it is different. If someone asks of a borderline case "Is he a child?," it is not that our problem in answering is the problem of knowing *which* question has been asked; there is only one possible question involved here. The problem is quite different: if the person of whom the question is asked is a borderline case, neither Yes nor No is a clearly correct answer. This question has a single vague meaning, and that is quite different from having two or more meanings.

Vagueness is a widespread feature of our thought. Consider the following list: "child," "book," "toy," "happy," "clever," "few," "cloudy," "pearl," "moustache," "game," "husband," "table" (see question 3.9). Is this feature ineradicable? Could we replace our vague expressions by precise ones? We will answer this question differently, indeed understand it differently, depending on what we take vagueness to be.

3.9

Show that each of the words in the list is vague by briefly sketching a borderline case. Can you think of any words that are not vague?

Arguably, it is important to the role of the concept of childhood in our lives that the word "child" be vague. For example, we take ourselves to have special duties to children that we do not have to adults, but these duties are not relinquished overnight; they gradually fade away, just as childhood itself does. There is no sharp boundary to the end of our duties, just as there is no sharp boundary to the end of childhood. If we were to replace "child" by some more precise term, say "minor," as we have to for certain legal purposes, it would no longer be possible to express the obligations we feel. Our duties to people as

children may end before or persist after their legal majority, depending on what age the law selects for coming of age, and depending on whether they are, as we say, "old for their age" or "young for their age."

Consider a second example of the same general kind of argument. It is important to the role of the concept of redness in our lives that the word "red" be vague. In particular, the vagueness is essential to the observational character of "red," to the fact that we can under favorable conditions tell just by looking whether something is red or not. If we replaced "red" by an expression having sharp boundaries, say one defined in terms of wavelengths, then the right way to approach applying it would be by an exact determination of the wavelength: in losing the vagueness, we lose the observational character.

These arguments are unsatisfactory twice over. First, they seem simply to presuppose a particular view of what vagueness is, without supplying a justification. In the argument about "child," it was explicit that vagueness was to be regarded as absence of sharp boundaries. The argument about "red" at least excluded some views of vagueness, e.g. (7) (vagueness is ignorance). Second, the arguments are fallacious. The uncontroversial premises in the case of "red" are these:

Under some conditions, and for some objects, we can confidently apply or deny "red" simply on the basis of observation.

For some objects, we cannot confidently either apply or deny "red."

These properties could hold of an expression which draws sharp boundaries, for example "is more than 6′ tall." Under some conditions, and for some objects, we could be confident, on the basis of observation alone, whether to apply or deny the predicate. This is consistent with the fact that, at least unless we have a ruler, there will be objects about which we are in doubt: those close in height to 6′. The argument erred in supposing that if a color word were defined in terms of exact wavelengths, we would have to use an exact wavelength-measurer to apply it. The argument failed to show that an absence of sharp boundaries was required for, or helped explain, the observational character of some vague expressions (see question 3.10).

3.10

Assess whether the following consideration might help reinstate an interesting connection between vagueness and observationality:

> In the case of "is more than 6′ tall" one could in principle decide all cases, if one had the help of a ruler. By contrast, in the case of "red," nothing could help one decide borderline cases.

The argument about childhood is also defective. To say that we have special duties of care to children is not in itself to say that how demanding those duties are is inversely proportional to how old a person is. If we do want to say the latter, then we seem to be able to say it straightforwardly, as I have just done, and in a way which does not require "child" to fail to have sharp boundaries. One can also accept that childhood fades gradually away without accepting that "child" draws no sharp boundaries. To show this by analogy: if one is under twenty-one, the time remaining until the moment of one's twenty-first birthday gradually diminishes, but not in a way that requires lack of sharp boundaries.

Whatever the explanation, there can be no doubt that vagueness is a very widespread phenomenon. Hence there are plenty of opportunities for constructing paradoxical sorites arguments. We must look at these more closely, and see what responses are possible.

3.2 Sorites paradoxes: some options

In this section, I will make more explicit one form of soritical reasoning, and will lay out some possible responses.

Sorites reasoning depends upon the supposition that vague expressions are "tolerant": small changes do not affect the applicability of the word. If someone is tall, so is a person a millimeter shorter; if a collection is a heap, so is one otherwise similar but with just one grain less. The paradox arises because big changes, ones which obviously do affect whether or not the word applies, can be constructed out of small changes.

Here is one explicit version of a sorites argument, showing the impact of the supposed tolerance of vague words. First premise:

(1) A 10,000-grained collection is a heap.

Second premise, manifesting the supposed tolerance of "heap":

(2) If a 10,000-grained collection is a heap, then so is a 9,999-grained collection.

Tolerance underwrites numerous further premises of this form, for example:

(3) If a 9,999-grained collection is a heap, then so is a 9,998-grained collection.

And so on. Let us call the first premise the *categorical* premise and the others the *conditional* premises. (A conditional statement is one of the form "If …, then …")

We are just as inclined to hold to these conditional premises when they concern small numbers as when they concern large numbers, and just as

inclined to hold to them for cases in which there is genuine doubt about whether a collection is heap-sized as when there is no such doubt. For example:

(10,000) If a 2-grained collection is a heap, so is a 1-grained collection.

We are firmly convinced that neither a 1-grained nor a 2-grained collection is a heap; but this does not stop us holding that *if* a 2-grained collection is a heap, then so is a 1-grained collection. This reflects our general conviction that taking away a grain cannot turn something from a heap into a non-heap. This also attaches to borderline cases, for example:

(9,925) If a 77-grained collection is a heap, so is a 76-grained collection.

We may be in genuine doubt about whether either collection deserves to be called a heap. Yet the conditional reflects our confidence that both or neither are heaps, and this is simply another way of putting the tolerance principle: a difference of a grain cannot be the difference between a heap and a non-heap.

We do not yet have a paradox. To get it, we apply to these premises a general principle of reasoning: given *p*, and a conditional of the form "if *p*, then *q*," we can derive *q*. This principle is still called by the Latin name it was given in the middle ages: modus ponendo ponens, or modus ponens, for short. Applying it to our first and second premises yields:

A 9,999-grained collection is a heap.

Applying the principle again to the above and premise (3) yields:

A 9,998-grained collection is a heap.

Continuing in the same way, we finally end up with the result that a 1-grained (or, for that matter, even a 0-grained) collection is a heap, and this is the absurdity.

As with any paradox there are three possible responses to consider:

(a) Accept the conclusion of the argument.
(b) Reject the reasoning as faulty.
(c) Reject one or more premises.

For some vague words, (a) may seem a totally unpromising possibility to explore. Could anything reconcile us to the suggestion that everyone is tall, or that a heap cannot be demolished grain by grain so that no heap is left, or that all colors are red? Some philosophers have used sorites reasoning to argue for conclusions of just this kind. I give one example in the next section.

3.3 Accepting the conclusion: Unger's view

Consider the following argument:

(1) Given just one gram of wood, you cannot make a table.
(2) If you cannot make a table out of n grams of wood, then you cannot make a table out of $n + 1$ grams.
(3) Hence you cannot make a table, no matter how many grams of wood you have.

Crucial to the argument is the idea that a gram cannot make the difference between not enough wood to make a table and enough wood. (If you feel doubts about this, try reducing the amount to a millionth of a gram.)

There is a close analogy between this reasoning and the reasoning of the previous section concerning heaps. That reasoning seemed to establish that heaps are indestructible: however many grains you take away from a heap, it remains a heap. The present reasoning runs in the reverse direction, and seems to establish that tables are uncreatable: however many bits you add, you still do not have a table. This conclusion amounts to the claim that there are no tables.

Nothing essentially depends upon the notion of literally "making" a table. We could use similar reasoning to show that there are no stones. If a region contains only one gram of solid material, it does not contain a stone. If a region contains n grams of solid material but not a stone, then a region as similar as possible except that it contains $n + 1$ grams of solid material also does not contain a stone. If we start with a region which we take, intuitively, to be occupied by a stone, and focus on a one gram part of it, and then successively expand our considerations outward by including, one at a time, a further adjacent region occupied by a gram of matter, we appear to be able to conclude, contrary to the original intuition, that however far we expand we will not get to a region containing a stone. Generalizing, the conclusion amounts to the claim that there are no stones.

These conclusions seem strange, but they have been advanced with at least apparent seriousness by Peter Unger (1979a). We can make them seem less mad by putting them in a certain perspective. Perhaps the conclusion we should draw from the existence of the sorites paradoxes is that vague concepts are deeply flawed: they commit us to absurdities. A flawed concept is one under which nothing can fall. So even though the world may contain all sorts of stuff, it is wrong to say that we can divide it up using concepts flawed by vagueness. Rather, we have to say that nothing matches these concepts: there are no tables, no stones, etc. (see question 3.11).

3.11

*How should someone who adopts Unger's viewpoint respond to the argument which concludes that even a 1-grained collection would make a heap?

One cannot dismiss this point of view as mad. However, it would not be wrong to regard it as something of a last resort, and Unger himself accepts that if there were some way of defusing sorites paradoxes, his own solution would be unattractive. We need to look at alternative responses.

Sorites reasoning appears to be extremely simple and to use only the fundamental logical principle of modus ponens, so the prospect of rejecting it as invalid (response (b) above) is not initially tempting.

The most obvious response to explore is (c): reject one or more premises. The next two sections explore two ways in which this might be motivated.

3.4 Rejecting the premises: the epistemic theory

The epistemic theory of vagueness (7 on p. 42) holds that vagueness is nothing but ignorance. So far as their semantics go, vague words are just like precise ones: they draw sharp boundaries. The epistemic theorist will see sorites paradoxical arguments as proving his view. When the arguments take the form displayed in the previous section, the epistemic theorist sees them as establishing by *reductio ad absurdum* the falsity of at least one premise. (An argument by *reductio ad absurdum* is one which starts by deducing an absurdity from some premises, in order to demonstrate that one of the premises must be false.) The conclusion (e.g. that there are no heaps) is plainly false, the reasoning is valid, so one of the premises must be false. The first premise (e.g. the one that says that a collection of 10,000 grains can make a heap) cannot be disputed. So it is the generalization exemplified by the second premise that must be false, the "principle of tolerance":

If an n-grained collection can make a heap, then so can an $n-1$-grained collection.

On the epistemic view, there are heaps and also non-heaps, and these are separated by a sharp line. In other words, for some number n, a collection of grains with n members can make a heap but a collection of grains with $n-1$ members cannot. The special twist the theory gives, from which it earns its name and also any plausibility it may have, is that we cannot, even

in principle, know where this sharp boundary lies. That is why borderline cases are confounding: they ask for a verdict in a situation in which we cannot know what the right verdict is.

The epistemic theory is generally greeted with incredulity. How could there be a sharp boundary to the heaps, the bald men, the tall men, childhood, and so on? We can agree with epistemic theorists that, if there are such boundaries, we cannot even in principle know where they fall; but in contrast to epistemicists, we might take this as a reason for thinking that there are no such boundaries. If they really exist, what could prevent us telling where they lie, given enough time and effort?

Epistemicists have two things to say in response. First, they will insist that they do not accept the *verificationist theory of meaning*, or *verificationism* for short. This theory, characteristic of the logical positivism which flourished in the middle years of the twentieth century, has it that any sentence we can understand is one which, in principle, we could tell to be true, or false, as the case may be. Verificationism is inconsistent with the epistemic theory, which holds that if an object is a borderline case for a vague word, we have no means of determining the truth or falsehood of an application of the word to the object. However, at least in the crude form presented here, verificationism is now not widely accepted, so few would wish to attack the epistemic theory from this angle.

Epistemicists have a second response: an explanation of why we are incurably ignorant in borderline cases. On the epistemic view, it is because our cognitive mechanisms, for example our senses, require a margin for error if they are to deliver knowledge. This is so in ordinary cases, ones not involving vagueness. For example, even if I believe, glancing round the stadium, that 3,973 people are present, when in fact this is the actual number, my true belief cannot count as knowledge, for there being one person more or less would not have affected my estimate. Even though I am right, this shows that I am right by accident, and so do not know. The ignorance that, on this view, is distinctive of vagueness is of a special kind: it relates to the precise meanings of the concepts I use (at least according to some epistemicists, e.g. Williamson). I actually use the concept *red*. Invoking this concept, let us suppose that a certain patch p is the last red patch on the wall we described earlier, where the red patches become less and less red, and are followed by orange patches. Suppose, moreover, that I *believe* that p is the last red patch in the series. The epistemic theorist argues that this belief, though true, does not count as knowledge, for there is no margin for error. I could easily have come to possess, in place of my actual concept *red*, a slightly different concept, say *red**, which drew the boundary in a slightly different place, so that the last red* patch would have been to the right of p. Just as there being

one person more or less in the stadium would have had no impact on what belief I formed, so my having this different concept would not have stopped me believing, falsely, that *p* is red*. Knowledge requires being non-accidentally right, and in cases like these, this in turn requires a margin for error. No margin for error, no knowledge (see question 3.12).

3.12

The version of epistemicism described here turns on a lack of margin for error concerning the concepts we use. Why should not an epistemic theorist instead rely on the margin for error needed by our perceptual systems? (We cannot distinguish the color of *p* from that of adjacent patches, so it was just luck we selected *p* as the last red.) First hint: compare the stadium case. Second hint: check out Williamson (1994: section 8.4).

Epistemic theorists may thus explain our ignorance, but they still have to establish the claim that there is anything to be ignorant of: that vague predicates draw sharp boundaries. Taking up this challenge in full would involve showing that alternative responses to sorites paradoxes are inadequate (see question 3.13).

3.13

Evaluate the following argument against the epistemic theory, as applied to color predicates:

> If a predicate stands for a manifest property (one which under some conditions detectably obtains), then under optimal conditions for manifestation, if there is a fact of the matter whether or not it obtains, that fact is detectable. We can view a region of the colored wall (one borderline for "red") under optimal conditions without being able to detect the presence or absence of redness; so there is no fact concerning whether that region is or is not red.

3.5 Rejecting the premises: supervaluations

One can reject one of the conditional premises without adopting the epistemic theory. One way is the *theory of supervaluations*. The underlying thought is that vague words are incomplete or defective in their meaning (see diagnoses 4 and 5 on p. 42), and this is manifest in borderline cases. For these cases, the meaning of the expression is inadequate to determine whether or not it is true of the object. One can specify the meanings of

such incomplete expressions by specifying the various ways in which the incompleteness could have been remedied, that is, the various ways in which the expression could have been made precise.

Many vague words have a meaning which depends upon some underlying ordering, one which induces principles like this: if anything is a heap, then anything otherwise similar except bigger is a heap too; and if anything is a non-heap, anything otherwise similar but smaller is also a non-heap. If a vague word like "heap" had a complete meaning, then for every candidate for being a heap, the word would either definitely apply or definitely not apply. The upshot would be a sharp cut-off: a smallest heap (a least point in the underlying ordering). On supervaluationist views, vagueness is lack of a complete meaning. The incompleteness shows up in borderline cases, to which the word neither definitely applies nor definitely fails to apply. We can envisage various ways in which the meaning of the word could be completed, various ways (consistent with the underlying ordering) that it could have been made precise. By indicating which these are, we can reveal a vague word's incomplete meaning, rather as we could reveal what Jack's unfinished house is like by showing all the various ways in which he could complete it. What has actually been constructed is what all these possible completions have in common. Likewise, what there is of the meaning of a vague word is what all the various ways of completing it agree on.

For supervaluationists, the crucial tool is the notion of a sharpening (sometimes called precisification). A sharpening, *s(w)*, of a vague predicate, *w* (a word like "heap" which can be true or false of objects), meets the following conditions:

- If *w* is definitely true of something, then *s(w)* is true of it.
- If *w* is definitely false of something, then *s(w)* is false of it.
- For each object, *s(w)* is either true of it or false of it.
- *s(w)* respects the underlying ordering (if there is one). For example, if *s*("tall") is true of someone 6′ tall, it is also true of someone 6′ 1″.

The meaning of a vague predicate is then to be described in terms of *all* the sharpenings which meet these conditions (analogously to *all* the ways Jack could complete his house). Where the sharpenings disagree, we have vagueness; where they agree, we do not (just as where the house models agree, we have finished construction, and where they disagree we do not).

Truth, for a supervaluationist, is identified with *truth on all sharpenings* and falsehood with *falsehood on all sharpenings*. What matters is what is really present, the incomplete meaning (likewise, on the analogy, the incomplete construction). A sentence which ascribes a vague predicate to a borderline case will be neither true nor false. This is the critical result which the supervaluationist uses to reject a premise of the sorites reasoning. The

first (categorical) premise, for example that with 10,000 grains one can make a heap, cannot be rejected. But many of the conditionals fail of truth, within supervaluationist frameworks. Let us see exactly how.

Suppose α and β are borderline for "heap," and that α has one more grain than β. (Perhaps α has 76 grains and β has 75.) On the basis of what has been said so far, "α is a heap" is neither true nor false: since α is a borderline case for "bald," *s*("bald") will be true of α for some but not all sharpenings *s*; so the sentence will be true on some sharpenings (and so not false) and false on some sharpenings (and so not true). Likewise for "β is a heap." Consider the conditional

If α is a heap, then β is a heap.

For this to be true, it needs to be true on every sharpening. Given the general conditions on sharpenings, there is a sharpening of "heap" on which "α is a heap" is true and "β is a heap" is false. This is a sharpening which makes a 76-grained collection the smallest heap. The conditional is false on this sharpening (a conditional with a true antecedent and false consequent is false). Hence the conditional is not true on all sharpenings and so is not true. (Neither is it false, as analogous reasoning would show (see question 3.14).) At least one such conditional is sure to feature somewhere in a sorites argument constructed in the style envisaged in section 3.2. Hence, according to supervaluation theory, not all the premises of this kind of sorites argument are true. This is enough to offer a resolution to the paradox. The theorist need not specify a particular non-true premise.

3.14

How could it be shown that the conditional is not false?

It is not that, on this account, *any* sentence in which a vague word is applied to an object in its penumbra is neither true nor false. For example, the sentence

α is either a heap or not a heap

is true on all sharpenings, and so true, even if α is a borderline case. Every sharpening draws the line somewhere. Wherever it draws it, "heap" will either be true of α on that sharpening, or false of α on that sharpening. Hence, for any sharpening, either "α is a heap" or "α is not a heap" is true on that sharpening. Therefore " α is a heap or α is not a heap" is true on every sharpening – that is, on the account, *true* (see question 3.15).

3.15

Does the supervaluational account entail the following: either "α is a heap" is true or "α is not a heap" is true?

What interesting point about the supervaluational treatment of "or" does your answer highlight?

To sum up, supervaluational accounts purport to dissolve the paradox by showing that not all the premises of the paradoxical argument are true. In particular, the principle of tolerance does not hold: we cannot infer that if someone is tall, so is someone just a millimeter shorter. One merit of the account is that it preserves standard logic. For example, it brings all instances of "either *p* or not *p*" out true. On this account, we need have no truck with the response I labeled (b): giving up some logical principles.

I shall consider four problems for supervaluational accounts. The first is that preserving classical logic may not be such a good thing. Vagueness has often been supposed to throw doubt on the logical principle called the Law of Excluded Middle, exemplified by:

Either he is an adult or he is not.

It might be suggested that given that there are borderline cases of adults, we should be reluctant to affirm this instance of the law. Affirming it might be exploited in an argument which could reasonably be held to be flawed:

Either he is an adult or he is not. If he is an adult, then watching the hard-porn movie will do him no harm. If he is not an adult, then he will not understand it, so, in this case too, watching it will do him no harm. Either way, it will do him no harm to watch it.

We might object: the argument fails to take account of the person on the borderline between childhood and adulthood. For him, it is not right to say "Either he is an adult or he is not." It is precisely because he is between childhood and adulthood that seeing the movie might harm him.

The second problem for supervaluational theories is that they assign intuitively the wrong truth value to a sentence which is central to sorites paradoxes:

For some number n, an n-grained collection is a heap but an $(n - 1)$-grained collection is not.

We intuitively tend to believe that this sentence is false (and likewise similar sentences for other vague words), for it seems to affirm the existence of a sharp boundary, and so fails to do justice to the vagueness of

"heap" (see question 3.16). On the supervaluational theory the sentence comes out as true, since it is true on every sharpening (see question 3.17).

3.16

Would the epistemic theorist agree with this judgment?

3.17

How is this demonstrated? (You might want to check out Sanford 1976.)

The third problem for supervaluational theories is that the underlying conception of vagueness as nothing but incompleteness of meaning is at a minimum insufficient. Predicates which, intuitively, are not vague may be incomplete in meaning. For example, we might define a minor by the following clauses:

(1) People who have not reached their seventeenth birthday are minors.
(2) People who have reached their eighteenth birthday are not minors.

This definition intuitively does not make "minor" vague. But its meaning is incomplete: it fails to speak to persons who are seventeen years old.

The clauses which define "minor" are entirely precise, and that is why we are right not to think of "minor" as vague. The problem relating to seventeen-year-olds is not vagueness or imprecision but incompleteness. We find the same phenomenon among familiar words. Perhaps "pearl" is true of anything made of a certain material and formed in an oyster and false of anything not of that material. This leaves undetermined the word's application to pearl-sized lumps of pearl-material synthesized outside of any oyster. I suggest we should see "pearl" not as vague, but as incomplete: there is a blank where there should be a rule. If this is right, supervaluation theory's guiding idea, that vagueness is a kind of incompleteness, is not enough to characterize vagueness. If vagueness is incompleteness of meaning, it must be a special kind of incompleteness.

The fourth alleged problem for the supervaluational account of vagueness presented here is that it fails to allow for "higher-order vagueness." The account presupposes that the notion of a sharpening is precise. Yet it seems that the notion is itself vague: might there not be a borderline case of a word being definitely true of something? "Is he *definitely bald*? I'm not sure; maybe so, maybe not." If there is vagueness here, then it would be wrong to assume that the notion of a sharpening is precise.

Let us suppose that is right: *sharpening* is vague. How does this amount to a problem for supervaluational accounts? One could not turn to supervaluationism to provide a precise description of vagueness; but that is consistent with its providing a correct description. Even if there is vagueness about which sentences are true on all sharpenings, it is not vague whether there are sentences in standard sorites arguments which are neither true on all sharpenings nor false on all sharpenings (see question 3.18). The supervaluationist's explanation of how standard sorites arguments should be resisted, by denying that all the premises are true, remains intact. The demand that vagueness be described in precise terms goes well beyond the demand for an adequate solution to sorites paradoxes. So perhaps there is no problem after all (see question 3.19).

3.18

This assumes that, even if "sharpening" is vague, each vague predicate has at least one clear case of a sharpening. What example could you give to illustrate this?

3.19

How would you assess the following objection to a version of the supervaluational account which allows that the notion of a sharpening is vague?

> If "sharpening" is vague, then no sentence can be definitely true. Truth involves appeal to *all* sharpenings; what is to count as a sharpening is vague, so it is not definitely true of any collection that these are all the sharpenings there are (with respect to a given predicate). However, it is absurd to suggest that "Yul Brynner was bald" is anything other than *definitely* true.

3.6 Rejecting the reasoning: degrees of truth

We envisaged three possible responses to the paradox:

(a) to accept the conclusion;
(b) to reject the reasoning; or
(c) to reject one or more premises.

We have taken (a) to be a last resort. We have explored two versions of response (c): the epistemic theory and the supervaluational theory. I now want to turn to what has sometimes been classified as a (b)-type response. (As we will see, this classification is open to question.)

We are strongly disinclined to allow that there could be anything wrong with modus ponens. Nevertheless, some theorists have tried to place the

blame on this principle of reasoning, and I will try to explain their grounds.

When asked to assess a claim to the effect that a sixteen-year-old is an adult, it is natural to say something like "That is *to some extent* true," or "There is *a certain amount of truth* in that"; likewise whenever a vague predicate is applied to a borderline case. The response to the paradoxical argument I now wish to consider takes this very seriously. The suggestion is that we introduce *degrees of truth*. If a predicate like "adult" definitely applies to an object we will say the application has the maximum degree of truth, conventionally 1. If a predicate definitely does not apply to an object we will say the application has the minimum degree of truth, conventionally 0. Borderline cases will be registered by intermediate degrees of truth. Ascribing "bald" to a man who nearly qualifies as bald will rate a degree of truth closer to 1 than applying it to a man who nearly qualifies as non-bald. A degree of truth theory thus takes very seriously the point that the meaning of a vague word says something about the borderline cases, by contrast with supervaluationists, who see borderlines as marking incompleteness of meaning. Degree theory seeks to represent *what* the meaning says by the various degrees of truth.

How can a degree theory dissolve the paradoxical argument? It must assign the highest degree of truth to the categorical premise of the argument and the lowest degree to the conclusion; how will it treat the conditional premises?

Suppose collections of grains of sand start becoming borderline for "heap" at around the 100 mark. What should a degree theorist say about the conditional

If this 95-grained collection is a heap, so is this 94-grained collection?

The antecedent of the conditional is

This 95-grained collection is a heap.

The consequent is

This 94-grained collection is a heap.

According to the degree theory, the antecedent is nearly but not quite true. Perhaps it is assigned the degree of truth 0.96. The consequent is also nearly true, but not quite so nearly true as the antecedent. Perhaps it is assigned the degree of truth 0.95. What degree of truth should be assigned to the conditional itself?

There is room for variation in detail, but the general idea is that if the antecedent of a conditional is truer than its consequent, then the conditional cannot be wholly true; thus the conditional in question needs to be

assigned a degree of truth less than 1. The justification for this lies in part with the analogy with the standard case in which degrees of truth are not taken into account: we say that a conditional whose antecedent is true and whose consequent is false cannot be true, because a conditional should not lead from truth to falsehood. Analogously, a conditional should not lead to a lower degree of truth. The greater the amount of truth lost in the passage from antecedent to consequent, the lower the degree of truth of the whole conditional.

So far, the degree theorist's response is of type (c): reject the premises. On this theory the conditional premises, though very nearly true, are not quite true. Hence we need not be fully committed to them. Hence the paradoxical argument does not commit us to the paradoxical conclusion. However, the account needs to go further. Even if we are not fully committed to the premises, they are very nearly true. The degree theorist has to explain how we can have premises that are all very nearly true, yet a conclusion that is wholly false.

One way to do this, though not the only one, is to deny that the conclusion follows from the premises, and thus deny the validity of modus ponens; this is a response of type (b). On the degree-theoretic account envisaged here, modus ponens does not preserve degree of truth: the conclusion of an argument of the form "If p, then q; p, therefore q" may have a lower degree of truth than any of the premises. The conditional mentioned earlier, about 95-and 94-grained collections, is extremely close to the whole truth; perhaps it has a truth degree of 0.99. The antecedent, we suggested, had a truth degree of 0.96. Yet applying modus ponens yields a conclusion with a truth degree of only 0.95, lower than the truth degree of either of the premises. Modus ponens is straightforwardly valid as applied to sentences with the extreme truth degrees, 0 or 1: that is, one cannot get a conclusion with degree less than 1 from premises of degree 1. However, in the intermediate degrees the application of modus ponens can lead to a "leakage" of truth. The leakage may be small for each application, but can be large if the number of applications is large, as in the case of the paradoxical argument; and it can be regarded as a sufficient condition for the invalidity of modus ponens.

We earlier thought that modus ponens was a principle that simply could not be abandoned. What the degree theorist suggests, however, may well be consistent with all we really believed about modus ponens. There are two reasons for this.

The first is that normally we have in mind only the cases in which modus ponens is applied to sentences that are (*completely*) true or (*completely*) false. For these cases, the degree theorist's view agrees with our intuitions.

The second is that it is arguably not correct to assume, as we have so far, that an argument's validity is properly understood in terms of preservation of degree of truth, in the sense that the conclusion of a valid argument must have a degree of truth no lower than that of the least true premise. Perhaps it is better (and it would certainly preserve tempting analogies with degrees of belief) to say that the conclusion of a valid argument cannot have a greater degree of falsehood than the sum of the degrees of falsehood of its premises (Edgington 1992). (If a number, n, measures a degree of truth, then $1–n$ measures a corresponding degree of falsehood.) Then degree theory need give us no reason to say that modus ponens is not valid (see question 3.20). If we adopt this position, we must reclassify the degree theory as one which rejects the premises of sorites arguments, or rather, does not fully accept them.

3.20

Why not?

A full defense of the degree of truth theory would require the consideration of a number of issues that I shall briefly mention. First, it is necessary to say something about what a degree of truth is. Second, some account must be given of the source and justification of the numbers that are to be assigned as degrees. Third, the full implications of the degree theory for logic must be set out and defended.

A degree theory treats vagueness as a semantic matter, a matter of meaning, and not as an epistemic one, a matter of knowledge or ignorance. The semantic theory of degrees of truth registers the semantics of a vague predicate as different in kind (since involving the intermediate degrees) from the semantics of a precise predicate. So the first thing a degree theorist would have to do is refute the epistemic theory. Let us suppose that this has somehow been accomplished. The remainder of the defense presupposes that vagueness involves there being no sharp boundaries.

A key property of truth is marked by the platitude that we aim to believe what is true. If we could show that degrees of truth had an analogous property, we would have gone some way toward explaining what a degree of truth is.

Suppose that you are fairly sure that Arkle won the Gold Cup in 1960. Your memory may fail you about some matters, but you are pretty reliable about the history of the turf. You reckon that you have a very much better than fifty-fifty chance of being right that it was Arkle. If you are at all attracted by gambling, it will be rational for you to bet on his having won, if

you can get odds as good as even; for if you follow this policy generally, you will win more than fifty times out of a hundred. A policy that will result in your winning more than you lose is a policy that it is rational for you to pursue. It is rational to perform a particular action that is required by the pursuit of a rational policy.

We want to believe what is true, but we do not always know what is true. The greater the confidence we have in a proposition, the more it affects us as if we believed it to be true. If we are almost certain that our house will not burn down, we will not spend much money insuring it against fire. If we are almost certain that we shall be alive tomorrow, we do not waste much time today making arrangements for our death.

It is rational for our beliefs to vary in strength, reflecting variations in our confidence, and thus variations in our assessment of the quality of our information. We may be less than totally confident because we are less than fully informed. The less than total confidence mirrors our epistemic deficiencies.

Vagueness may also lead to less than total confidence. Suppose you know, having had it on impeccable authority, that all and only red mushrooms are poisonous. You wish to kill Jones. All other things being equal, you would prefer to poison him, and you would prefer to do so using mushrooms, so that it will look like an accident. However, the only mushroom you can find right now, though reddish, is not a clear case of red. Will you use it to try to poison Jones? It depends upon how important it is to you to succeed, how important it is to succeed at first attempt, and how soon Jones must die if his death is to be of service to you. How reasonable it is to use the mushroom depends upon the weights you assign to these other factors, and upon your degree of confidence in the redness of the mushroom. The more confident you are that this mushroom is really red, the more reasonable it is to use it; the less confident, the less reasonable. In the context, this confidence affects your action in the same way as would a lack of confidence springing from lack of information, from fear that your memory fails you, or whatever. From the point of view of action, it is rather as if you had less than total confidence in the statement "This mushroom will do the job."

There is also a sharp contrast. Less than total confidence springing from incomplete evidence or fear of unreliability mirrors our epistemic deficiencies; but according to degree theorists, less than complete confidence springing from an appreciation of vagueness does not. If the mushroom is a borderline case, it is not your fault that you are unsure whether it should be counted as red; indeed, you would be at fault if you took yourself to be required firmly to classify it either as red or as not red. Continuing our assumption that the epistemic theory has been rejected, we can go further: no matter how

perfect your memory and senses, no matter how infallible your reasoning, the mushroom stays on the borderline. On the question of whether the mushroom is red, an omniscient being could do no better.

Where we have incomplete information, or unreliability, there is a chance of improvement: we can in theory raise our confidence by getting more information. Where we have vagueness, there may be no chance of improvement. Given your language and the way the world is, you can do no better than have partial confidence in "This mushroom is red." Truth is what we seek in belief. It is that than which we cannot do better. So where partial confidence is the best that is even theoretically available, we need a corresponding concept of partial truth or degree of truth. Where vagueness is at issue, we must aim at a degree of belief that matches the degree of truth, just as, where there is no vagueness, we must aim to believe just what is true.

The second part of a defense of a degree of truth theory is to explain and justify the origin of the numbers that are assigned as degrees of truth. Suppose there are two mushrooms, both borderline cases of red, but one redder than the other. If you want to commit the poisoning, and you have full confidence in the information that all and only red ones kill, you should choose the redder if you choose either. The redder one must be closer to being a definite case of red. This suggests how we could justify assigning degrees of truth: we have to assign a higher degree to a redder object; or, if we are dealing with "heap," we must assign a higher degree to borderline cases the more numerous they are, that is, the closer they come to being definite cases for "heap." In short, the source and justification of assignments of degrees of truth would lie in our comparative judgments involving borderline cases (see questions 3.21, 3.22).

3.21

How would you respond to the following objection?

It is one thing to say that the comparative form of "red," namely "redder than," is to be used as the basis for assigning degrees in connection with "red"; but it is quite another thing to apply this to "heap." The basis for the assignments would be comparisons involving "heaper than"; but this is nonsense.

3.22

How would you respond to the following objection?

I agree that there are degrees of redness, but I cannot see that this means that there are degrees of truth.

The third part of the defense of a degree theory would involve justifying how one ascribes degrees of truth to logically complex sentences. Degree theories of the kind under consideration, in which the degree of truth of a complex sentence is determined by the degree of truth of its constituents, depart from ordinary, so-called classical, logic. Whereas classical logic has it that all sentences of the form "*p* and not-*p*" are false, and all sentences of the form "*p* or not-*p*" are true, some degree theorists demur. On their view, when *p* has only a medium degree of truth, "*p* and not-*p*" will not be completely false, and "*p* or not-*p*" will not be completely true. We have already seen, in the case of the argument about the harmful effects of watching hard-porn movies, that there is at least some case for holding that, if *p* is vague, "*p* or not-*p*" is not without qualification true. Furthermore, the naturalness of "It is and it isn't," as a response to the question whether a borderline-case mushroom is red, gives at least a preliminary indication that the degree theorist is right to recognize that not all instances of "*p* and not-*p*" are completely false.

However, there are problems. Standard forms of degree theory assign a conjunction a degree of truth equal to that of the least true conjunct. This means that, if *x* is a clear case of something red, *y* is a borderline case of red, and both *x* and *y* are, to the same degree, intermediate cases of something small, the conjunctions "*x* is red and *x* is small" and "*y* is red and *y* is small" will be assigned the same intermediate degree of truth, which is unintuitive (see question 3.23). Again, suppose Eve is definitely female and a borderline case for being an adult, so that "Eve is an adult" is assigned an intermediate degree of truth. Assume that it is definitely true that a woman is an adult female. Standard degree theory cannot distinguish between "Eve is an adult or Eve is a woman" and "Eve is an adult and Eve is not a woman," assigning both the same intermediate degree (Fine 1975).

3.23

Which conjunction do you think should have the higher degree of truth? Cf. Edgington (1992).

The degree theory, like the supervaluation theory, appears to be at risk from considerations relating to higher-order vagueness. On the colored wall discussed previously, red gradually gives way to orange from left to right. At the left end, a sentence "This patch is red" will presumably be assigned degree 1. However, which is the last patch relative to which the sentence is assigned degree 1? If we allow that there is a last such patch,

then it seems that we have introduced a borderline after all: between the definite reds and the others. If we say that there is no such patch, we seem to be committed to the absurdity that even when the sentence is applied to an orange patch it is still awarded the maximum degree of truth.

Part of the problem is that we tend to think of semantic theories, like supervaluational theory or degrees of truth theory, as themselves expressed in a sharp language (the metalanguage). This means that we tend to think that vagueness can be described in precise terms. The phenomenon of higher-order vagueness suggests (though it does not entail) that we are mistaken in this tendency.

If we try to describe a vague language by a vague one, we may not have conquered the problems concerning vagueness with which we began. Sorites paradoxes may threaten the language in which our semantics is expressed. If "has degree of truth 1" is precise, then it seems to draw a sharp boundary where no boundary should be. If it is vague, we will be tempted to suppose that if applying "red" to one of two indistinguishable patches merits degree of truth 1, so does applying it to the other, and we will be set on the familiar slippery sorites slope.

3.7 Vague objects?

Are there vague *objects*, or is vagueness something that arises, not from the way the world itself is, but rather from how we describe it? This question is answered by the epistemic theory: the home of vagueness lies in our cognitive faculties, not in the world. An answer is also (optionally) suggested by the degree theory: if the degrees are based on the responses of subjects to borderline cases, perhaps vagueness is a matter of human psychology rather than a feature of the world (Schiffer 2003). Those who believe that vagueness is properly to be described by saying that borderline cases are ones in which there is no fact of the matter might lay the blame on the meanings of the words; but they also might lay the blame on the world. Can we make sense of the idea that the world itself is vague?

We can start by reverting to an earlier question (3.6). The argument for discussion went like this:

> Mountains are part of reality, but they are vague. They have no sharp boundaries: it is vague where the mountain ends and the plain begins. So it is easy to see that vagueness is a feature of reality, and not just of our thought and talk.

Even if we like the conclusion, we should not accept this argument for it. Given our language, which contains words like "mountain," we can ask a vague question: does this spot belong to the mountain or to the plain?

However, if we could give a complete description of the world without making use of such a vague expression, we would have no inclination to infer from the vagueness of the question to the vagueness of the world. We would not seem to be greatly handicapped, in describing the world, if we lacked the word "mountain": we could just draw the contour lines on our maps. In many cases, for example "heap," our use of the vague word is guided by sharp underlying facts, for example, by (in part) how many grains the collection contains. Each collection has a definite number of members, and we could in principle give more information about a collection by the sharp fact of how many members it has than by the vague matter of whether or not it is a heap. So the displayed argument does not undermine the view that vagueness comes from our thought and talk, rather than being an objective feature of the world.

Let us consider an old story. Theseus had a ship. When a plank rotted, it was replaced, and thanks to the repair the ship remained in service. After a while, none of the original planks were left. Likewise for the other kinds of parts of the ship – masts, sails, and so forth. Did Theseus' ship survive? There was a ship in continuous service, and we incline to hold that this is indeed Theseus' ship, much repaired. But suppose that someone had kept the rotted planks and other parts and then reassembled these into a (possibly unseaworthy) ship. Does this have a better claim to be the original ship of Theseus? There is vagueness of some kind here. The question is: is the ship *itself* vague, or does the vagueness end with the word "ship," leaving the ship itself uncontaminated?

It seems to me that the second answer is the right one. In such a case, we can give an agreed and relatively precise account of the "facts of the matter." We know just what happened. It is a verbal question to which object, if any, we ought to apply the phrase "the ship of Theseus." So the vagueness comes from words.

This view could be supported by a two-stage argument. First, show that *identity* is not a vague relation; that is, show that questions of the form "Is this thing the same as that thing" have definite answers. If so there should be a definite answer to whether the long-serving ship that Theseus still sails is the same as the ship later reassembled from the discarded parts of Theseus' original. The suggestion is that, quite generally:

If β is α, then β is definitely α; and if β is not α, then β is definitely not α.

Suppose β is α. It seems indisputable that, for any object x,

x is definitely x.

So "is definitely α" is true of α. If β is α, anything true of α is true of β. So "is definitely α" is true of β. Hence:

β is definitely α (see question 3.24).

3.24

Can you establish the rest of what is needed for the non-vagueness of the identity relation, namely:

if β is not α, then β is definitely not α?

For help with this tricky question, consult Evans (1978) and Wiggins (1986).

Suppose we think that it is vague whether the ship Theseus still sails is the same as the one assembled from the discarded parts; that is, that they are not definitely identical. The above argument would show that we should conclude that the ships are distinct.

The second stage of the argument involves showing that if identity is not a vague relation, then objects are not vague. The idea is that if an object were vague, it would be a vague matter what object it is identical with. Since the first part of the argument has supposedly shown that identity is not vague, the conclusion is drawn that objects are not vague.

I close with four qualms. First, the second step of the argument is not cogent as stated. It is not unreasonable to suppose that, if there are any vague objects, collections with vague membership conditions are included. So if there are uniformly colored blocks of various colors and weights on my table, including various red ones and orange ones, the collection of red blocks on my table has a vague membership condition, and for some block, it may be a vague matter whether it belongs to the collection or not. The collection seems as good a candidate as any for being a vague object. Yet we might insist that the identity conditions for such collections are completely precise. If we ask whether the collection of red blocks is the same collection as the collection of heavy blocks, we might insist that an affirmative answer requires that the collections coincide in every respect: all the definite members of the red collection must be definite members of the heavy collection and vice versa; all the definite non-members of the red collection must be definite non-members of the heavy collection and vice versa; and so on through any further distinctions. There would then never be any vagueness about whether this collection is the same as or different from that collection. We thus have not excluded the possibility of vague objects without vague identity.

Second, denying that there are vague objects seems to presuppose that the "facts themselves" are precise. I said that, in the case of Theseus' ship, the facts of the matter are "relatively precise." They are precise relative to

the vagueness of "ship," since they can be stated without using that word. However, other words, such as "plank," have to be used. This is just as vague as "ship." Can we be sure that there is a range of ultimate facts that can be described without using any vague expressions at all? Such a belief would demand careful justification.

The third qualm is this: identity over time, as discussed in the case of the ship, must be governed by principles such as: replacing some, but not too many, parts of an artifact does not destroy it, but leaves you with the very same artifact. Such principles are vague. How could the identity relation, which they determine, be precise (see question 3.25)?

3.25

How would you evaluate the following argument?

> If you exist at all, you are a vague object, for we believe that a molecule more or less cannot make the difference between whether you exist or do not. On this basis, we can construct a paradoxical argument: taking away one molecule will not make you cease to exist, taking away one more will not make you cease to exist, and so on; thus you can exist even if no molecules of you do. This shows that you are as vague as a heap. However, there are no vague objects, therefore you do not exist.

For essentially this argument, see Unger (1979b).

The final qualm is more technical: the best implementation I know of the semantic approach mentioned at the end of the previous section, in which a vague language is used to describe vagueness, posits the existence of vague objects, namely, vague sets. This might give one a starting point from which to extend belief in vague objects more widely.

Suggested reading

The *Stanford Encyclopedia* is likely to be the first port of call for those wishing to explore vagueness further. Two relevant entries are Sorensen's (2006a) on vagueness (plato.stanford.edu/entries/vagueness/) and Hyde's (2005) specifically on sorites paradoxes (plato.stanford.edu/entries/sorites-paradox/). Both these entries have good bibliographies. For overviews of vagueness, which also contain defenses of the authors' own views, see Williamson (1994, ultimately defending epistemicism) and Keefe (2000, ultimately defending supervaluationism). The present chapter does not mention contextualist approaches to vagueness and the sorites. Starting points would be Kamp (1981), Raffman (1994) and Graff-Fara (2000). For

a very comprehensive bibliography, see the Arche website at St. Andrews, in particular: www.st-andrews.ac.uk/~arche/projects/vagueness/bibliography.shtml.

Section 3.1

An introductory article is Black (1937); see also Dummett (1975). For a survey of treatments of sorites paradoxes, see Sainsbury and Williamson (1995). For an argument for the utility of vagueness, see Wright (1975). For his more recent views see Wright (1987) and (2001).

Section 3.3

For the view that there are no tables, see Unger (1979a).

Section 3.4

The epistemic theory can be traced back to Chrysippus (*c.* 280 – *c.* 207 BC). Its current main defenders are Sorensen (1988) and Williamson (1992a, 1992b, 1994) (they defend different versions of the theory). See also Cargile (1965). Among many critics, see Wright (1995) and Ludwig and Ray (2002).

Section 3.5

A classic text for the supervaluation theory is Fine (1975); see also Van Fraassen (1966); Kamp (1975). These papers involve some technicalities. For a less formal account of the underlying idea, see Dummett (1975, esp. pp. 256–7). For a brief history, a critical discussion, and a development of the theory to accommodate higher-order vagueness, see Williamson (1994); also Keefe (2000: ch. 7). For the relation between supervaluation theory and higher-order vagueness, see also Fine (1975, esp. §5).

The criticism of supervaluation theory given in the text, that it verifies the intuitively unacceptable "For some number n, an n-grained collection is a heap but an (n–1)-grained collection is not," has been attacked. Various writers have suggested that the quoted sentence is acceptable, as long as we realize that its truth does not require that there be a number n such that the following is true:

An n-grained collection is a heap but an (n –1)-grained collection is not.

See, e.g. Dummett (1975, pp. 257–8); and, for opposition, Kamp (1981, pp. 237ff.). The claim depends upon two features: (1) a view of "there is"

according to which it is like "or" (so that to say that there is a student who smokes is to say that either Sally smokes, or Michael smokes, or …, and so on through all the students); and (2) a view of "or" according to which a statement "*p* or *q*" can be (definitely) true even though neither "*p*" nor "*q*" is. A standard alleged example of the latter is "This is orange or red," said of a borderline case. See again Dummett (1975, p. 255). It is not at all clear whether the combination of (1) and (2) is less paradoxical than the paradox of the heap. It involves claiming that "there is a such-and-such" can be true, even if "that is a such-and-such" is false of each thing in the universe.

For a defense of supervaluationism from the objections fielded in the text see Keefe (2000).

For higher-order vagueness, see Wright (1992), the symposium between Hyde (1994) and Tye (1994b), and Williamson (1999).

Section 3.6

A classic source for degree theory is Zadeh (1965). I had mostly in mind the kind of theory advanced in Goguen (1969). For a version which does not treat the connectives as degree functional see Sanford (1975). For a more philosophical and less technical account of degree theory, see Peacocke (1981). For criticisms see Edgington (1992), Sanford (1976), and, on the resulting logic, Fine (1975), and Williamson (1994). For an account of validity and degrees according to which modus ponens remains valid, see Edgington (1992).

For the notion of partial belief, see Ramsey (1926) and Jeffrey (1965, chs. 3 and 4). A substantive question is whether a similar argument could be used to underwrite objective probabilities. If the answer is affirmative, as Mellor (1971) argues, then a question of crucial importance would be whether one can give a satisfactory account of why the arguments reach different destinations: degrees of truth in one case, objective probabilities in the other. Edgington (1992) suggests that less than full confidence induced by vagueness behaves differently from less than full confidence induced by lack of information. Recent discussions of some versions of degree theories can be found in Schiffer (2003) and MacFarlane (2007) (a hybrid view, combining degrees of truth with some aspects of epistemicism).

For a semantics with a vague metalanguage, see Tye (1994a).

Section 3.7

On vague objects, see Evans (1978), Lewis (1988), Copeland (1994), Salmon (1982, pp. 243ff.), Tye (1990), Wiggins (1986) and Williamson (2003).

4 Acting rationally

4.1 Newcomb's paradox

You are confronted with a choice. There are two boxes before you, *A* and *B*. You may either open both boxes, or else just open *B*. You may keep what is inside any box you open, but you may not keep what is inside any box you do not open. The background is this.

A very powerful being, who has been invariably accurate in his predictions about your behavior in the past, has already acted in the following way:

He has put $1,000 in box *A*.
If he has predicted that you will open just box *B*, he has in addition put $1,000,000 in box *B*.
If he has predicted that you will open both boxes, he has put nothing in box *B*.

The paradox consists in the fact that there appears to be a decisive argument for the view that the most rational thing to do is to open both boxes; and also a decisive argument for the view that the most rational thing to do is to open just box *B*. The arguments commend incompatible courses of action: if you take both boxes, you cannot also take just box *B*. Putting the arguments together entails the overall conclusion that taking both boxes is the most rational thing and also not the most rational thing. This is unacceptable, yet the arguments from which it derives are apparently acceptable.

An argument for opening both boxes goes like this. The powerful being – let us call him the Predictor – has already acted. Either he has put money in both boxes or he has put money in just box *A*. In the first case, by opening both boxes you will win $1,001,000. In the second case, by opening both boxes you will at least win $1,000, which is better than nothing. By contrast, if you were to open just box *B*, you would win just $1,000,000 on the first assumption (i.e. that the Predictor has put money in both boxes) and nothing on the second assumption (i.e. that the Predictor has put money just in box *A*). In either case, you would be $1,000 worse off than had you opened both boxes. So opening both boxes is the best thing to do.

An argument for opening just box *B* goes like this. Since the Predictor has always been right in his previous predictions, you have every reason for thinking that he will be right in this one. So you have every reason to think that if you were to open both boxes, the Predictor would have predicted this and so would have left box *B* empty. So you have every reason to think that it would not be best to open both boxes. Likewise, you have every reason to think that if you choose to open just box *B*, the Predictor will have predicted this, and so will have put $1,000,000 inside. Imagine a third party, who knows all the facts. He will bet heavily that if you open just box *B* you will win $1,000,000. He will bet heavily that if you open both boxes you will get only $1,000. You have to agree that his bets are rational. So it must be rational for you to open just box *B*.

This paradox has been used to compare two different principles for determining how it is rational to act. One principle is this: act so as to maximize the benefit you can expect from your action. In stating this principle, "benefit" is usually replaced by the technical term "utility." Part of the point of the technical term is to break any supposed connection between rationality and selfishness or lack of moral fibre. A benefit or "utility" consists in any situation that you want to obtain. If you are altruistic, you may desire someone else's welfare, and then an improvement in his welfare will count as a utility to you. If you want to do what is morally right, an action will attract utility simply by being in conformity with what, in your eyes, morality requires, even if from other points of view, say the purely material one, the consequences of the action are not beneficial to you.

There is obviously something appealing in the principle that it is rational to act so as to *maximize expected utility* – MEU for short. Consider gambling: the bigger the prize in the lottery, the more money it is rational to pay for a ticket, everything else being equal; the larger the number of tickets, the less money it is rational to pay. The MEU principle tells you to weigh both these factors. If there are 100 tickets and there is just one prize of $1,000, then you will think that you are doing well if you can buy a ticket for less than $10. (If you could buy them *all* for less than $10 each, then you could be certain of gaining $1,000 for an expenditure of less than $1,000.) If the tickets cost more than $10, you may have to think of the lottery as a way of raising money for a charity that you wish to support, if you are to buy a ticket.

Such an example contains a number of quite unrealistic assumptions. Some of these are inessential, but at least one is essential if the MEU principle is to compare any possible pair of actions for their degree of rationality. This is the supposition that utilities and probabilities can be measured (see questions 4.1, 4.2, 4.3). If they can, then we can simply

compute which of the actions open to us have greatest expected utility: we multiply the measure of utility by the measure of the probability of that utility accruing.

4.1

Suppose that on Monday you are penniless and starving, but that on Tuesday you win $1,000,000 in a betting pool. Do you think that the number 5 can be used to measure the utility of $5 to you on each of these days?

4.2

Suppose you have four courses of action open to you, (a)–(d), associated with rewards as follows: (a) $1, (b) $6, (c) $10,000, (d) $10,005. Do you think that the number 5 can be used to measure both the difference between the utilities of (a) and (b) and the difference between the utilities of (c) and (d)?

4.3

* Discuss the following view:

Although people want things other than money, we can use numbers to measure how much they want things, by finding out how much they would be willing to pay, supposing, perhaps *per impossibile*, that what they want could be bought. If a man says he wants a happy love affair, we can measure the utility of this upshot to him by finding out how much money he would be willing to give up to get what he wants. $1000? $10,000? Would he be willing to sell his car to raise the money? His house? All that is needed is the ability to imagine things being other than they are: to imagine that things that in fact cannot be bought can be bought.

Lotteries are useful examples of how probability can be measured. If a lottery is fair, each ticket has an equal chance of winning: the chance of any ticket in an n-ticket lottery winning is simply $1/n$. Probability theorists generalize this notion in the following way. Suppose that there are just x relevant and exclusive possibilities (e.g. that ticket number 1 is the unique winner, that ticket number 2 is the unique winner, and so on), and that of these y have the property P (e.g. that the winning ticket is mine). Then the probability that what actually happens has property P (e.g. that my ticket wins) is given as y/x. This means that the highest measure of probability is

1 (when $x = y$, e.g. if I buy all the tickets) and the lowest 0 (if $y = 0$, e.g. if I buy no tickets), with the intermediate measures lying in between.

Lotteries are also useful examples of how utilities can be measured, if we allow ourselves the simplifying assumption that the cash value of the win represents its utility. So we can readily use lotteries to exemplify expected utility. Suppose that there are two lotteries, one with 1,000 tickets and a single \$1,100 prize, and one with 900 tickets and a single \$1,000 prize. Expected utility is the chance you think you have of winning times the utility of the win. For the first lottery this can be represented as $1/1{,}000 \times 1{,}100 = 1.10$. For the second, it can be represented as $1/900 \times 1{,}000 = 1.11$ (approx.). The expected utility of the second lottery is a shade higher than that of the first. So if the tickets for both lotteries cost the same, and you are going to buy a ticket for one, the MEU principle recommends that you choose the second.

The MEU principle does not recommend that you buy a ticket in either lottery. There may well be many alternative ways of spending your money with expected utilities higher than those associated with either lottery. The principle only tells you that *if* you are going to buy a ticket for either, it should be for the second (see question 4.4).

4.4

* Could the MEU principle register a general dislike of gambling, as opposed to other ways of spending money? If so, how?

The notion of utility was introduced in terms of what upshot an agent wants. What someone wants sometimes means what he or she wants all things considered. If I decide to go to the dentist, then typically I want to go – that is, want to go all things considered. However, what a person wants can also mean anything to which he attaches some positive value. In this sense, it is true of me, when I freely and willingly go to the dentist, that I want not to go: not going has the positive value of sparing time and the discomfort associated with the dentist's chair. If I go, it is because this want is trumped by another: I want to avoid decay, and for the sake of that benefit I am prepared to put up with the loss of time and the discomfort. The appropriate connection between utility and wanting should exploit not what an agent wants overall, but rather that to which he attaches any positive value.

The situation that gives rise to Newcomb's paradox is represented in table 4.1. The expected utility of opening both boxes is calculated as follows. Knowing the Predictor's record, you regard it as very likely that the Predictor will have correctly predicted your choice. Hence if you open

Table 4.1 *Newcomb's paradox.*

	The Predictor has *not* put money in *B*	The Predictor *has* put money in *B*
you open *A* and *B*	$1,000	$1,001,000
you open just *B*	$0	$1,000,000

both boxes you must think that it is very likely that the Predictor will have predicted this and so will have put no money in box *B*. So the expected utility is some high ratio close to 1, call it *h*, measuring the likelihood of this outcome, multiplied by 1,000, measuring its utility (the total sum you will collect). Analogously, the expected utility for you of opening just box *B* is the same high ratio, measuring the likelihood of the Predictor having correctly predicted that this is what you would do, and so having put $1,000,000 in box *B*, multiplied by 1,000,000, measuring the utility of that outcome. Since, whatever exactly *h* may be, 1,000 × *h* is much less than 1,000,000 × *h*, MEU recommends opening just box *B*.

I will set out the calculations in more detail. (Readers who feel they have a good enough hang of them should skip to the next paragraph.) The expected utility of an action is calculated as follows. First, you determine the possible outcomes O_i. Each is associated with a probability, conditional upon doing *A*, and a utility. The expected utility of an *outcome*, relative to an action *A*, is the product of its utility and its probability given *A*. The expected utility of an action *A* is the sum of the expected utilities of its outcomes relative to *A*:

$$EU(A) = [prob(O_1/A).U(O_1)] + [prob(O_2/A).U(O_2)] + \cdots$$

Here "EU *(A)*" stands for the expected utility of *A*, "prob(O_i/*A*)" for the probability of outcome O_i given *A*, and "U(O_i)" for the utility of that outcome. Applied to Newcomb's paradox, using *B* for the action of opening only box *B*, and *A*&*B* for the action of opening both boxes, we have:

$$\begin{aligned} EU(B) &= [prob(B \text{ is empty}/B).\ U(B \text{ is empty})] + [prob(B \text{ is full}/B).\ U(B \text{ is full})] \\ &= (1-h).0 + h.1{,}000{,}000. \\ EU(A\&B) &= [prob(B \text{ is empty}/A\&B).U(B \text{ is empty and } A \text{ is full})] \\ &\quad + [prob(B \text{ is full}/A\&B).U(B \text{ is full and } A \text{ is full})] \\ &= h.1{,}000 + [(1-h).1{,}001{,}000]. \end{aligned}$$

Setting $h = 0.9$ makes EU*(B)* = 900,000 and EU*(A&B)* = 101,100, giving a nearly ninefold advantage to taking just box *B*.

The MEU principle underwrites the argument for opening just box *B*. To resolve the paradox, one would need to show what was wrong with the other argument, the argument for opening both boxes. Those who are persuaded that it is rational to open both boxes will regard the fact that the MEU principle delivers the contrary recommendation as a refutation of the principle.

One attractive feature of MEU is that it is a quite general, and independently attractive, principle. Are there any other principles of rational action that are also attractive, yet that deliver a different recommendation? There are. One example is the so-called *dominance principle* – DP for short.

According to DP, it is rational to perform an action α if it satisfies the following two conditions:

(a) Whatever else may happen, doing α will result in your being no worse off than doing any of the other things open to you.
(b) There is at least one possible outcome in which your having done α makes you better off than you would have been had you done any of the other things open to you.

DP has commonsensical appeal. If you follow it you will act in such a way that nothing else you could do would have resulted in your faring better, except by running the risk of your faring worse.

Table 4.1 may be taken to show that opening both boxes satisfies DP, and that opening only box *B* does not (see question 4.5). Whatever the Predictor has done, you are better off opening both boxes than opening just one. In either case, you stand to gain an extra $1,000 as compared with the other course of action open to you. Hence DP and MEU conflict: they commend opposite courses of action.

4.5

* Suppose that we think of the outcomes in a different way, as shown in table 4.2.

Table 4.2

	The Predictor has predicted correctly	The Predictor has predicted incorrectly
You open *A* and *B*	$1,000	$1,001,000
You open just *B*	$1,000,000	$0

Opening both boxes appears no longer to dominate opening just one box. How should one respond?

One way to diagnose Newcomb's paradox is precisely as the manifestation of this conflict of principle. The constructive task is then to explain how the principles are to be restricted in such a way that they cease to conflict, while retaining whatever element of truth they contain.

How is the Predictor so good at predicting? Suppose it worked like this. Your choice would cause the Predictor to have made the correct prediction of it. To take this alleged possibility seriously, we have to take seriously the possibility of "backward causation": that is, a later event (here your choice) causing an earlier one (here the Predictor's prediction). Let us for the moment take this in our stride. If one knew that this was how things worked, surely there could not be two views about what it would be rational to do. One should open just box *B*, for this would cause the Predictor to predict that this is what one would do, which would lead to his having put $1,000,000 in box *B*. Not making this choice, by contrast, would lead to his not having put the $1,000,000 in box *B*. Clearly it would be crazy not to choose to open just box *B*.

The original case was, perhaps, underdescribed. Perhaps it did allow for the possibility (if there is such a possibility) of backward causation. To prevent confusion, let us stipulate that the *original case* is one that excludes backward causation. It is instructive, however, to consider this *other* case, where there is supposed to be backward causation. Perhaps the attraction of opening just box *B* in the original case sprang from thinking of it as the backward causation case. More generally, perhaps the paradox strikes us as paradoxical only to the extent that we confuse the original case with the backward causation case. To the extent that we think of the case as involving backward causation, we are tempted by MEU. To the extent that we think of it as excluding backward causation we are tempted by DP. What strikes us as conflicting views of the same case are really views of different cases.

In the original case, one might suppose that the Predictor bases his decision on general laws, together with particular past facts. These might all be physical, or they might be psychological. For example, the laws might be laws of psychology, and the particular facts might concern your personality. There is no question of backward causation. Then the basis for the prediction consists in facts that lie in the past. Rejecting backward causation means that nothing you can now do can affect the basis for the prediction. Hence nothing you now do can make any difference to whether there is or is not money in box *B*. So you should open both boxes.

There is a complicating factor. Suppose that determinism is true. Suppose, in particular, that the psychological laws, together with data about your character up to the time at which the Predictor made his prediction, determine how you will now act, in the sense of making it impossible for you to do anything other than what, in fact, you will do.

This may totally undermine the idea of rational decision, and so make the whole question of what is best to do one that cannot arise. In short, there is a case for saying that if there were such a Predictor, then there could be no question about which choice is rational. I shall ignore this case, and argue that it is rational to open both boxes. Those who are moved by it could read my conclusion as hypothetical: if we can make sense of rationality at all in the deterministic Newcomb situation, then the rational thing is to open both boxes.

In defending this conclusion, I need to consider whether, as claimed earlier, it is rational for the onlookers to bet heavily on the following two conditionals:

(a) if you select both boxes, box *B* will be empty;
(b) if you select just box *B*, it will contain $1,000,000.

If they are rational, the onlookers will bet in accordance with their expectations. Their expectations are the same as yours. They have very strong reason to believe the two conditionals, given the Predictor's past successes. How can this be reconciled with my claim that if the Predictor bases his prediction on past evidence, then it is rational to open both boxes? If it is rational for the onlookers to expect the conditionals to be true, it must be rational for you to expect the same. However, you have a *choice* about which conditional will count. It is rational, surely, to make the second conditional count, and you can do this by opening just box *B*. How can one reconcile the rationality of belief in the conditionals with the rationality of opening both boxes?

Let us look more closely at the basis of the rationality of belief in the conditionals. We can do this by looking at it from the point of view of the onlookers. They reason as follows. The Predictor has always been right in the past. Since he has already filled the boxes, his prediction is based on knowing some past facts about you and your circumstances, and applying some generalizations. Our best evidence for what he has predicted is what you choose to do. This is why we believe the conditionals. Your opening just box *B* is evidence that the Predictor has predicted this and, hence, by the way the problem is set up, is evidence that he has filled box *B* with $1,000,000. Likewise for the other possibility.

The rationality of these beliefs does not entail the rationality of opening just box *B*. This is most easily seen if we switch to the subject's point of view: to *your* point of view, as we are pretending. The Predictor makes his choice on the basis of past facts about you, together with some generalizations. To simplify, let us say that there are two relevant possible facts about what sort of person you were at the time the Predictor made his prediction: either you were a one-boxer – that is, a person disposed to

open just box *B* – or you were a two-boxer – that is, a person disposed to open both boxes. If you find yourself tempted to open both boxes, that is bad news (see question 4.6). It is evidence that you are now a two-boxer and, all other things being equal, is thereby evidence that you were a two-boxer at the time when the Predictor made his decision. Hence it is evidence that he will have predicted that you will open both boxes, and so it is evidence that there will be no money in box *B*. However, there is no point in trying to extirpate this disposition, and it would be a confusion to think that you could make any difference to the situation by resisting it. There is no point trying to extirpate it *now*, since the Predictor has made his prediction and either he noticed your two-boxing disposition or he did not. If he did notice it, getting rid of it now is closing the stable door after the horse has bolted. Nothing you can now do can make any difference as to whether or not you were a two-boxer at the time the Predictor made his prediction. Thus it can be rational to open both boxes even if it is also rational to believe that if you open both boxes, box *B* will probably be empty. It is rational to believe this conditional because it is rational to believe that if you open both boxes you are a two-boxer, and if you are now a two-boxer you probably were one when the Predictor scanned you, in which case he has probably predicted that you will open both boxes, in which case he has probably not put anything in box *B*.

4.6

How would you respond to the following argument?

It would come as wonderful news to learn that I am a one-boxer, for then I will be able to infer that I will soon be rich. However, I can give myself that news simply by deciding to be a one-boxer. So this is what I should decide to do.

If you found in yourself an inclination to open just box *B*, that would be good news, for analogous reasons; but it is an inclination that it would be more prudent to resist. By resisting it and opening both boxes, you cannot make the money that you can reasonably presume is already in box *B* go away, and you will gain the extra $1,000 in box *A*.

Here is an objection. If this is where the reasoning were to end, would not a really good Predictor have predicted this, and therefore have ensured that there is nothing in box *B*? Furthermore, had you taken the reasoning through a further twist, using the fact just mentioned as a reason for in the end taking just box *B*, the Predictor would have predicted this too, and so would have filled box *B*. So is not this what you should do?

The original difficulty remains and cannot be overcome. No matter what twists and turns of reasoning you go in for now, you cannot affect what the Predictor has already done. Even if you could make yourself now into a one-boxer, it would not help. What mattered was whether you were a one-boxer, or a person likely to become a one-boxer, at the time when the Predictor made his prediction. You cannot change the past (see questions 4.7, 4.8).

4.7

We have envisaged the choice before you being a once-in-a-lifetime chance. However, suppose you knew that you were going to be allowed to make this choice once a week for the rest of your life, and suppose the facts about the Predictor remain the same. What is the most rational policy to pursue?

4.8

Consider a variant of the problem – let us call it the "sequential Newcomb." The difference is that you are allowed to make your choice in two stages: you can elect to open box *B*, reserving your choice about box *A* until you find out what is in box *B*. Suppose you open *B* and there is nothing inside. Should you elect also to open *A?* Suppose you open *B* and there is $1,000,000 inside. Should you elect also to open *A?* Do your answers have any implications for the original Newcomb?

We have said that the Predictor has *always* been right in the past (see question 4.9). Let us imagine, in particular, that he has always been right about Newcomb problems. We shall suppose that each person is confronted with the problem only once in his life (there is no second chance), and that the Predictor has never been wrong: never has a two-boxer found anything in box *B*, and never has a one-boxer found box *B* empty. Most of your friends have already had the chance. The one-boxers among them are now millionaires. You wish above all things that you were a millionaire like them, and now your chance has come: you are faced with the Newcomb problem. Is it not true that all you have to do is choose just box *B?* Is that not a sure-fire way to riches? So how could it be rational to refuse it?

4.9

Consider a variant in which he has *mostly* been right in the past. Would this make any difference to the argument? Try working out what the MEU recommends if we set the probability of the Predictor being right at 0.6.

So far, this raises no new considerations: the two-boxer's reply still stands. However, I have put the matter this way in order to add the following twist. Being of good two-box views, you think that the Predictor is, for some crazy reason, simply rewarding irrationality: he makes one-boxers rich, and one-boxers are irrational. Still, if you want to be rich above all things, then is not the *rational* thing to do to join the irrational people in opening just box *B*? Sir John Harington (1561–1612) wrote that

> Treason doth never prosper; what's the reason?
> For if it prosper, none dare call it treason.

Likewise, if "irrationality" pays, then it is not irrationality at all! You want to be rich like your millionaire friends, and if you think as they do you will be. It is rational to adapt means to ends, so it is rational to think the way they think.

This suggestion can be represented as involving two main points. The first is that one might reasonably want to be a different sort of person from the sort one is: here, a less rational sort. Some people committed to lucidity and truth as values find this suggestion unpalatable (see question 4.10). However, a second point is needed: that if it is reasonable to want to be a different sort of person, then it is reasonable, even as things are, to act as that other sort of person would have acted. The second point is what secures the passage from envying the one-boxers to the claim that it would be rational to follow their lead. Once clearly stated, this second point can be seen to be incorrect: given that you are not a "natural" one-boxer, given that you are persuaded by the argument for two-boxing, nothing can as things stand make it *rational* for you to one-box. (Quite likely, most of us would succumb to irrational forces and one-box, but that does not show we would be rational to do so.)

4.10

What are your own views on this point? Some people say that they wish they could believe in life after death. If this wish involves wishing that they could cease to be moved by the evidence *against* life after death, it is an example of the sort of desire whose reasonableness or rationality is in question.

A clear perception of the advantages of being a one-boxer cannot give you a *reason* for becoming one – even if that were in your power. Atheists might clearly perceive the comfort to be derived from theism, but this does not give them a *reason* for believing that God exists. The light of reason cannot direct

one toward what one perceives as irrational. To adopt a position one regards as irrational one needs to rely on something other than reason: drugs, fasting, chanting, dancing, or whatever (see question 4.11).

4.11

Would you feel differently if there was just $1 in box *A* (with the same arrangements for box B)? Would you feel differently if there was $900,000 in box A (with the same arrangements for box B)? For discussion, see Nozick (1993, pp. 44–5).

This way of dealing with the paradox takes no account of the two principles, MEU and DP. Is either to be accepted? MEU cannot be correct, since it commends taking just box *B*. DP cannot be correct since, in the other version of the paradox, in which backward causation was admitted, DP wrongly recommended taking both boxes (see questions 4.12, 4.13). However, we may be able to see why the principles let us down when they did, and this may lead to ways of suitably restricting them.

4.12

Consider some familiar gambling game (e.g. roulette or poker). Can DP be used to say which bets in your selected game are rational? Assume that the only aim is to win as much money as possible.

4.13

Israel is wondering whether to withdraw from territories it occupies. Egypt is wondering whether or not to go to war with Israel. From Israel's point of view, the utilities are shown in table 4.3:

Table 4.3

	Egypt declares war	Egypt does *not* declare war
Israel withdraws	0	2
Israel remains	1	3

Show how this example can be used to demonstrate that DP does not always give correct results. (See Bar-Hillel and Margalit 1972.)

In the backward causation case, it is no accident that DP gives the wrong result. It has no means of taking into account the fact that your choice will affect what is in the boxes, by affecting the Predictor. More generally, it gives the wrong result because it makes no provision for the ways in which one's acting can affect the probabilities of outcomes. The backward causation case alone shows that DP cannot serve as it stands as a universally correct principle of rational action: it cannot be rational to act in such a way as to cause a diminution in the likelihood of someone else doing something that would increase one's benefits.

Equally, it is no accident that MEU gives the right result for the backward causation case. The rationale of MEU is given by the thought that it is rational to act in ways one takes to be likely to *promote* one's benefits. In the backward causation case, one has reason to believe that how one acts will affect one's benefits by affecting the Predictor's decision. In this case, the conditional probabilities reflect the probability of one's action genuinely promoting one rather than another outcome.

By contrast, this does not hold in the original case. The conditional probabilities obtain, but in a way that fails to reflect the underlying rationale of the MEU. The probability that if you open both boxes, box *B* will be empty is indeed high; but it is not high because your opening both boxes will have any causal role in bringing it about that box *B* is empty. The right restriction on MEU, so far as Newcomb's paradox goes, is that one should act on the principle only when the conditional probabilities reflect what one believes one's actions will *produce*.

We have seen that DP is not an acceptable principle of rational action, since it takes no account of conditional probabilities. This fact explains why it happens to give the right result in the original case. Here, because the probabilities are irrelevant, in that they do not reflect the likely effects of the possible actions, it is right to ignore them. So far as this case goes, one appropriate restriction on DP is that it can be used only when there is no relevant difference in the probability of the various possible outcomes.

Though these considerations explain away Newcomb's paradox, they leave a great deal of work to be done within the wider task of understanding the nature of rational action. A first point to consider would be whether the modified versions of MEU and DP are mutually *consistent*: whether, that is, they would agree on their verdicts in every case (see question 4.14). One would have to go on to ask whether they are *correct*: whether either delivers for all cases a correct account of how it is rational to act. It is unlikely that any such simple principles would be adequate to this task. Indeed, many philosophers are skeptical concerning many

of the notions upon which this discussion has been based. It is not at all plausible to think that the values that are at issue in deciding what to do are measurable in the way that has been presupposed. It would be important to consider whether any substantive principles of rationality can be formulated that do not rest on this supposition. A wider issue is whether we have any right to a supposedly objective, culture-independent notion of rationality as a scale against which any action at all can be measured. Perhaps there are species of rationality, or perhaps rationality is simply one value among others. In the next section, I consider one alleged threat to the coherence of the notion of rationality.

4.14

How would you respond to the following argument?

The dominance principle DP cannot conflict with the MEU principle, if by this is meant that there is a situation in which an action with maximum expected utility would fail to be preferred by the dominance principle. For any upshot, the probability of its occurring is the same regardless of the action, so the only relevant fact, for each upshot, is the utility. So MEU and DP cannot diverge. Table 4.4 makes this plain.

Table 4.4

	P_1	P_2
A_1	5	2
A_2	4	2

A_1 and A_2 are actions open to you. The possible outcomes are P_1 and P_2. If you do A_1 and P_1 is the outcome, your utility is measured by the number 5. Likewise for the other cells in the table. The dominance principle commends A_1 in preference to A_2. The MEU either does likewise or else is indifferent between A_1 and A_2, and in either case the principles do not conflict. To show this, let us call the agent's probabilities of P_1 and P_2 respectively π_1 and π_2. We do not know what these values are, but we can be sure that $5 \times \pi_1$ is greater than $4 \times \pi_2$, and that $2 \times \pi_2$ is not greater than $2 \times \pi_2$. So MEU must either recommend A_1 or else be neutral.

4.2 The Prisoner's Dilemma

You and I have been arrested for drug running and placed in separate cells. Each of us learns, through his own attorney, that the district

attorney has resolved as follows (and we have every reason to trust this information):

(1) If we both remain silent, the district attorney will have to drop the drug-running charge for lack of evidence, and will instead charge us with the much more minor offense of possessing dangerous weapons. We would then each get a year in jail.
(2) If we both confess, we shall both get five years in jail.
(3) If one remains silent and the other confesses, the one who confesses will get off scot-free (for turning state's evidence), and the other will go to jail for ten years.
(4) The other prisoner is also being told all of (1)–(4).

How is it rational to act? We build into the story the following further features:

(5) Each prisoner is concerned only with getting the smallest sentence for himself.
(6) Neither has any information about the likely behavior of the other, except that (5) holds of him and that he is a rational agent.

There is an obvious line of reasoning in favor of confessing. It is simply that whatever you do, I shall do better to confess. For if you remain silent and I confess, I shall get what I most want, no sentence at all; whereas if you confess, then I shall do much better by confessing too (five years) than by remaining silent (ten years). We can represent the situation by table 4.5, and the reasoning in favor of confessing is the familiar dominance principle (DP).

In table 4.5 <0, 10> represents the fact that on this option I go to prison for zero years, and you go for ten years; and so on. The smaller the number on my side of the pair (the left side), the better I am pleased. It is easy to see that confessing dominates silence: confessing, as compared to silence, saves me five needless years if you confess, and one if you do not.

Since you and I are in relevantly similar positions, and, by (6), we are both rational, presumably we shall reason in the same way, and thus perform the same action. So if it is rational for me to confess, it is rational for you to do likewise; but then we shall each go to prison for five years. If

Table 4.5 *The Prisoner's Dilemma.*

	you confess	you don't confess
I confess	<5,5>	<0,10>
I don't confess	<10,0>	<1,1>

we both remain silent, we would go to prison for only one year each. By acting supposedly rationally, we shall, it seems, secure for ourselves an outcome that is worse for both of us than what we could achieve.

On this view, rational action in some circumstances leads to worse outcomes than other courses of action. Even if this is depressing, it is not as it stands paradoxical: we all know that irrational gambles can succeed. What is arguably paradoxical is that the case is one in which the failure of rationality to produce the best results is not a matter of some chance intervention, but is a predictable and inevitable consequence of so-called rational reasoning. How, in that case, can it be rational to be "rational"? (Compare: how could it be rational to be a two-boxer, if being a one-boxer would ensure that one would be a millionaire?) The allegedly unacceptable consequence of the apparently acceptable reasoning is that rational action can be seen in advance to make a worse outcome highly likely.

If this is a paradox, then the correct response, I believe, is to deny that the consequence is really unacceptable. The unacceptability is supposed to consist in the fact that if we were both to act in a certain way, we would be better off than if each were to follow the supposed dictates of rationality. Hence rationality is not the *best* guide to how to act, in that acting in the other way would lead to a better outcome for both. The trouble with this suggestion is that any guide to action has to be available to the agent's decision-making processes. To be guided by the thought that we would both be better off if both remained silent than if both confessed, I would need to know that you would remain silent. What it is rational to do must be relative to what we know. If we are ignorant, then of course acting rationally may not lead us to the best upshot. Here, the ignorance of each concerns what the other will do; and this, rather than some defect in rationality, is what yields the less than optimal upshot.

However, we are now close to a paradox of a different sort. I have said that there is a compelling argument for the rationality of confessing. Yet it appears that there is also a strong case for the rationality of remaining silent. If this case is good, then two apparently acceptable arguments lead to conclusions that, taken together, are unacceptable.

The argument for silence goes like this. We both know we are rational agents, because that is built into the story. We therefore know that any reason one of us has for acting will apply to the other. Hence we know that we shall do the same thing. There are two courses of action that count as doing the same thing: both confessing or both remaining silent. Of these, the latter is preferable for both of us. So of the available courses of action, it is obvious which each of us must rationally prefer: remaining silent.

This argument invites a revenge. Suppose silence is the rational choice, and that I know it is. Then, in knowing that you are rational, I know that it

is the choice you will make. So I know that you will remain silent. However, in that case it must be rational for me to confess, thus securing my preferred outcome: getting off scot-free. On the other hand, I know that you can reason like this too; therefore, if you are rational, you will not keep silent. In that case it is again – but more urgently – rational for me to confess. The way of silence is unstable. Thus the hypothesis that keeping silent is the rational choice is refuted.

The previous paragraph undermines the argument for the claim that silence is the rational policy, and thus removes the threatened paradox that both silence and confession are the rational policy. We should not, however, be content with simply showing this: we should ask how the case connects with the principles of rational action already discussed.

The MEU principle, as stated in section 4.1, claimed that the rational act was whatever maximized expected utility, where this was to be understood in terms of two factors: the desirability of a certain outcome and its probability, conditional upon the performance of a given act. In connection with Newcomb's paradox, I originally said that the relevant probabilities were

(a) the probability of there being \$1,000,000 in box *B, given that* I choose to open both boxes, and
(b) the probability of there being \$1,000,000 in box *B, given that* I choose to open just box *B*.

In the discussion, I claimed that these are not the right probabilities to consider if they do not reflect the tendency of my action to *produce* the outcome in question. So what are the right probabilities to consider? One suggestion is that they are

(a′) the probability of my bringing it about that there will be \$1,000,000 in box *B* by choosing both boxes, and
(b′) the probability of my bringing it about that there will be \$1,000,000 in box *B* by choosing one box.

Assuming that there is no backward causation, both of these probabilities are 0. If we stipulate that anything known to be true anyway (having a probability of 1, regardless of my action) is something that *anything* I do counts as bringing about, then opening box *A* has an expected utility equal to the utility of \$1,000, and opening only box *B* has an expected utility of 0. The version of MEU that considers the probabilities (a′) and (b′), rather than (a) and (b), supports two-boxing.

Let us apply the contrast between these two versions of MEU to the Prisoner's Dilemma. The probability of you confessing, given that I confess, is high and equal to the probability of you remaining silent, given that I remain silent. Moreover, the probability of you confessing, given that I

remain silent, is low, and so is the probability of the converse. These conditional probabilities follow from my knowledge, built in to the example, that you and I will reason in similar ways, since we are both rational (see question 4.15). With suitable utilities, there will be versions of the dilemma in which MEU, in its original form, designates silence as the rational course of action (see question 4.16).

4.15

Clarify the assumption behind the remark that if two people are rational, then for any problem both will reason about it in similar ways. Is the assumption justifiable?

4.16

Using the numbers in table 4.5 as the utilities of the various outcomes (prefix each with a minus sign to show that the outcomes are mostly undesirable), how could you assign conditional probabilities (using numbers between 0 and 1) in such a way as to make the expected utility of silence higher than that of confession? Your assignment will justify the sentence in the text.

In its modified form, MEU was to take into account the probability of an action *producing* the relevant outcome. Since that probability is by stipulation 0 in the present case, because each of us makes his decision before knowing what the other has decided, the modified MEU does not give us any guidance: all the expected utilities, understood in this way, are the same. However, this fact would be a telling reason for applying DP: if you have no idea what outcomes your actions will bring about, choose that action that will make things better for you whatever the other person does.

It has been suggested that Newcomb's paradox is simply a version of the Prisoner's Dilemma. In Newcomb's paradox, the crucial matter – whether or not there is anything in box B – is one of match: the money is in the box if and only if my action matches the prediction. It does not matter whether the prediction occurs before or after the act of choice; what matters is that the act of choice should have no effect on the content of the prediction. Likewise, match is of the essence in the Prisoner's Dilemma: knowing that we are both rational, I expect my action to match yours, just as I expect the prediction to match my choice. Moreover, just as I cannot affect the prediction, so I cannot affect your choice. Table 4.6 sets out the similarities.

I have to choose whether to do X (confess, take both boxes) or Y (remain silent, take just one box). The column indicates what the other

Table 4.6 *Similarities between Newcomb's paradox and the Prisoner's Dilemma.*

		silent/one box	confess/two box
X	confess/two box	1	3
Y	silent/one box	2	4

character in the story may do: the other prisoner remains silent or confesses; the Predictor predicts that I shall one-box or else that I shall two-box. My preferences among the outcomes, from best to worst, are in this order: 1, 2, 3, 4. The "matching possibilities" are shaded: the other prisoner does what I do, the Predictor predicts correctly. I know that my choice cannot affect whether or not a match will occur. I know that a match is much more likely than a mismatch.

In a nutshell, the two arguments we have considered are these:

(A) Do *X*, since you are better off, whatever the other does, than you would be if you were to do *Y*: 1 is better than 2, and 3 is better than 4.
(B) Do *Y*, since match is the most likely kind of outcome, and of these 2 is better than 3.

If this analogy is correct, then one has *consistent* views on the problems only if one is either a two-boxer and a believer in confessing, or else a one-boxer and a believer in silence. My view is the first.

The Prisoner's Dilemma is a simplified version of a well-known conflict: if cooperating means forgoing something that one would otherwise have preferred, then cooperation appears not to be in one's best interests. What serves my purposes best is to secure cooperation from you, while not being cooperative in return. In the Prisoner's Dilemma, what would be best for me is that you remain silent, perhaps under the influence of persuasion, threats, or promises from me, while I, perhaps reneging on undertakings to you, confess. If the first view is correct, and *X*-ing is the rational thing to do, then if we both pursue our interests rationally, we shall end up serving these interests less well than we might. This is not really unacceptable, for it is true; but it may seem depressing.

In the case we have considered, there is just a single situation requiring a decision. Suppose instead that we are confronted with a "multiple Prisoner's Dilemma": suppose that there are a series of choices, and that we each know this – in particular, we each know that this is not the last time we shall be playing the game with each other – and that we also know that the other will remember, and no doubt be guided by, what has happened on previous occasions. There is an argument to the effect that

this new situation would push me in the direction of silence. Suppose you get the idea that I am the sort of person who generally confesses. Then I know that this will make you confess too, to protect yourself from the disastrous consequences of silence, and the overall result will be less than the best for me, time after time. So I have an interest in getting you to believe that I am the sort of person who generally remains silent. One way I can propagate this view is by in fact remaining silent. (We all know, from our knowledge of used-car salesmen and politicians, that this is not the only way to try to achieve this kind of effect.) I also know that you will follow the same policy. So in this situation the cooperative policy of silence would appear to be the rational one (see question 4.17).

4.17

* Suppose that all parties know in advance how many times they will be in this situation – fifty times, say. How would you state the case for the view that the most rational policy is always to confess?

There is fascinating evidence that this is not far from the truth. In some computer simulations of Prisoner's Dilemma situations, the following strategy did better than any other: start by remaining silent; thereafter do what the other player did on the previous round. In suitable circumstances, this will lead to a situation of stable cooperation. Since the multiple Prisoner's Dilemma corresponds more closely to more of real life than the single case, it may be that the upshot of the discussion ought not to be so depressing: perhaps rational self-interest is not doomed to lead to a non-optimal outcome.

Suggested reading

Newcomb's paradox (also known as Newcomb's problem) first appeared in print in Nozick (1969). Nozick says that the paradox was invented by Dr William Newcomb of the Livermore Radiation Laboratories in California. As far as I know, Newcomb himself has not written about his paradox. Nozick (1993) makes new contributions to the discussion of this paradox and the general issues that it raises.

The discussions to which I am most indebted are Mackie (1977) and Gibbard and Harper (1978). For a defense of one-boxing see Bar-Hillel and Margalit (1972).

On probability see Ramsey (1926), Jeffrey (1965); and, for wider applications, Kyburg (1961) and Levi (1967).

Hollis and Sugden (1993) provide a good recent overview of many of these issues. See also Campbell and Sowden (1985) for a useful collection of essays and a good editor's introduction.

Systematic studies of rational decision include Jeffrey (1965) and Anand (1993).

A response to Newcomb's paradox not explored in the text is this: the arguments for one-boxing and for two-boxing are equally compelling, but inconsistent, so there can be no such Predictor (compare, there can be no such Barber). See Schlesinger (1974b) and a critical discussion by Benditt and Ross (1976).

For a discussion of what probabilities should properly be considered in rational action, taking causation into account, see Gibbard and Harper (1978).

For a good article on the Prisoner's Dilemma, see Steven Kuhn's (2007) contribution to the *Stanford Encyclopedia*: plato.stanford.edu/entries/prisoner-dilemma/.

The issues raised by the Prisoner's Dilemma connect closely with practical problems: see e.g. Parfit (1984, chs. 2–4).

For the similarity between Newcomb's paradox and the Prisoner's Dilemma see Lewis (1979). For a similar problem, see Selton (1978).

For an account of the results of computer simulation of various strategies for playing multiple Prisoner's Dilemma games, see Axelrod (1984).

5 Believing rationally

This chapter concerns problems about knowledge or rational belief. The first main section, called "Paradoxes of confirmation," is about two paradoxes that might be called "philosophers' paradoxes." Let me explain.

Most of the paradoxes in this book are quite straightforward to state. Seeing what is paradoxical about them does not require any special knowledge – you do not have to be a games theorist or a statistician to see what is paradoxical about Newcomb's paradox or the Prisoner's Dilemma, nor do you have to be a physicist or sportsman to see what is paradoxical about Zeno's paradoxes. By contrast, the paradoxes of confirmation arise, and can only be understood, in the context of a specifically philosophical project. Therefore these paradoxes need some background (section 5.1.1) before being introduced (in sections 5.1.2 and 5.1.3). The background section sets out the nature of the project within which the paradoxes arise.

The last three main sections of the chapter (5.2–5.4) concern the paradox of the Unexpected Examination. Although it is hard to resolve, one form of it is easy enough to state. (More complex forms, discussed in sections 5.3–5.4, involve technicalities: these sections can be omitted without loss of continuity.) This paradox has been used to cast doubt on intuitively natural principles about rational belief and knowledge.

5.1 Paradoxes of confirmation

5.1.1 Background

We all believe that there is a firm distinction between strong, good, or reliable evidence on the one hand, and weak, bad, or unreliable evidence on the other. If a stranger at the racetrack tells you that Wolf-face will win the next race, and you have no other relevant information, you would be a fool to bet heavily on Wolf-face. The evidence that he will win is extremely thin. However, had the trainer given you the same tip, that would have provided you with much stronger evidence. It would be stronger still if you

knew that the trainer was a crook who believed that you were on to him, and if you also knew that he thought a good tip would buy you off.

Most of our actions are guided by scarcely conscious assessments of how good our evidence is for the relevant beliefs. When we choose what film to see or what restaurant to patronize we are often guided by past experience: by whether the director or actors have good track records, or whether the restaurant has produced good food in the past. We are also guided by what other people say: we weigh their testimony, trusting some – good reviewers, or people we know to be good judges of food – more than others. In such everyday cases, our assessment of the quality of the evidence is unreflective: we recognize good judges and bad judges, good signs and bad signs; but we never normally ask ourselves what *constitutes* a good judge or a good sign.

The philosophical project within which the paradoxes of section 5.1 arise is to state general principles determining what counts as good evidence. Such principles sometimes surface outside philosophy departments. In law courts, for example, explicit categorizations of the evidence ("circumstantial," "inadmissible") are used to grade it; and in scientific investigations, in particular those involving certain kinds of numerically structured data, there are elaborate and sophisticated statistical theories bearing on the question of the extent to which data support a hypothesis.

The branch of philosophy in which philosophers have tried to articulate general principles determining the quality of evidence is called "confirmation theory." These attempts have given rise to surprising paradoxes. Understanding them will lead to a better idea of the nature of evidence.

If a body of propositions constitutes *some* evidence (however slight) for a hypothesis, let us say that these propositions *confirm* the hypothesis (see question 5.1). (This philosopher's use of the word "confirm" is weaker than ordinary usage.) From this starting point one might hope to develop an account of what one should believe. For example, one might think that one ought to believe, of all the relevant hypotheses that one can conceive, those best confirmed by all of one's data. Be that as it may, there are problems enough even with the starting point, let alone with what one might develop from it.

5.1

It may be useful to think that for one proposition to confirm another is for the former to raise the probability of the latter, in the sense that if the first is true, the second is more probable than it would have been otherwise.

Can you think of any circumstances in which there might be a divergence between confirmation as probability-raising and confirmation as providing "evidence" (however slight)?

A very natural thought is that the following principle will play some fundamental role in an account of confirmation:

G1: A generalization is confirmed by any of its instances.

Here are some examples of generalizations:

(1) All emeralds are green.
(2) Whenever the price of cocaine falls, its consumption rises.
(3) Everyone I have spoken to this morning thinks that the Democrats will win the next election.
(4) All AIDS victims have such-and-such a chromosome.

G1 asserts that these propositions are confirmed by their instances – that is, respectively, by this, that, or the other emerald being green; by cases in which the price of cocaine falls and its consumption increases; by the fact that I spoke to Mary this morning, and she thinks that the Democrats will win; and by the fact that Frank, who has AIDS, also has this chromosome. G1 does not assert, crazily, that an instance can *establish* a generalization. A single instance can *confirm*, according to G1, but obviously that does not settle the matter. A single instance does not even show that it is rational to believe the hypothesis, let alone that it is true.

I have spoken both of objects (like emeralds) and of facts (like the fact that Frank has AIDS and also this chromosome) as instances of generalizations, and I shall continue to do so. However, on state occasions I shall say that an instance of a generalization is itself a proposition. When the generalization has the form

All *A*s are *B*s,

an *instance* of it is any proposition of the form

This *A* is a *B*.

Thus

This emerald is green

is an instance of

All emeralds are green.

A *counterinstance* of a generalization "All *A*s are *B*s" is a proposition of the form

This *A* is not a *B*.

So

This emerald is not green

is a counterinstance of "All emeralds are green." Just as we may, on non-state occasions, speak of green emeralds as instances of this latter proposition, so we can speak of non-green emeralds as counterinstances of it.

The opposite of confirmation is *disconfirmation*. A hypothesis is disconfirmed by propositions that tend to show it to be false. An extreme case is *falsification*: a generalization is falsified by any counterinstance of it.

The principle G1 is to be understood to mean that any proposition that is an instance of a generalization confirms that generalization. It is not always clear how this is meant to link up with the notion of good evidence. Obviously, one AIDS victim with a certain chromosome does not alone constitute good evidence for the hypothesis that all AIDS victims have it; but perhaps a large number of instances, and no counterinstances, do add up to good evidence. If so, we shall think of each instance as making a positive contribution to this good evidence, and this is what philosophers have in mind by the notion of confirmation. G1 does not say, absurdly, that an instance of a generalization would, in and of itself, give us good reason to believe that generalization. Rather, it says that an instance makes a positive contribution, however slight, and however liable to be outweighed by other factors, toward constituting good evidence. The idea is that if we come to know an instance of a generalization, we have taken one small step toward having good evidence for that generalization, even though other things we know may undermine this evidence. Indeed, our other knowledge might include a counterinstance of that same generalization.

The quality of evidence is a matter of degree. Some evidence is stronger, other evidence weaker. One way we might try to build toward this from the idea of confirmation, together with G1, is by saying that your evidence for a generalization is stronger the more instances of it your total body of knowledge contains, provided that it contains no counterinstances. However, one must beware of supposing that it is at all easy to arrive at a correct account. The following shows that what has just been suggested is indeed wrong. One could well have come across many instances, and no counterinstances, of the generalization

All places fail to contain my spectacles

(one has searched high and low without success); yet, assuming one does indeed have spectacles, one would be quite right to be certain that this generalization is false.

The appeal of G1 comes in part from the thought that *extrapolation* is reasonable. If all the things of kind A that you have examined have also been of kind B, then you have some reason to extrapolate to the hypothesis that all things of kind A are of kind B. Of course, the evidence may be slight, and it may be outweighed by other evidence.

We are not usually interested in confirmation (in the technical sense used here) in cases of generalizations such as "Everyone I met this morning said that the Democrats would win." If I had met a reasonably small number of people, I might say in the afternoon: "I don't need evidence – I already *know* that it's true." The idea is that my own experience already determines the truth of the generalization. The contrast is with generalizations such as "Whenever the price of cocaine falls, its consumption increases." You may know that this has held so far, but this does not settle that the proposition is true, for it speaks to future cases as well as past ones. This is the sort of generalization for which we feel we need evidence: a generalization not all of whose instances one has encountered (see question 5.2).

5.2

There are generalizations of which one could not be sure that one had encountered all the instances. What are some examples?

Inductive reasoning, as philosophers call it, consists in arguing from evidence or data to hypotheses not entailed by these data. One traditional philosophical problem has been to justify this process: to show that it is at least sometimes legitimate to "go beyond the data." Let us call this the problem of *justification*. Another philosophical problem is to give a general account of the kinds of inductive reasoning we *take* to be legitimate (without necessarily pronouncing on whether or not they are really legitimate). Let us call this the problem of *characterization*. We take it that it is legitimate to argue from the fact that the sun has risen every day so far to the conclusion that it will, probably, rise every day in the future; or, at least, to the conclusion that it will, probably, rise tomorrow. By contrast, we do not think that it is legitimate to argue from these same data to the conclusion that the sun will sometime cease to rise, or to the conclusion that it will not rise tomorrow (see question 5.3). The problem of characterization is to give an illuminating general account of the features of evidence that make us count it as *good* evidence, as a legitimate basis for the hypothesis in question.

5.3

Victims of the so-called Monte Carlo fallacy dispute this. They hold that the longer the run of successive reds on a fair roulette wheel the *less* likely it is that red will come up on the next spin. What, if anything, is wrong with this view? Is there anything right about it?

An initial answer to the problem of characterization is that inductive reasoning is generally taken to be legitimate when it is a case of extrapolation: when one reasons on the assumption that what one has not experienced will resemble what one has. G1 is connected with this initial suggestion, for it specifies a way of extrapolating.

These problems of induction are akin to problems already encountered. Earlier, we asked "Under what conditions are data good evidence for a hypothesis?" If we can answer this question in some illuminating way (and not merely by saying, for example, "When they are"), we shall thereby be close to having solved the problem of justification – for we shall then be close to showing that it *is* sometimes legitimate to go beyond the data (see question 5.4). Moreover, if we could answer the question "Under what conditions are data *taken to be* good evidence for a hypothesis?", we would have answered the problem of characterization.

5.4

* How might answering this question fail to show that inductive reasoning is sometimes legitimate, and thus fail to be a complete answer to the problem of justification?

We shall be concerned only with the characterization problem: not the question of whether there is any genuinely legitimate inductive reasoning, but rather the question of what sort of inductive reasoning we (rightly or wrongly) take to be legitimate. Though this seems the easier problem, attempts to answer it lead quickly to contradictions.

5.1.2 The paradox of the Ravens

Despite the initial appeal of G1, it leads, in conjunction with other apparently innocuous principles, to a paradox discovered by Carl Hempel (1945), and now generally known as the paradox of the Ravens.

In order to derive the paradoxical consequence, we need just one other principle:

> **E1**: If two hypotheses can be known *apriori* to be equivalent, then any data that confirm one confirm the other.

Something can be known *apriori* if it can be known without any appeal to experience. For example, one does not have to conduct any kind of social survey to discover that all women are women: indeed, one could not discover this by a survey. What can be known *apriori* can be known simply on the basis of reflection and reasoning.

Two hypotheses are equivalent just on condition that if either one is true, so is the other, and if either one is false, so is the other. E1 asks us to consider cases in which two hypotheses can be known *apriori* to be equivalent. An example would be the hypotheses

R1: All ravens are black

and

There are no ravens that are not black

and also

R2: Everything non-black is a non-raven.

Any two of these three hypotheses are equivalent, and this can be shown simply by reflection, without appeal to experience; so the equivalence can be known *apriori*. For example, suppose R1 is true: all ravens are black. Then, clearly, any non-black thing is not a raven, or, as R2 puts it, is a non-raven. So if R1 is true, so is R2. Now suppose that R1 is false; then some ravens are not black. However, this means that some things that are not black are ravens, so R2 is false, too. Thus R1 and R2 are equivalent, and this can be known *apriori*.

We can now show how the paradox of the Ravens is derived from G1 and E1. By G1, R2 is confirmed by its instances – for example, by a white shoe, or (using the state-occasion notion of an instance) by, for example:

P1: This non-black (in fact, white) thing is a non-raven (in fact, a shoe).

Instance P1 confirms R2, but R2 can be known *apriori* to be equivalent to R1. So, by E1, P1 confirms R1, "All ravens are black." This, on the face of it, is absurd. Data relevant to whether or not all ravens are black must be data about ravens. The color of shoes can have no bearing whatsoever on the matter. Thus G1 and E1 – apparently acceptable principles – lead to the apparently unacceptable conclusion that a white shoe confirms the hypothesis that all ravens are black. This, finally, is our paradox.

The principles of reasoning involved do not appear to be open to challenge, so there are three possible responses:

(a) to say that the apparently paradoxical conclusion is, after all, acceptable;
(b) to deny E1; or
(c) to deny G1.

Hempel himself makes the first of these responses. One could argue for it as follows. First, we must bear in mind that "confirm" is being used in a technical way. It does not follow from the supposition that a white shoe *confirms* that all ravens are black that observing a white shoe puts you in a

position reasonably to believe that all ravens are black. Second, there are cases in which it seems quite natural, or at least much less absurd, to allow that P1 confirms that all ravens are black – that is, that P1 could make a positive contribution to some good evidence for the hypothesis. Suppose that we are on an ornithological field trip. We have seen several black ravens in the trees and formulate the hypothesis that all ravens are black. We then catch sight of something white in a topmost branch. For a moment we tremble for the hypothesis, fearing a counterinstance – fearing, that is, that we have found a white raven. A closer look reveals that it is a shoe. In this situation, we are more likely to agree that a white shoe confirms the hypothesis. Hempel tells a similar story for a more realistic case. Investigating the hypothesis that all sodium salts burn yellow, we come across something that does not burn yellow. When we discover that the object is a lump of ice, we regard the experiment as having confirmed the hypothesis.

The first point appeals to the idea that some complicated story must be told in order to link confirmation to having good reason to believe. Furthermore, in the telling, it will be apparent why observing white shoes, despite their confirmatory character with respect to the hypothesis that all ravens are black, does not normally contribute to giving one good reason to believe the hypothesis. We cannot assess the suggestion until we know the details of this story.

The second of these points emphasizes that confirmation, as we normally think of it, is not an absolute notion but is relative to what background information we possess. Making this point leaves unstarted the task of specifying how the technical notion of confirmation – which, so far, has been taken as absolute – should be modified so as to take account of this relativity.

Perhaps these points can be developed so as to justify the first response, (a); but I shall now turn to the other possible responses.

Response (b) is to deny E1. For example, one might simply insist that anything that confirms a generalization must be an instance of it. This avoids the paradox, but it is very hard to justify. For example, suppose that we are investigating an outbreak of legionnaires' disease. Our hypothesis is that the source of the infection was the water at St. George's school, consumed by all the children who attended last week. Will only an instance of the generalization "All pupils attending St. George's last week contracted legionnaires' disease" confirm it? Imagine that we find some St. George's children who are free from the disease, but that it then turns out that they missed school last week. We would normally count this as evidence in favor of our hypothesis – some potential and highly relevant counterinstances have been eliminated – and yet these children are not instances of the hypothesis.

There is a more general argument against the rejection of E1. Suppose we find some data that confirm two hypotheses, H1 and H2. It is standard practice to reason as follows: H3 is a consequence of H1 and H2, so to the extent that H1 and H2 are confirmed, so is H3. For example, if we had data that confirmed both the hypothesis that all anorexics are zinc-deficient and the hypothesis that everyone who is zinc-deficient is zinc-intolerant, the data would surely confirm the hypothesis that all anorexics are zinc-intolerant. However, if we allow that data confirm the *apriori* knowable consequences of hypotheses they confirm, we have in effect allowed E1 (see question 5.5).

5.5

This argument for E1 presupposes the principle that anything that confirms a hypothesis confirms its consequences. As John MacFarlane pointed out (personal communication), this is open to counterexample. Consider:

A = The card on the table is a face card.
B = The card on the table is a red jack.
C = The card on the table is red.

(C) is an *apriori* consequence of (B), and (A) confirms (B), at least in the sense of raising its probability. But (A) does not confirm (C).

Is the counterexample decisive? If not, explain why not. If so, is there any better argument for E1?

The third possible response to the paradox is to reject G1. This is both the most popular response, and also, I believe, the correct one. The paradox of the Ravens already gives us some reason to reject it, if the other responses are unsatisfactory. The paradox of "grue," to be considered in the next section, gives a decisive reason for rejecting it. Moreover, there are quite straightforward counterexamples to it. Consider, for example, the hypothesis that all snakes inhabit regions other than Ireland. According to G1, a snake found outside Ireland confirms the hypothesis; but however we pile up the instances, we get no evidence for the hypothesis. Quite the contrary: the more widespread we find the distribution of snakes to be, the more unlikely it becomes that Ireland is snake-free. A non-Irish snake does not confirm the hypothesis, since it makes no positive contribution to the evidence in favor of the hypothesis, and may even count against it.

Rejecting G1 resolves the paradox, but it leaves us in a rather unsatisfactory position regarding confirmation. We have made very little progress toward uncovering the principles that underlie our discrimination between good and bad evidence. The next paradox brings to light more difficulties in the path of this project.

5.1.3 "Grue"

According to G1, green emeralds confirm the hypothesis that all emeralds are green. Now consider the predicate "grue," invented by Nelson Goodman (1955) with an eye to showing the inadequacy of G1. The meaning of "grue" is stipulated so as to ensure that a thing *x* counts as *grue* if and only if it meets either of the following conditions:

Gr1: *x* is green and has been examined, or
Gr2: *x* is blue and has not been examined.

The class of grue things is thus, by definition, made up of just the examined green things together with the unexamined blue things. All examined emeralds, being all of them green, count as grue, by Gr1. It follows from G1 that the hypothesis that all emeralds are grue is confirmed by our data: every emerald we have examined is a confirming instance because it was green. This is absurd. If the hypothesis that all emeralds are grue is true, then unexamined emeralds (supposing that there are any) are blue. This we all believe is false, and certainly not confirmed by our data. G1 must be rejected.

What is paradoxical is that a seeming truth, G1, leads, by apparently correct reasoning, to a seeming falsehood: that our data concerning emeralds confirm the hypothesis that they are all grue (see question 5.6). The paradox relates to the problem of *characterization* – of saying what kinds of evidence we take to be good, or what sorts of inductive argument we take to be legitimate – because we need to say what makes us treat green and grue differently. G1 does not discriminate between the cases.

5.6

* An alternative presentation of the paradox identifies the apparently unacceptable conclusion as being that the same body of data can confirm the *inconsistent* hypotheses that all emeralds are green and that all emeralds are grue. Is it unthinkable that a body of data should confirm inconsistent hypotheses?

The conclusion is unacceptable even if we recall that "confirms" is being used in a technical sense: it is not equivalent to "gives us good reason to believe," but means only something like "would make a positive contribution to a good reason for believing." It strikes us as unacceptable to suppose that an examined green emerald makes any contribution at all to giving a good reason for supposing that all emeralds are grue.

We have already seen in connection with the Ravens (section 5.1.2) that there is a case for rejecting G1; that case is strengthened by the present

Grue paradox. If we reject G1, then both paradoxes are, for the moment, resolved, for we shall have said that an apparently acceptable premise is not really acceptable. What can we put in its place? It would seem that something like G1 must be true. Is there an appropriate modification? If not, then the Grue paradox remains not fully resolved; for to say that there is no appropriate modification of G1 is to say that there are no principles governing what makes a body of data confirm a hypothesis. This seems as unacceptable as the view that green emeralds confirm the hypothesis that all emeralds are grue.

Most of the suggested responses can be seen as falling into one of two patterns:

(1) The blame is placed on the word "grue," which is said to be of a particularly nasty or "pathological" or gruesome kind, rather than on the structure of G1. All we need is a general principle for excluding the *grue*-like words, and G1 will be acceptable for the remainder.
(2) The blame is placed not so much on "grue" as on the attempt to formulate a principle, like G1, that takes no account at all of *background information* – information that is always in play in any real-life case of evidence or confirmation.

If we try the first response, the difficulty is to say exactly what is nasty about "grue." It is not enough to say that "grue" is an invented word, rather than one that occurs naturally in our language. Scientists often have to invent words (like "electron") or use old words in new ways (like "mass"), but it would be extravagant to infer that these new or newly used words cannot figure in confirmable generalizations.

It is more appealing to say that what is wrong with "grue" is that it implicitly mentions a specific time in its definition. Its definition appeals to what has *already* been examined, and this refers to the time at which the definition is made. In this respect, "grue" and "green" differ sharply, for there is no *verbal* definition of "green" at all, and even if there were it would certainly not involve reference to a particular time.

If we were to restrict G1 to generalizations in which there is no reference to a time, it would be too restrictive. For example, the generalization "In Tudor times, most agricultural innovations were made in the north of the country" is one that could be confirmed or disconfirmed on the pattern of G1. In addition, G1 would not be restrictive enough. The structure of the Grue paradox is preserved if we can find a way of picking out just the emeralds we have already examined. We might do this by giving each one a name, e_1, e_2, … ; or it might be that all and only the emeralds so far examined have come from a certain emerald mine (now exhausted); or something of the kind. Then we could define a predicate equivalent to

"grue" without mentioning a time. We could say that it is to apply to any of e_1, e_2, … just on condition that that thing is green, and to anything else just in case it is blue; or we could say that it is to apply to everything taken from a certain mine just on condition that it is green, and to anything else just on condition that it is blue. It is not of the essence of the paradox that the definition of "grue" mentions a time.

There are other ways of trying to say what is nasty about "grue." Goodman's own attempt has at least superficial similarities with one I rejected earlier. He says that what is wrong with "grue" is that it is not "well-entrenched"; that is, the class of entities to which it applies is a class that has not been alluded to much – indeed, at all – in the making of predictions (see Goodman 1955, esp. pp. 97ff.). To treat being poorly entrenched as sufficient for being incapable of figuring in confirmable generalizations seems to put an intolerable block on scientific innovativeness. Though Goodman is well aware of this problem, and provides a more sophisticated response than my brief description would suggest, there is room for doubt about whether he deals with it successfully.

I now want to consider a response of the other kind I mentioned: not restricting G1 by limiting it to generalizations that do not contain words sharing the supposed nasty features of "grue," whatever these features may be; but rather restricting G1 by appeal to background information. Intuitively, what is wrong with supposing that our information about examined emeralds gives us any reason for thinking that all emeralds are grue is that we know that the examined ones are grue only by virtue of having been examined. We do not believe that our examining the emeralds had any "real" effect on them. We believe that if they had not been examined they would not have been grue. What makes it so absurd to suppose, on the basis of our data, that all emeralds are grue is that we know that the unexamined ones lack the property by virtue of which the examined ones *are* grue: namely, having been examined. An initial attempt to formulate this thought might look like this:

G2: A hypothesis "All *F*s are *G*s" is confirmed by its instances if and only if there is no property *H* such that the *F*s in the data are *H*, and if they had not been *H*, they would not have been *G*.

We might try to support G2 by the following case, which in some respects looks similar to the grue emeralds. Suppose that we are gathering evidence about the color of lobsters, but unfortunately we have access only to boiled ones. All the lobsters in our sample are pink. Moreover, we know that the lobsters in the sample are pink only by virtue of having been boiled. Then it would be absurd for us to think that our sample confirms the hypothesis that all lobsters are pink. Here the hypothesis is "All lobsters (*F*) are pink (*G*)," and *H* is the property of having been boiled. Because the lobsters in

the sample are boiled, and had they not been boiled would not have been pink, the data do not meet the condition imposed by G2 for confirming the hypothesis (Jackson 1975).

The lobster case brings to light a difficulty, or series of difficulties, connected with G2. We start to uncover them if we ask: how do we know that the lobsters in the sample would not have been pink had they not been boiled? It would seem that if we know this, then we know that some lobsters are not pink at all times, and thus we are in a position to know that the hypothesis is false.

This shows that we can explain, without appealing to G2, why the evidence for the hypothesis that all lobsters are pink was deficient. A body of evidence fails to confirm any hypothesis to which it contains a counterinstance. In addition, the case brings to light something more fundamental: G2, as it stands, does not require our *body of data* to contain the proposition that there is no *H* such that the examined *F*s would not have been *G* had they not been *H*. It requires only that this proposition be true. What would be relevant to G2 would thus be a variant of the lobster case in which all observed lobsters are pink, but we, the observers, do not realize that they are pink only because they have been boiled. G2 rules that, in this state of ignorance, our data do not confirm the generalization that all lobsters are pink. Is this acceptable?

This raises an important issue. If it sounds wrong to say that the person who has observed only pink lobsters, and who knows nothing of the connection between boiling and color (and perhaps does not even know that the sample lobsters have been boiled), lacks data that would confirm the hypothesis that all lobsters are pink, this is because we intuitively feel that evidence should be *transparent*. By this I mean that we intuitively feel that if a body of data is evidence for a hypothesis, then we ought to be able to tell that this is so merely by examining the data and the hypothesis: one ought, in other words, to be able to tell that this is so *apriori*. This intuitive feeling might be supported by the following argument. Suppose that no evidence is, in this sense, transparent. Then, a claim to the effect that a body of data *D* confirms a hypothesis *H* will itself be a hypothesis needing confirmation. We shall need to cast around for data to confirm, or disconfirm, the hypothesis that *D* confirms *H*. It looks as if we are set on an infinite regress, and that we could never have any reason to suppose that anything confirms anything unless evidence is transparent.

Not all evidence is transparent. Spots can confirm the hypothesis that the patient has measles, but one needs medical knowledge to recognize that the data, the spots, are thus related to the hypothesis: one needs to know that only people, or most people, with spots of this kind have measles. The most the argument of the preceding paragraph could show

is that *some* evidence needs to be transparent, since this is all that is needed to block the alleged regress.

If we feel that some evidence should be transparent, we shall surely feel that an example is the limiting case in which everything that *can* be included among the data *has* been included. In this case, we shall feel that one ought to be able to tell *apriori*, without further investigation, which hypotheses these data confirm. However, this is not guaranteed by G2, for two reasons.

First, for some hypotheses of the form "All *F*s are *G*s," our data may include plenty of instances and no counterinstances but fail to contain either the proposition "There is no *H* such that all examined *F*s are *H* and would not have been *G* had they not been *H*" or its negation. In this case, G2 would not allow us to tell *apriori* whether our data confirm the hypothesis, since we could not tell whether the condition it places on the instances of the hypothesis obtains or not.

Second, it is a debatable question whether this condition *could* properly be included among our data. One might hold that all data must, in the end, be observations, and that a condition such as "There is no *H* such that all examined *F*s are *H* and would not have been *G* had they not been *H*" is not immediately available to observation, and so cannot be a datum.

These objections point in controversial directions. The second objection presupposes a form of *foundationalism*, which is highly controversial. Perhaps, contrary to the presupposition, there is nothing in the intrinsic nature of a proposition that qualifies it as a datum or as a non-datum; thus, on occasion, the counterfactual condition could count as a datum. If this is allowed, then we could envisage a variant of G2 that meets the first of the two objections.

G3: A hypothesis "All *F*s are *G*s" is confirmed by a body of data containing its instances, and containing no counterinstances, if and only if the data do not say, of some property *H*, that the *F*s in the data are *H*, and if they had not been *H* they would not have been *G*.

This rules that "All emeralds are grue" is not confirmed by its instances, for our data do say, concerning the property *being examined*, that the emeralds in the data have been examined, and had they not been examined they would not have been grue. It has the further merit of being consistent with transparency: whether or not a body of data confirms a hypothesis depends only on the data and the hypothesis themselves, and not on other, perhaps inaccessible, facts. However, it has the apparently dubious feature that smaller bodies of data can confirm more than can larger bodies.

To see how this works, imagine two people, both confronted with pink boiled lobsters, and both concerned to consider the question of whether

their data confirm "All lobsters are pink." One person does not realize that all the lobsters he has seen have been boiled, or else does not realize that boiling them affects their color. If G3 is correct, that person's data do confirm the hypothesis "All lobsters are pink." The other person, by contrast, knows that the lobsters would not have been pink had they not been boiled. G3 entails that that person's data do not confirm the hypothesis that all lobsters are pink. If you know more, your data may confirm less.

This feature is one that should come as no surprise. A body of data full of instances of a generalization, and containing no counterinstances, may confirm the generalization, though the same body enriched by a counterinstance would not. Still, G3 needs refinement. For one thing, it still leads, in conjunction with E1, to the Ravens paradox (see question 5.7). For another thing, we need to modify it somewhat in the light of examples like the following.

5.7

Sketch a proof of the Ravens paradox, using G3 rather than G1.

Suppose you find, year after year, that although all the other vegetables in your garden are attacked by pests, your leeks are always pest-free. Would it be reasonable to conclude that leeks are immune to pests? Let us suppose that you know no proposition to the effect that your leeks would not have been pest-free had they not possessed some property *P*. According to G3, the hypothesis that all leeks are immune to pests is confirmed by your data; but I think that we should not, in fact, put much confidence in the hypothesis, given the data. Even if one knows no proposition of the relevant kind, one may strongly suspect that *there is* one, even though one does not know it. One knows in a general way that susceptibility to pests is likely to be affected by such factors as the nature of the soil, how strongly the plant grows, and what other vegetation is around. Even though your data do not include a proposition that selects a factor that explains the pest-free quality of your leeks, you might well believe that *there is* a proposition of this kind. If so, you should not put much faith in the hypothesis that all leeks, including those grown in very different conditions, are immune to pests.

If it is to deliver the results we want in such cases, G3 would need to be revised, so that:

G4: A hypothesis "All *F*s are *G*s" is confirmed by a body of data containing its instances, and containing no counterinstances, if and

> only if the data do not say that *there is*, or even that *there is quite likely to be*, a property, *H*, such that the examined *F*s are *G* only in virtue of being *H*.

All other things being equal, the fact that we think it quite likely that there are conditions under which leeks suffer from pests is enough to diminish, or even perhaps cancel, the confirmatory impact of our pest-free leeks; G4 is intended to do justice to this fact.

G4 entails that the hypothesis that all emeralds are grue *is* confirmed by the data consisting just of propositions of the form "This is a green emerald," "This has been examined," and so on; but it does not entail that this hypothesis is confirmed by the body of data, including background information, that we in fact possess. That body of data includes the proposition that the examined emeralds would not have been grue had they not been examined; so it entails that there is a property *(having been examined)* such that the emeralds in the data are grue only in virtue of possessing that property.

Does G4 rule out enough (see question 5.8)? One might in particular have doubts about whether it should allow that the grue hypothesis is confirmed by the narrower body of data. These doubts might be to some extent assuaged by reflecting that a body of data can confirm a hypothesis that they do not make it rational to believe. We could insist that sometimes instances confirm, in the sense of making a positive contribution to good grounds for belief, while not on their own constituting such grounds.

5.8

What are the advantages (if any) and disadvantages (if any) of the following variant (call it G5)?

> A hypothesis "All Fs are Gs" is confirmed by a body of data containing its instances, and containing no counterinstances, if and only if the data say that no property, *H*, is such that the *F*s in the data are *H*, and if they had not been *H* they would not have been *G*.

The following example is designed to put this view to the test. We have to try to imagine a case in which we have just the instances of a generalization, and absolutely no relevant background information at all. Suppose you come across a very large sack of marbles. You cannot see into the sack, but you manage to take out one marble at a time. You do this for a while, and all those you take out are green. By G4, the hypothesis that all the marbles are green, including the unseen ones still in the sack, is confirmed. However, I claim that you still do not have good reason to

believe that all the marbles in the sack are green. The case is, by stipulation, one in which you are not allowed to bring to bear any background information in the form of suppositions about how the marbles came to be in the sack. The hypothesis that they were put there by a collector of marbles must be, for you, no more likely than that they were put there by a philosopher wanting to make a point about confirmation theory. Moreover, you can form no reasonable belief about whether you are, or are not, selecting the marbles "at random." Perhaps the only way to extract a marble is to press a lever at the side of the sack. You have no idea whether they are being dealt from the top or in some other order, or whether the mechanism selects them in some genuinely indeterministic fashion. Under these circumstances, it seems to me quite doubtful whether the run of green marbles gives us good grounds for believing that all the marbles in the sack are green. In particular, there does not seem to be much difference in the justification of the view that they are all green and the view that they are all grue. It must be stressed that such situations are bound to be rare. Perhaps we have to imagine ourselves on an alien planet, governed by unknown physical laws, to ensure that we are really not bringing background information to bear, as we normally would.

We perhaps incline to think that if *all* we knew about emeralds consisted in a large number of green samples, it is not merely that the hypothesis that all emeralds are green would be confirmed by our data: in addition, we would be justified in believing it. This, I think, is an illusion. In forming this view, I think we unconsciously bring to bear background information concerning the color constancy of many gemstones, together with the supposition that emeralds are gemstones. Take this supposition away, and there can be no move from confirmation to justified belief. One can see this by comparing the case in which all the tomatoes I have ever come across are green. This would give me good reason to believe that all tomatoes are green only if I had good reason to think that the sample tomatoes were somehow typical; but of course I can have no such reason (see question 5.9).

5.9

How, if at all, should G4 be modified in the light of this apparent counterexample?

> "All water boils at 100°C (at sea level)" is confirmed by a body of data containing its instances, and containing no counterinstances, if and only if the data do not say that the examined samples of water boil at 100°C only by virtue of being composed of H_2O. (From a counterexample proposed by John Heil.)

The Grue paradox has been held to have more distant ramifications. To gesture toward these, let us consider a corollary that Goodman stresses:

Regularities are where you find them, and you can find them anywhere.

The old idea – found, for example, in David Hume in the eighteenth century – was that to reason from experience in a way we take to be legitimate is to extrapolate regularities obtaining within our experience. One thing that Goodman's "grue" shows is that this is, at best, a highly incomplete account, for it does not answer the question: what is a regularity? The regular connection between being an emerald and being green? *And* the regular connection between being an emerald and being grue? Our original problem reemerges in this form: either we can give no account of what a regularity is, in which case it is useless to describe our inductive practice as extrapolating experienced regularities; or else we give an account of regularity that includes the undesirable emerald–grue regularity as well as the desirable emerald–green one.

This relatively narrow point about confirmation suggests a deeper metaphysical one: whether a series of events counts as a regularity depends upon how we choose to describe it. This has suggested to some a quite thoroughgoing conventionalism, according to which there is no separating how the world is in itself from the conventions we bring to bear in describing and classifying it. To others, it has had the effect of deepening their skepticism about the legitimacy of inductive reasoning. If there are endless regularities that we could have extrapolated, what makes it rational to pick on the ones we in fact do? It is bad enough having to justify extrapolating a regularity, but it is worse when one must, in addition, justify selecting one rather than any of the countless other regularities in the data to extrapolate.

To yet others, the Grue paradox has suggested that there is something quite undetermined, at least at the individual level, about our concepts. Wittgenstein asked us to consider someone who, having added 2 to numbers all the way up to 1,000, continues this way – 1,004, 1,008, … – and yet protests that he is "going on in the same way." We could define a gruelike operator "$+^\star$" as follows: $x +^\star 2 = x + 2$, if $x < 1{,}000$; otherwise $x +^\star 2 = x + 4$. It has been suggested that there are no facts, or at least no individual facts, that make it true of us that we use concepts like *green* and $+$ rather than concepts like *grue* and $+^\star$.

The impact of grue goes well beyond the problems of finding a non-paradoxical account of our notion of confirmation.

5.2 The Unexpected Examination

The teacher tells the class that sometime during the next week she will give an examination. She will not say on which day, for, she says, it is to be a

surprise. On the face of it, there is no reason why the teacher, despite having made this announcement, should not be able to do exactly what she has announced: give the class an unexpected examination. It will not be totally unexpected, since the class will know, or at least have good reason to believe, that it will occur sometime during the next week. However, surely it could be a surprise, or unexpected, in this sense: that on the morning of the day on which it is given, the class will have no good reason to believe that it will occur on *that* day, even though they know, or have good reason to believe, the teacher's announcement. Cannot the teacher achieve this aim by, say, giving the examination on Wednesday?

The class reasons as follows. Let us suppose that the teacher will carry out her threat, in both its parts: that is, she will give an examination, and it will be unexpected. Then the teacher cannot give the examination on Friday (assuming this to be the last possible day of the week); for, by the time Friday morning arrives, and we know that all the previous days have been examination-free, we would have every reason to expect the examination to occur on Friday. So leaving the examination until Friday is inconsistent with giving an *unexpected* examination. For similar reasons, the examination cannot be held on Thursday. Given our previous conclusion that it cannot be delayed until Friday, we would know, when Thursday morning came, and the previous days had been examination-free, that it would have to be held on Thursday. So if it were held on Thursday, it would not be unexpected. Thus it cannot be held on Thursday. Similar reasoning supposedly shows that there is no day of the week on which it can be held, and so supposedly shows that the supposition that the teacher can carry out her threat must be rejected. This is paradoxical, for it seems plain that the teacher *can* carry out her threat.

Something must be wrong with the way in which the class reasoned; but what?

The class's argument falls into two parts: one applies to whether there can be an unexpected examination on the last day, Friday; the other takes forward the negative conclusion on this issue, and purports to extend it to the other days.

Let us begin by looking more closely at the first part. On Friday morning, the possibilities can be divided up as follows:

(a) The examination will take place on Friday and the class will expect this.
(b) The examination will take place on Friday and the class will not expect this.
(c) The examination will not take place on Friday and the class will expect this.
(d) The examination will not take place on Friday and the class will not expect this.

When we speak of the class's expectations, we mean their rational or well-grounded ones. It is not to the point that they may have expectations to which they are not entitled, or lack expectations to which they are entitled. For example, it is not to the point that the class might irrationally (without entitlement or justification) believe that the examination would take place on Wednesday. Even if it then did take place on Wednesday, this would not show that the teacher's announcement was false, for what she said, as we have interpreted it, was that the class would have *no good reason* to believe that it would occur when it did.

The overall structure of the class's argument is meant to be a *reductio ad absurdum*: they take as a supposition that the teacher's announcement is true, then aim to show that this leads to a contradiction, and hence that the supposition must be rejected. In this first part of the argument, the supposition is used to show that the examination cannot occur on Friday. This is extended to every day of the week in the second part of the argument, so that, in the end, the supposition is rejected. Thus the teacher's announcement is disproved.

Given that the examination has not occurred on the previous days, at most possibility (b) is consistent with the truth of the teacher's announcement. The class's argument aims to show that (b) is not a real possibility.

The idea is that the class can infer that if the examination occurs on Friday, then the announcement is false, contrary to the supposition. The inference is based on the consideration that the class will know that Friday is the last possible day for the examination. So, given the supposition that the teacher's announcement is true, they would expect the examination, were it to occur on Friday; but this is inconsistent with the truth of the announcement. If we hold on to the supposition, the examination cannot take place on Friday.

This argument is not straightforward, as can be brought out in the following way. Imagine yourself in the class, and it is Friday morning. There is surely a real question, which you may well feel that you do not know how to answer: has the teacher forgotten or changed her mind, or will the examination indeed take place today? It would seem that this doubt is enough to ensure that if it does take place that day, it will be unexpected: the class was not entitled to expect it.

The class's argument is meant to circumvent this difficulty by using the truth of the teacher's announcement as a supposition – one that, in the end, is going to be rejected. Given this supposition, the class on Friday morning can rule out the non-occurrence of the examination. On the other hand, it can also rule out its occurrence – and this is what is meant to show that, if the supposition is true, the examination cannot occur on Friday.

However, it is a mistake to think that the supposition merely of the *truth* of the teacher's announcement will do the required work. To see this, imagine ourselves once more among the class on Friday morning. Suppose that the teacher's announcement is true but that we do not know or even believe this. Then we may not believe that the examination will occur. This is enough to make the truth of the announcement possible: if the examination does occur, we shall not have expected it. This shows a fallacy in the reasoning as so far presented. Merely supposing, for *reductio*, that the teacher's announcement is true is not enough to establish that the examination will not be held on Friday. At that point in the argument, we need as a supposition that we *know* that the teacher's announcement is true (Quine 1953).

If we are to have a paradoxical argument worth discussing, we need to make some changes. The details are quite complicated: to make the discussion manageable, we shall soon need to use some abbreviations. Those whose distaste for symbols exceeds their thirst for knowledge might prefer to step off the ride at this point.

5.3 Revising the Unexpected Examination

One modification we could make is to leave the announcement unchanged but alter the structure of the argument. Instead of taking the announcement itself as our supposition, we shall suppose that the class *knows* the truth of the announcement. This supposition is refutable, on Friday, by the considerations outlined. If on Friday we know that the announcement is true, we know that the examination will occur on Friday. If we know that the examination will occur on Friday, the announcement is not true. If the announcement is not true, then we do not know that it is true. The supposition that we know that it is true entails its own falsehood, and so can be rejected. Applying similar reasoning to the other days of the week, the upshot would be that the class can show that it cannot *know* that the announcement is true. This may seem paradoxical: intuitively, we want to say that we knew, from the announcement, that there would be an examination sometime, though we did not know when, and so it was unexpected.

An alternative modification involves changing the announcement to include the fact that the class will not know, on the basis of the announcement, that the examination will take place on the day that it does. In a way that can only be made clear by some abbreviations, this will give us a valid argument for the conclusion that the announcement is false. If this is paradoxical, it is because it seems intuitively obvious that such an announcement could be true.

Let us call the original version of the argument OV, the first proposed modified version MV1, and the second proposed modified version MV2. Since the number of days of the week is irrelevant, let us simplify by supposing that there are just two possible examination days, Monday or Tuesday. For OV and MV1, I shall abbreviate the announcement as:

> **A1**: I shall give you an examination on either Monday or Tuesday, and you will not know – or have good reason to believe – on the morning of the examination that it will occur that day.

The other abbreviations are as follows:

M for "the examination occurs on Monday";
T for "the examination occurs on Tuesday";
$K_M(\ldots)$ for "the class knows on Monday morning that …";
and
$K_T(\ldots)$ for "the class knows on Tuesday morning that …"

We can express A1 symbolically as:

([M and not-K_M(M)] or [T and not-K_T(T)]) and not both M and T.

(That is, either there will be an examination on Monday and the class does not know this on Monday morning, or there will be an examination on Tuesday and the class does not know this on Tuesday morning; and there will be an exam on at most one morning (see question 5.10).)

5.10

Would it be better to have A1 abbreviate the following?

(M or T) and not-K_M(M) and not-K_T(T).

OV can be represented as follows:

1.	Suppose A1	
2.	Suppose not-M	
3.	K_T (not-M)	(from 2 + memory)
4.	If not-M, T	(by the definition of A1)
5.	K_T (T)	(from 3 + 4)
6.	If K_T (T) and not-M then not-A1	(by the definition of A1)
7.	not-Al	(from 2, 5, + 6)
8.	M (and so not-T)	(from 1, 2, and 7)
9.	K_M (M)	(from 8 + A1)
10.	If K_M (M) and not-T, then not-A1	(definition of A1)
11.	not-A1	(from 8, 9, + 10)
12.	not-A1	(from 1 + 11)

The overall shape of the argument is *reductio ad absurdum*: one makes an assumption in order to show that it leads to a contradiction and so must be rejected. In the present case, the supposition of A1 is supposed to lead eventually to the conclusion that A1 is false. (Indentation is used to show that – and how – some steps of the argument occur within the scope of a supposition. For references and further details see the suggested reading at the end of this chapter.) It seems that we intuitively hold that A1 can be true; and that clash constitutes the paradox.

OV suffers from the defect that no adequate justification is provided for step (5). The idea is meant to be this: if A1 is true, then the examination will occur on Tuesday if it does not occur on Monday; so if we know the examination did not occur on Monday, we know that it will occur on Tuesday. However, this is not a sound inference: we would also need to *know* that the examination will occur on Tuesday if it does not occur on Monday (see question 5.11).

5.11

Can you give a simple example which brings out why this pattern of inference is not sound? Is the step at line (9) unsound for the same reason? Can the argument be repaired by including an assumption to the effect that the class does know that if the examination does not occur on Monday, then it will occur on Tuesday?

MV1 can be represented as follows:

1.	Suppose K(A1)	
2.	Suppose not-M	
3.	K_T(not-M)	(from 2 + memory)
4.	If not-M, T	(by definition of A1)
5.	K_T (If not-M, T)	(by supposition 1)
6.	K_T(T)	(from 3 + 5)
7.	If K_T then not-Al	(definition of A1)
8.	If not-A1, then not-K(A1)	(only truth is known)
9.	If K_T(T) then not-K(A1)	(from 7 + 8)
10.	M (and so not-T)	(from 1, 6 and 9)
11.	K_M(M)	(from 10) (see question 5.12)
12.	If K_M(M) and M, then not-Al	(definition of A1)
13.	not-Al	(from 10, 11, + 12)
14.	If not-A1, then not-K(A1)	(only truth is known)
15.	not-K(A1)	(from 13 + 14)
16.	not-K(A1)	(from 1 + 15)

5.12

The argument is suspect at this point. We have supposedly proved M [at (10)] on the basis of K(A1). It is quite plausible to hold that this means that we can know the corresponding conditional, namely:

If K(A1), then M.

To obtain $K_M(M)$ from K[If K(A1), then M] would appear to require as a premise not merely K(A1), but K[K(A1)]. We should avoid obtaining the latter by the dubious principle:

If K(φ), then K[K(φ)].

Why is this principle dubious? Might the argument manage with something weaker?

Even if MV1 is valid (and question 5.11 gives a reason for doubt on this point), it is questionable whether there is anything paradoxical in this conclusion. To have a paradox, we would need also to have an argument for the conclusion that K(A1). Perhaps it is just intuitively obvious that K(A1), given, if you like, the class's knowledge of the teacher's unimpeachable reputation for veracity and constancy of purpose; but suppose someone failed to share this intuition?

If not-K(A1), then it is very easy for A1 to be true: the class will not on the basis of A1 have any expectations, since the students can establish that they cannot know Al. This gives the teacher plenty of scope for surprising them.

However, the class can also go through the reasoning of the preceding paragraph: "Our proof that A1 cannot be known shows us how easy it is for A1 to be true. If it *can* be true, then, given the teacher's proven veracity and determination, we have every reason to believe that it *is* true." If the class is led by this consideration to believe the announcement, then there is a case for thinking that their belief amounts to knowledge. So it seems that *if* the argument is valid, we have a paradox (see question 5.13).

5.13

Once one starts thinking about knowledge, one can rather easily convince oneself that there is less of it than one might have thought. So I would not be surprised if someone were to say "We could not *know* that the teacher would carry out her threat, however reliable we know her to have been in the past. The most we would be entitled to is the justified belief that she would."

Rework MV1 in terms of justified belief rather than knowledge. (You will probably find you have to make an inference from "It was rational for the class to believe the teacher's announcement when it was made" to

5.13 (cont.)

"It would be rational for the class to believe the teacher's announcement on the morning of the last day, if the exam had not yet been given." Is this inference sound? Is the parallel inference in the case of knowledge sound? At what points was it assumed in the arguments displayed above?)

MV2 requires a different announcement:

> **A2**: Either [M and not-K_M(If A2, then M)] or [T and not-K_T(If A2, then T)].

(That is, the examination will take place on Monday or Tuesday, but you will not know on the basis of this announcement which day it will be.) A2 differs from A1 in a striking respect: the specification of A2 refers to A2 itself; in other words, A2 is a *self-referential* announcement.

MV2 can be represented as follows:

1.	Suppose A2	
2.	Suppose not-M	
3.	K_T(not-M)	(from 2 + memory)
4.	K_T(If not-M, then if A2, then T)	(the class understands A2)
5.	K_T(If A2, then T)	(from 3 + 4)
6.	not-A2	(from 2 + 5)
7.	M	(from 1, 2 + 6)
8.	If A2, then M	(summarizing 1–7)
9.	K_M(If A2, then M)	(the proved is known)
10.	If K_M(If A2, then M), then if A2, then not-M	(from definition of A2)
11.	If A2, then not-M	(from 9 + 10)
12.	not-A2	(from 8 + 11)

MV2 purports to prove that A2 is not true. This is paradoxical only if we have some good reason to think that A2 is, or could be, true. We seem to have some reason: have we not all been exposed to such threats of unexpected examinations? The form of A2 admittedly has the self-referential feature already noticed, but it is not clear that this should detract from its possible truth. When the teacher says that the examination is to be unexpected, what is clearly intended is that it be unexpected on any basis, including on the basis of this present announcement. So the intuitions that told us that A1 could be true, and could be known, should also tell us that A2 could be true. However, intuition may be less than wholly confident when faced with the apparent validity of MV2.

5.4 The Knower

Using a self-referential type of announcement, one can construct a further announcement, call it A3, that is certainly paradoxical. It has come to be called the Knower paradox:

A3: K(not-A3).

(As we might put it: "The class knows that this very announcement is false.")

We can represent the argument that establishes both A3 and not-A3 as follows – call it MV3:

1.	Suppose A3	
2.	K(not-A3)	(definition of A3)
3.	not-A3	(what is known is true)
4.	If A3, then not-A3	(summarizing 1–3)
5.	not-A3	(from 4)
6.	not-K(not-A3)	(from 5 + definition of A3)
7.	K(not-A3)	(5 + what is proved is known)

Lines (6) and (7) are contradictory.

In view of this result, we must examine carefully (a) the nature of the announcement and (b) the epistemic principles – the principles involving the nature of knowledge – used to reach the paradoxical conclusion. If there is anything wrong with the principles, then we may have to revise our views about the earlier arguments, for they, too, rest on these principles.

(a) We cannot satisfy ourselves merely by saying that A3 is contradictory. A contradiction is false, whereas A3, if the argument MV3 is sound, is demonstrably true (see line 7, bearing in mind how A3 is defined). More hopeful would be to say that A3 is *unintelligible*, perhaps in part because of its self-referentiality. What, we might ask, does it *say*? What is it that it claims cannot be known? If we say it claims that it itself cannot be known, we seem to be flailing in thin air rather than genuinely answering the question.

Some of this doubt might be removed by changing the example. Suppose now that we have two teachers, *X* and *Y*. *X* says "What *Y* will say next is something you can know to be false." *Y* then says "What *X* has just said is true." It looks as though we have to count both utterances as intelligible, since in other contexts they certainly would have been intelligible, and even in this context we can understand *X*'s without knowing what *Y* will say, and can understand *Y*'s without knowing what *X* has said. However, in the context *Y*'s announcement appears to be equivalent to A3. We could argue informally for the contradiction like this. Suppose *Y* is true (let *X* and *Y* now also abbreviate the respective teachers' remarks).

Then *X* is true, so you can know *Y* to be false, so it is false. So the supposition that *Y* is true leads to the conclusion that it is false. Hence we can conclude that it is false (cf. MV3, line 5). Hence we can conclude that *we can know Y to be false*. However, if *Y* is false, then *X* is false; i.e. *we cannot know Y to be false*. So it seems we have an argument that has the essential features of A3, but which has a defense against the charge that the announcement is unintelligible. (Compare Burge 1978, p. 30.)

(b) Let us isolate the three epistemic principles concerning knowledge appealed to in MV3. The first – call it EK1 – is what licenses the move from (2) to (3) in MV3. In its most general form it is that what is known is true. We could write it:

EK1: If K(φ), then φ.

The other point at which appeal to epistemic principles is made is the move at (7) from (5). It cannot be true that anything provable on the basis of no matter what assumptions is knowable. Given the assumption that $5 > 7$, I could perhaps prove that $5 > 6$ on that assumption, but obviously I could not *know* that $5 > 6$. So the principle that we need at this point is that anything proved from known assumptions (or from no assumptions) is known (see question 5.14). We could write this as:

EK2: If C is provable from (P_1…, P_n) and K(P_1, …, P_n), then K(C).

What assumptions (corresponding to P_1 etc.) are in play in the move from (5) to (7)? Just one: EK1. So, in order to apply EK2, we need to add:

EK3: K(EK1).

5.14

Compare with the principle sometimes called "epistemic closure":

If K(if φ, then ψ) and K(φ), then K(ψ).

Is EK2 entailed by the closure principle? Does the converse entailment hold?

Are these three principles plausible? Expressed informally they are the following:

EK1: What is known is true.
EK2: What is provable from things known is known.
EK3: It is known that what is known is true.

The first principle has sometimes been doubted on the grounds that, for example, people once knew that whales were fish. This doubt should be dispelled by the reflection that the correct account of the matter is that

people *thought* they knew this, although they really did not. How could they have known it if it is not even true?

EK2 does not hold generally: we do not know all the infinitely many things that could be proved from what we know; we do not even believe all these things, if only because it would be beyond our powers to bring them all to mind. However, this implausible aspect of EK2 is not required for the paradox, which only needs a much narrower claim: that at least one person who has constructed a correct proof of not-A3 from a known premise knows that not-A3.

The third principle cannot be seriously questioned, once we have granted the first. So the only doubt about the premises attaches to EK2. We could circumvent this by using an even weaker and very hard to controvert principle: what is provable from something known is *capable* of being known by a fully rational subject. With appropriate modifications to A3, we shall be able to prove a contradiction from principles that appear indubitable, together with the admission of the intelligibility of the teacher's announcement (see question 5.15).

5.15

Provide the modified A3 (call it A4) and the appropriate argument, setting out the epistemic principles in detail.

It is hard to know what to make of this paradox. One promising suggestion sees a similarity between it and the Liar paradox (see sections 6.2–6.9 below). Knowledge quite clearly involves the notion of truth, and the Liar paradox shows that this notion can lead to paradox. So perhaps what is at fault in the concept of knowledge is the concept of truth it contains, as displayed in EK1; and perhaps the remedy consists in applying to knowledge whatever non-paradoxical elaboration of the notion of truth we can extract from consideration of the Liar paradox.

The suggestion cannot be quite right for the following reason. Unlike knowledge, belief does not entail truth; yet a paradox rather like the Knower – we could call it the Believer – can be constructed in terms just of belief. Consider the following:

B$_1$: α does not believe what B$_1$ says (see question 5.16).

5.16

* Construct a version of the Believer analogous to the "What I am now saying is false" version of the Liar. (Look ahead to section 6.4.)

Does α believe B_1 or not? If α does believe B_1, then he can see that he is believing something false. There is no gap between seeing that something is false and not believing it, so if α believes B_1, he does not believe it. Equally, if α does not believe B_1, then he can see that B_1 is true. There is no gap between seeing that something is true and believing it, so if α does not believe B_1, he believes it.

The paradox depends on at least two assumptions:

(1) that α *can* see that, if he believes B_1, it is false, and if he does not believe it, it is true;
(2) that what α can see he *will* see.

Neither assumption would be capable of true generalization. For (1) to hold of α requires, among other things, that he be able to see that he is α. One could arguably envisage this not being true, if α had an unusually low level of self-awareness. For (2) to hold of α requires a positive level of intellectual energy: one does not always take advantage of one's epistemic opportunities. However, we have a paradox if we can make the following highly plausible assumption: that there is at least one person with the self-awareness and energy required to make (1) and (2) true of him (or her), at least in respect of B_1.

We can represent the argument to the contradiction, and the assumptions upon which it depends, in a manner analogous to the representation of the Knower paradox. (For a different version, see Burge 1978, esp. p. 26.) We abbreviate "α believes what ( ) says" as "B( )"; so B_1 = not-B(B_1).

1.	Suppose B(B_1)	
2.	If B(B_1), then B[B(B_1)]	(self-awareness)
3.	B[B(B_1)]	(from 1 + 2)
4.	B[If B_1, then not-B(B_1)]	(α understands B_1)
5.	If B[B(B_1)], then not-B[not-B(B_1)]	(rationality)
6.	not-B[not-B(B_1)]	(from 3 + 5)
7.	not-B(B_1)	(4, 6, + closure)
8.	If B(B_1), then not-B(B_1)	(summarizing 1–7)
9.	not-B(B_1)	(from 8)
10.	B[not-B(B_1)]	(from 9 + self-awareness)
11.	B(B_1)	(from 10 +definition of B_1)

The unconditionally derived lines (9) and (11) are contradictory.

Let us examine the assumptions upon which the argument depends. The first principle to be used is what I have called "self-awareness." It involves at least this, which is enough to vindicate the move to line (2):

EB1: If B(φ), then B[B(φ)] (see question 5.17).

5.17

Is EB1 sufficient to justify the move to line (10) of the immediately preceding argument?

This is not very plausible. If it were true, then having one belief, say φ, would involve having infinitely many: that you believe that φ, that you believe you believe that φ, and so on. However, no such contentious general principle is required for the paradox. It is enough that some person, α, should meet these two conditions: if α believes B_1, under circumstances that can be as favorable as you like to self-awareness, then he will believe he does so; and if α does not believe B_1, then he will believe he does not. It seems impossible to deny that there could be a person of whom this is true.

The second assumption is that α understands B_1 and therefore realizes (and so believes), from the definition of B_1, that if B_1 then not-B(B_1). Again, it seems impossible to deny that there could be a person who has this belief.

At line (5), the argument appealed to something called "rationality." A generalization would be the following:

EB2: If B(φ) then not-B(not-φ).

Put so generally, this is not plausible, since people in fact have contradictory beliefs without realizing it; but we need only impute a fairly modest degree of rationality to α in order for the weakest premise needed at line (5) to obtain.

A generalization of the closure principle is this:

EB2: If B(if φ, then ψ) and B(not-ψ), then B(not-φ).

For normal persons, this is not a plausible principle: we do not believe all the consequences of things we believe. However, it again seems easy to imagine that α verifies the particular case of the principle needed in the above argument.

Let us step back. A suggestion was that the Knower paradox should be treated like the Liar paradox, on the grounds that knowledge entails truth, and the Liar paradox shows that the notion of truth requires special treatment. The point of introducing the Believer paradox was to challenge this suggestion. Belief does not entail truth, yet belief gives rise to a paradox quite similar to the Knower.

The conclusion is that the reason given for treating the Knower and the Liar in similar ways is deficient. However, there is another similarity between the Knower, the Believer, and the Liar: they all involve

self-reference, as does any version of the Unexpected Examination which (like MV2) involves a self-referential announcement. We shall see in sections 6.6–6.8 that there is a case for thinking that some kinds of self-reference prevent apparently intelligible utterances from being genuinely intelligible. If the case is good, it should be considered as the basis of a possible response to the Knower, the Believer, and self-referential versions of the Unexpected Examination.

Suggested reading

Section 5.1.1

David Hume assumed that it was easy to answer the problem of characterization in the way envisaged in this section: the arguments we take to be legitimate are those in which it is assumed that the future will resemble the past, that is, that "nature is uniform." This suggestion is shown to be inadequate by the Grue paradox. Hume (1738, book I, part III) remains essential reading on the problem of the justification of induction, and on various connected issues, notably causation. For a modern account, see James Vickers' encyclopedia article: plato.stanford.edu/entries/induction-problem/.

Section 5.1.2

Hempel (1945) gives a classic account of the paradox of the Ravens. I have departed from Hempel's formulation of E2. The equivalence relation he uses is that of *logical* equivalence, not *apriori* equivalence. Two propositions are logically equivalent just on condition that some system of formal logic has a theorem saying that either both propositions are true, or else both are false. In familiar classical logic, "Tom is a bachelor" and "Tom is a bachelor or the earth is round or not-round" are logically equivalent, but "Tom is a bachelor" and "Tom is an unmarried man" are not logically equivalent (though they can be known *apriori* to be equivalent). The intuitive motivation for the equivalence principle is this, in my view: if P and Q are in the appropriate sense equivalent, then if we can find evidence supporting P, we need no further empirical data to see that Q is thereby supported to the same extent. If this motivation is accepted, it seems clear that the appropriate equivalence relation is wider than logical equivalence, and is, precisely, *apriori* equivalence.

For a brief introduction to confirmation theory, see Schlesinger (1974a). A useful collection of early papers is in Foster and Martin (1966).

Section 5.1.3

The classic source of the Grue paradox is Goodman (1955). There has been controversy about how Goodman defines "grue." He writes that "grue" is to be introduced so that:

> it applies to all things examined before *t* just in case they are green but to other things just in case they are blue. (1955, p. 74)

The time *t* is arbitrary. In giving my account, I have imagined ourselves being at that time: if all emeralds are now grue, the examined ones are now green and the unexamined ones blue; there is no question of an emerald having to change color to stay grue. For discussion of some alternative interpretations, see Jackson (1975).

My G2–4 were inspired by Jackson, whose own proposal is:

> certain *F*s which are *H* being *G* does not support other *F*s which are not *H* being *G* if it is known that the *F*s in the evidence class would not have been *G* if they had not been *H* (Jackson 1975, p. 123).

These proposals contain a subjunctive conditional ("If it had not been that …, it would not have been that …"), and so would not be acceptable to Goodman, whose overall project is to give an account of such conditionals. However, there is nothing in the nature of the characterization problem as such (section 5.1.1) which precludes the use of subjunctive conditionals in its solution.

The view that instances alone cannot make it reasonable to believe a generalization has been advocated by Foster (1983).

The example of "+*" derives from Wittgenstein (1953), esp. pp. 185ff., and has been revived by Kripke (1982). Goodman's own philosophical development has been influenced by what he would regard as ramifications of the Grue paradox – see Goodman (1978, p. 11).

Section 5.2

The paradox goes back at least to Scriven (1951); a good early discussion is Quine (1953). Recent work includes Bar-Hillel and Margalit (1983), (1985), Janaway (1989), Koons (1992) and Williamson (1992a). See also "The Grid" and "The Designated Student" in appendix I of the present volume, and compare Sorensen (1982). For a good overview, see Roy Sorensen's encyclopedia article at plato.stanford.edu/entries/epistemic-paradoxes/.

G. E. Moore considered that it was paradoxical for me to assert "*p*, but I do not believe that *p*," despite the fact that the quoted sentence is consistent and, indeed, the fact it expresses might be true of me. (E.g. you

could truly say of me: "*p*, but he does not believe it.") This may be connected with the Unexpected Examination: see Wright and Sudbury (1977).

Section 5.3

The proofs are set out in a style invented by Fitch (1952). The style is used in several more recent books, e.g. Thomason (1970). I hope that their intended structure will be self-explanatory, but some observations may be useful.

For example, what is the difference between lines (7), (11), and (12) on p. 111? Each has the same conclusion, but it has been reached from different suppositions, as the different degrees of indentation show. At (7), the argument claims that we have reached not-A1 on the basis of supposing that A1 is true and that not-M is true. This would mean that we have reached a contradiction: since anything entails itself, the supposition of both A1 and not-M leads to the contradiction that A1 and not-A1. We must therefore reject at least one of these. Line (8) claims that if we hold on to A1, we must reject not-M (equivalently, T). At line (11), not-A1 depends only on the supposition of A1 itself. In other words, at this point we have shown that A1 entails its own negation. This is enough to show that, on the basis of *no* suppositions at all, we can infer the falsehood of Al, since anything entailing its own negation is false, and this is what (12), by having *no* indent, expresses.

Section 5.4

The classic statement of the Knower is Montague and Kaplan (1960). A similar paradox is in Buridan's Sophism 13; see Hughes (1982). For the Believer, and a comparison with the Liar, see Burge (1984). See also Asher and Kamp (1986).

6 Classes and truth

The paradoxes to be discussed in this chapter are probably the hardest of all, but also the most fecund. Russell's paradox about classes, which he discovered in 1901, led to an enormous amount of work in the foundations of mathematics. Russell thought that this paradox was of a kind with the paradox of the Liar, which in its simplest form consists in the assertion "I am now (hereby!) lying." The Liar paradox has been of the utmost importance in theories of truth. Everything to do with these paradoxes is highly controversial, including whether Russell was right in thinking that his paradox about classes and the Liar paradox spring from the same source (see section 6.9).

6.1 Russell's paradox

If Socrates is a man, then he is a member of the class of men. If he is a member of the class of men, then he is a man. Can *classes* be members of classes? The answer would seem to be Yes. The class of men has more than 100 members, so the class of men is a member of the class of classes with more than 100 members. By contrast, the class of the Muses does not belong to the class of classes having more than 100 members, for tradition has it that the class of the Muses has just nine members.

Most classes are not members of themselves. The class of men is a class and not a man, so it is not a member of the class of men, that is, not a member of itself. However, some classes are members of themselves: the class of all classes presumably is, and so is the class of all classes with more than 100 members. So is the class of non-men: the class of all and only those things that are not men. No class is a man, so the class of non-men is not a man, and is therefore a member of the class of non-men.

Consider the class of all classes that are not members of themselves. Let us call this class *R*. The necessary and sufficient condition for something to belong to *R* is that it be a class that is not a member of itself. Is *R* a member of itself?

Suppose that it is. Then *R* must meet the (necessary) condition for belonging to *R*: that it is not a member of itself. So if it is a member of itself, it is not a member of itself.

Suppose that it is not. Then, being a non-self-membered class, it meets the (sufficient) condition for belonging to *R*: that it is not a member of itself. So if it is not a member of itself, it belongs to *R* and so is a member of itself.

Summarizing: *R* is a member of itself if and only if it is not a member of itself. This is contradictory (see question 6.1).

6.1

The Class paradox, as Russell saw, is very similar to one about properties. Most properties are not applicable to themselves. The property of being a man is a property and not a man, so it does not apply to the property of being a man; that is, it is not self-applicable. Some properties are self-applicable: the property of being a property presumably is, and so is the property of being true of more than 100 things; etc.

How would you spell out the contradiction about properties?

To have a contradiction is not necessarily to have a paradox. Recall the Barber paradox from the introduction. The Barber shaves all and only those who do not shave themselves. Who shaves the barber? By reasoning similar to that used to derive Russell's paradox, we find that the barber shaves himself if and only if he does not.

We respond to the Barber paradox simply by saying that there is no such barber. Why should we not respond to Russell's paradox simply by saying that there is no such class as *R*? The difference is this: nothing leads us to suppose that there is such a barber; but we seem to be committed, by our understanding of what it is to be a class, to the existence of *R*. We are forced by the paradox to accept that there cannot be such a class. This is paradoxical because it shows that some very compelling views about what it is for a class to exist have to be abandoned.

The first paragraph of this section was supposed to introduce the natural or intuitive view of classes, which I must now make more explicit. I said that if Socrates *is a man*, then he is a member of the class of men. Let us use "condition" for what is expressed by, for example, the italicized phrase just used. Being a man is a condition, and one that Socrates satisfies, although Mont Blanc does not. The natural view of classes includes this principle of Class Existence:

> **CE**: To every intelligible condition there corresponds a class: its members (if any) are all and only the things that satisfy the condition.

Corresponding to the condition of being a man, there is the class of men. Even when a condition is contradictory – for example, the condition of being both square and not square – there is a corresponding class; though since nothing meets the condition, this is a class with no members (the empty or "null" class).

CE appears to lead to Russell's paradox. It entails that there is such a class as *R* if there is the intelligible condition: being a class that is not a member of itself. The condition appears intelligible; yet we have already seen that there cannot be such a class as *R*.

We could put this point in a more symbolic and more perspicuous way as follows. Let us use "ε" to abbreviate "is a member of" (and "belongs to"). Then we can rewrite CE as follows:

> **CE**: For every intelligible condition, *F*, there is a class, *x*, such that: for anything *y*, *y* ε *x* if and only if *y* satisfies *F*.

For "*y* satisfies *F*" we can write, simply, "*y* is *F*"; for "if and only if" we can write "iff". Putting "*R*" for Russell's paradoxical class, "¬" for "not", and "¬ is a member of itself" for "*F*", CE yields:

For anything *y*, *y* ε *R* iff ¬(*y* is a member of itself).

For something *y* to be a member of itself is for it to belong to *y* so we can rewrite the above as:

For anything *y*, *y* ε *R* iff ¬ (*y* ε *y*).

What holds for anything must hold for *R*, so we get the explicitly contradictory:

RP: *R* ε *R* iff ¬ (*R* ε *R*).

A natural suggestion is that the condition "being non-self-membered" is not genuinely intelligible – and this is, in effect, what most responses to Russell's paradox propose. If we follow this suggestion, CE can be preserved, as long as we take a sufficiently narrow view about what constitutes a condition. However, if we are also to preserve some well-known results in mathematics, it is far from obvious what this narrower view ought to be. In particular, reasoning and assumptions very like those that occur in the derivation of Russell's paradox also occur in a famous proof by Cantor, which I shall now set out. Indeed, it was studying Cantor's proof that led Russell to the discovery of the paradox. It is hard to see how to block the paradox while allowing the proof.

What is to be proved is that the power class of any class has more members than the class. (For a definition of *power class*, see the suggested

reading at the end of this chapter.) Cantor's proof could be sketched informally as follows:

1. A class must have *at least* as many subclasses as members, since for each member the unit class to which it alone belongs is a subclass.
2. So either there are as many subclasses as members or more.
3. Suppose there are as many. This means that there is a one–one function f correlating members of the class with its subclasses.
4. Now form the following subclass S: $x \in S$ iff $\neg[x \in f(x)]$. By the supposition at (3), for some α, $S = f(\alpha)$. Applying the definition of S we get: $\alpha \in S$ iff $\neg [\alpha \in f(\alpha)]$.

Therefore, given $S = f(\alpha)$:

($\star$) $\alpha \in S$ iff $\neg.(\alpha \in S)$.

5. This contradiction shows that we must reject the supposition at (3). Hence we must adopt the other alternative available in (2): a class has more subclasses than members.

In (4) we have a contradiction which resembles that involved in Russell's paradox (compare the asterisked line with RP), but here used to serious and informative effect in the proof. The proof assumes that *if* there is a function f, then there is a subclass S. If we are to use the current suggestion that an overliberal interpretation of CE is to blame for the paradox, and yet preserve Cantor's proof, we need to find a restriction on the notion of a *condition* that allows, via CE, the hypothetical existence of the class S but disallows the existence of R. Moreover, to be philosophically satisfying, there must be a philosophical justification for the restriction, enabling us to understand the origin of the paradox and to feel that we have something better than an ad hoc blocking manoeuvre. Russell attempted to provide precisely this in his Vicious Circle Principle (VCP). His idea is that the condition involved in the specification of R is viciously circular, and therefore not intelligible. He thought that the VCP also explained away a number of other paradoxes including, most importantly, the Liar paradox. I shall discuss the VCP (in sections 6.8 and 6.9) in the course of the discussion of the Liar (beginning in section 6.2).

For Russell, the VCP provided the philosophical motivation for his Theory of Types, according to which classes are arranged in a hierarchy, in such a way that every class is on a higher level than any of its members. The theory ensures that no class belongs to itself: no expression of the form $x \in x$ counts as meaningful. Simplified versions of Russell's Theory of Types have been the dominant tradition in mathematical work on classes. One must distinguish between the Theory of Types on the one hand, a device designed to ensure that paradoxes do not inhibit mathematical work, and, on the other

hand, a justification for such a theory, like that which, according to Russell, is provided by the VCP. The justification ought to help us understand what it is about classes that calls for a Theory of Types. The search for understanding of this kind is a distinctively philosophical project, upon which the working mathematician can reasonably decline to engage.

6.2 The Liar: semantic defects

The material of the next four sections is hazardous. (Recall the fate of Philetas, mentioned in the introduction.)

A relatively recent version of the Liar paradox appears in St. Paul's Epistle to Titus (1, 12–13) (see question 6.2). This version involves the island of Crete and the notion of lying, and lying involves an intention to deceive. These features are irrelevant to the paradox. Eliminating such irrelevancies, we get something like this:

What I am now saying is false.

6.2

It is not clear that the saint sees any logical, as opposed to moral, problems. The relevant text is as follows:

12. One of themselves, even a prophet of their own, said, The Cretans are always liars, evil beasts, slow bellies.
13. This witness is true. Wherefore rebuke them sharply, that they may be sound in the faith.

St. Paul's version depends on the assumption that all the other Cretans are liars. Construct an explicit argument for the contradiction (perhaps modeled on that given below for L_1) that makes this dependence plain.

The simplest version of all, which will be the starting point of the discussion, is

L_1: L_1 is false.

Here we have a sentence, called L_1, that is supposed to say of itself that it is false. One can derive something apparently paradoxical as follows. Suppose it is true; then it is as it says it is – false. So it is false. Suppose that it is false. Well, *false* is just what it says it is, and a sentence that tells it the way it is is true. So it is true. To sum up: if L_1 is true, it is false; and if it is false, it is true.

Is this paradoxical? Perhaps it sounds as if it is, but let us look more carefully. We have two conditional claims:

(a) If L_1 is true, then it is false.
(b) If L_1 is false, then it is true.

We assume that anything that is false is not true, and anything that is true is not false; so (a) and (b) yield:

(a′) If L_1 is true, then it is not true.
(b′) If L_1 is false, then it is not false.

If a sentence implies its own negation, then we can infer that negation. (This principle is called *consequentia mirabilis*. It amounts to the validity of the sequent: $A \rightarrow \neg A \vDash \neg A$.) Both (a′) and (b′) offer inputs to this principle. The first assures us that "L_1 is true" implies its negation, so the principle tells us that we can infer that L_1 is not true. The second, in an exactly parallel way, enables us to infer that L_1 is not false. So standard reasoning guarantees that L_1 is not true and is also not false. Let us summarize this as follows:

G: L_1 is neither true nor false.

Is *this* paradoxical? Not unless we have some independent reason to suppose that L_1 is either true or false. For example, we might be able to justify some *principle of bivalence*, perhaps to the effect that *every* sentence, and so in particular L_1, is either true or false. Otherwise we might simply *accept* G, saying that L_1 lies in a *gap* between truth and falsehood (hence the "G"). This would not in itself offer a complete account of the paradox, for it would remain to discover general principles to explain why L_1 should fail to be true and fail to be false. But accepting G would at least fix the general approach.

We could not accept G if there were some irresistible reason for supposing that L_1 had to be either true or false. Might there be an irresistible reason for accepting some principle of bivalence? The version given in the preceding paragraph is certainly not true. Questions are expressed in sentences, but no question is either true or false. Suppose then we restrict the principle to declarative, indicative sentences. Still, there are putative counterexamples, for example:

You have stopped beating your wife.

If you have never beaten your wife, the sentence is certainly not true; but to say it is false, or to say that you have not stopped beating your wife, arguably suggests that you are still beating her. Again, consider a case in which someone says

That elephant is about to charge

when there is no elephant in the offing. We certainly cannot count the sentence as true; but can we count it as false? If we did this, should not the following sentence be true?

That elephant is *not* about to charge.

Yet, if there is no elephant, this seems as poor a candidate for truth as the previous one.

Despite the apparent counterexamples, it is hard not to feel the pull of the thought that some principle of bivalence, no doubt suitably refined, ought to be correct. The underlying idea might be expressed like this: any non-defective representation of how things are in the world must be either accurate or inaccurate, true or false. Some sentences, like questions and commands, are not designed to represent the world, so there is no question of them representing it correctly or incorrectly. Other sentences, though designed to represent the world, fail to count as representations at all, correct or incorrect, in virtue of some semantic defect. The case of the missing elephant is a putative example. For that sentence to represent the world at all, it must refer to an elephant (or so it is plausible to suppose). Since it fails to refer, it counts as semantically defective, and so counts as neither true nor false.

In sum, one natural and immediate response to the Liar is to accept the reasoning which leads to the conclusion that the paradoxical sentence, L_1, is neither true nor false (see question 6.3). Since it is hard to suppose that a semantically non-defective sentence could fail to be either true or false, this approach carries with it the obligation to explain wherein the defectiveness of L_1 consists. It would be totally unexplanatory to say that the defect consists in the sentence's potential for paradox, for that potential is precisely what we need to understand.

6.3

* Show how the principles of reasoning in this section can apparently be used to derive that L_1 is both true and false. This derivation shows that one should not regard L_1 as a basis for a straightforward proof of G.

Most accounts of the Liar paradox endeavor to establish plausible general principles upon which the Liar sentences are defective. We will consider some of these in detail in subsequent sections.

6.3 Grounding and truth

One approach to identifying a semantic defect in L_1 starts from the idea that the truth of a sentence must be grounded in something outside the sentence itself. We can make the thought vivid by imagining how one might introduce someone to the notion of truth.

We take the learner to understand most of English, but not the word "true." We could try to explain the notion of truth using the following recipe:

You should call a sentence true iff you are willing to assert it.

("Iff" abbreviates "if and only if.") The learner could use this explanation to respond to, for example, "Snow is white" by saying "True!," and to respond to "Grass is red" by saying "Not true!" However, he could not in the first instance use the explanation to find out how to respond to a sentence like:

(1) "Snow is white" is true.

Until he has already understood "true," he cannot know what it would be to be willing to use this sentence to assert something. At a later stage, once he becomes aware that he is to respond with "True!" to "Snow is white," he will be able to see that he should assent to (1), and hence see that (1) is itself something to which "true" applies. The picture is of someone climbing a ladder. At the base there are sentences not containing the word "true," to which he can learn to apply the word. As he does so, he can thereby come to see how to apply the word to sentences such as (1) on the next rung up; sentences applying "true" to sentences at the base. He can work his way up this ladder indefinitely. Where *S* is a sentence not containing "true," he can use this process to understand any sentence of the form

… "S is true"' … is true,

where the second ellipsis stands in for any number of further occurrences of "is true," (see question 6.4).

6.4

What does the first ellipsis indicate?

Learning how to apply the concept of truth requires there to be sentences that do not themselves invoke the concept: these are the *base* sentences. The learning situation mirrors a putative metaphysical fact: *truth depends on something outside itself*. One might defend the claim that L_1 is neither true nor false on the grounds that L_1 does not respect this fact. Accordingly, it is semantically defective. Let me try to explain this.

Whether or not "Snow is white" is true depends on whether or not snow is white. In this simple case, whether or not something is true depends quite directly on a fact that can be expressed without invoking the concept of truth: on whether or not snow is white. This is an example of how truth depends upon something outside itself. In more complex cases, the

dependence is less direct. For example, consider (1) again ("'Snow is white' is true"). Whether (1) is true depends on whether or not "Snow is white" is true. This in turn depends on whether or not snow is white. So whether or not (1) is true depends, but at one remove, on whether or not snow is white. In the end, we get back to a non-truth-involving question. In this reflection, we travel down the ladder toward the base. In considering learning, we were traveling upward from the base – same ladder, different direction.

To reinforce the suggestion, consider this series of sentences:

(S2) (S1) is true
(S3) (S2) is true
(S4) (S3) is true
⋮

Can we genuinely make sense of such a series? Everything depends upon what (S1) is. If it is, for example, "Snow is white," then there is no problem: we reach base. However, we would never reach base if (S1) were, for example,

(S1) (S4) is true.

Here truth wanders in a circle, without ever touching the ground. In this case we need to say that none of the sentences is true – and also, for the same reason, none is false (Kripke 1975).

This line of thinking gives a general reason for accepting G (L_1 is neither true nor false). L_1 can never reach base: there is no getting to a non-truth-involving fact on which the truth or falsehood of L_1 could depend. We come back always to L_1 itself, which is not a base sentence. To summarize: the trouble with L_1 is that it is *ungrounded*.

The same account also applies well to

T_1: T_1 is true.

Here is a sentence that seems to say of itself that it is true. It is not paradoxical. The supposition that it is true does not lead to the conclusion that it is not; the supposition that it is not true does not lead to the conclusion that it is. Still, intuitively there is something wrong with T_1, and L_1 shares the defect. The account just given purports to identify this defect: T_1 is ungrounded. Like L_1, it does not make contact with a non-truth-involving base, so both sentences are neither true nor false.

Thus G can be defended. We can provide reasons, independently of threat of paradox, for thinking that L_1 is neither true nor false: it is ungrounded. But even if all this is accepted, paradox remains.

6.4 The Strengthened Liar

G says that L_1 is neither true nor false, and thus accepts the reasoning we considered at the beginning of section 6.2. However, G itself appears to support a paradox.

G entails that L_1 is not false. This is the negation of L_1 itself. So G entails

not-L_1: L_1 is not false.

So not-L_1 is true (using the principle that anything that entails a sentence entails the truth of that sentence). This in turn entails that L_1 is false (using the principle that any sentence whose negation is true is false). So G appears to entail a contradiction: that L_1 is not false and L_1 is false (see question 6.5). Hence it cannot constitute a resolution of the paradox.

6.5

* Show how one can (apparently) *derive* that L_1 is not false without appeal to G.

A related difficulty is that G is unable to deal with a related paradoxical sentence:

L_G: L_G is either false, or else neither true nor false.

We can reason as follows: suppose L_G is neither true nor false; then it is true (since it is an or-statement one of whose alternatives is true), and so it is either true or false. We can then reason as we did with L_1 to show that it is neither true nor false. Combining results, we show that it is both neither true nor false, and also either true or false.

The easiest way to see what is going on in reasoning of this sort is to consider yet another paradoxical sentence:

L_2: L_2 is not true.

Suppose L_2 is true. Then it is as it says it is, namely, not true; so it is not true. Suppose that it is not true. Well, *not true* is just what it says it is, and a sentence that tells it the way it is is true; so it is true. To sum up: if L_2 is true, it is not true; if it is not true, it is true.

This appears to be a genuine contradiction, and one which cannot be assuaged by G. If L_2 is, as G affirms, neither true nor false, then in particular it is not true. But the reasoning just advanced purports to show that one can refute this claim: if L_2 is not true, then it is true (since *not true* is just what it says itself to be). If G is supported by ideas about grounding, the problem is simply that an ungrounded sentence is

not true: that was the whole idea. So if L_2 is ungrounded, *L_2 is not true*; we are committed to the problematic L_2 itself.

One may be tempted to try to modify G, for example to something entailing

G′: L_2 is neither true nor not true.

First, this appears to be a contradiction. Standard reasoning would enable us to infer from G that L_2 is both true *and* not true (see question 6.6). Second, like G, G′ implies directly that L_2 is not true: it entails the paradoxical sentence itself.

6.6

The reasoning depends on the equivalence between

neither *P* nor *Q*

and

both not-*P* and not-*Q*.

On what other principle does the reasoning depend?

L_2 and the attendant reasoning is sometimes known as the "Strengthened Liar." A standard view is that it shows the inadequacy of theories like the one based on the notion of grounding as resolutions of Liar paradoxes. More generally, it might be taken to show that any approach which tries to resolve the paradox by finding some semantic defect in L_2 is doomed, since what is semantically defective is not true.

Leaving this claim in abeyance for the moment, I now turn to an approach to resolving these paradoxes, derived from Tarski, for which the Strengthened Liar poses no special problem, no problem not already posed by the ordinary Liar. This approach also finds something semantically defective in sentences like L_1 ("L_1 is false") and L_2 ("L_2 is not true"), but of a quite different kind.

6.5 Levels

In deriving apparently unacceptable conclusions from L_1 and L_2, we relied upon two principles:

if a sentence is true, then things are as it says they are;
if things are as a sentence says they are, then the sentence is true.

Tarski stressed the feature of truth these principles capture. He expressed it somewhat more formally. Let us use "σ" to stand in for a name of any

sentence, and "*p*" to stand in for a sentence. Then, Tarski claimed, for any acceptable language we must accept every instance of

T: σ is true iff *p*

provided that the sentence named by σ means the same as the sentence that replaces *p*. In the limiting case, these can be the same sentence; so an instance of T (putting "'Snow is white'" for "σ" and "snow is white" for "*p*") is

"Snow is white" is true iff snow is white.

T may seem utterly platitudinous; but the Strengthened Liar shows that it has contradictory instances. Putting "L_2" for "σ" and "L_2 is not true" for "*p*," we get:

(⋆) L_2 is true iff L_2 is not true.

Since L_2 is "L_2 is not true," (⋆) presumably meets the requirement that the sentence named by "L_2" (namely "L_2 is not true") means the same as the sentence which replaces *p* (namely "L_2 is not true").

One aspect of the problem posed by the Liar is that the apparently platitudinous T leads by apparently correct reasoning to the contradictory (⋆). Tarski's response is that the ordinary concept of truth, the one we use every day, is incoherent and must be rejected. According to Tarski, it needs to be replaced by a series of concepts of truth, hierarchically arranged, and each expressed in a language different from any natural language (i.e. from any language that has evolved naturally).

Suppose some language λ_0 contains a predicate "Tr_1" that applies to all and only the true sentences of λ_0. Suppose also that λ_0 contains a sentence σ that says of itself that it is not Tr_1. Then, granting T, we have a version of the Liar: if Tr_1 applies to σ, then σ says truly that Tr_1 does not apply to it; but if Tr_1 does not apply to it, then, since this is what it says, it is true, and so Tr_1 does apply to it. Tarski took the contradiction to refute the supposition that σ belongs to λ_0. The natural explanation of how this could be is that Tr_1 is not an expression of λ_0. Hence, no sentence belongs to λ_0 if it contains Tr_1. This blocks the paradox in the following sense: the proposed language, since it does not contain a predicate true just of its true sentences, is one in which the paradoxical sentence cannot be formulated. One can write down the words, but they are claimed to have no significance: they are semantically wholly defective.

We can enlarge a language by adding new expressions. In particular, we could enlarge λ_0, taken to contain no occurrence of "Tr_1," by adding "Tr_1." We could call the newly formed language λ_1: it contains all the sentences of λ_0 together with all sentences which can be formed from these

by using "Tr_1"; so it contains σ. Paradox is still avoided: σ does not belong to λ_0, and since Tr_1 is defined only for λ_0 sentences, there is no question of Tr_1 applying to σ. The expression σ (= "σ is not Tr_1") does not belong to λ_0, and so it is not one of which "Tr_1" can be significantly affirmed or denied.

It is not that there is no predicate true of just the sentences of λ_1. There is: call it "Tr_2" (see question 6.7). However, for familiar reasons, it cannot belong to λ_1 (see question 6.8). In general, a predicate Tr_n cannot belong to a language λ_{n-1} but only to a language of level at least n.

6.7

Is σ Tr_2?

6.8

How does the supposition that Tr_2 belongs to λ_1 lead to paradox?

No paradoxical Liar sentence can be formulated in any of the languages in Tarski's hierarchy. How is this supposed to provide a "solution" to the paradox? The paradox arises in our language, so a proper defusing of it must say something about our language, and not merely offer a replacement.

What Tarski says about our language is that the Liar shows it to be incoherent. We must replace our actual, but incoherent, concept of truth by a family of new concepts, each fixed to a level in the hierarchy, in the way just described. Many people have sought something less radical, a response that preserves more of our ordinary thought and talk.

One such less radical response draws on a Tarskian notion of hierarchy, but claims that this is already implicit in our actual use of "true." Unlike Tarski's account, which claimed that ordinary language is irremediably defective, this alternative claims that the defects are mere appearance: the underlying reality is that we already use a Tarski-like hierarchy of concepts of truth.

A major difficulty with this suggestion is that there would appear to be nothing in our usage reflecting the appropriate sensitivity to Tarski-style, fixed-in-advance levels. For example, suppose I say:

What you said just now is not true.

On the face of it, anyone, including myself, could quite well know what I have said without knowing what you have said. (Imagine a game on the lines of paper, stone, and scissors, in which two players have to make a simultaneous

declaration. The task of one is to say whether the other has declared something true. Normally things work well: you declare "Snow is white" and I declare "Not true!" and you win. But what happens when I declare "Not true" and you declare "You win"? My declaration is intelligible in advance of knowing the content of yours.) On a hierarchical view in which levels are fixed in advance, something in my use of this sentence determines an association between "true" and some level. Presumably the normal (default) level would be 1. If you have said "Snow is white," there is no problem. But suppose you have said "What Mark will say is true." On the present theory, the intelligibility of my utterance requires my "true" to be on a higher level than yours; but if my utterance can be understood without knowing what you have said, its level of truth must get fixed independently of the content of what you have said. This suggests that it will be difficult to apply this kind of hierarchy response to natural language. (However, compare Burge 1979.)

So far we have considered two main ways of making good the claim that Liar-paradoxical sentences are semantically defective. One used the notion of grounding, in terms of which there seemed to be some hope of defending the view that L_1 is neither true nor false; though the hope that this would lead to a resolution of all versions of the Liar was apparently dashed by the Strengthened Liar (L_2: L_2 is not true). The other was Tarski's claim that any non-hierarchical notion of truth is incoherent. The Strengthened Liar creates no special problem for this view (see question 6.9). However, it has difficulties. To jettison our ordinary concept of truth seems too radical; yet it seems incorrect to suppose that our concept already contains, implicitly, the required segregation into levels. Where else might one look for an account of the semantic defectiveness of Liar sentences?

6.9

Why not? You might like to answer by criticizing one or both versions of the following reasoning:

Version 1:

Even when levels of truth are made explicit, as Tarski requires, we can formulate a Liar sentence, e.g.:

L_N : L_N is not true_n.

If this is defective, through infringing levels requirements, then it is not true_n; but since this is what it says it is, it must be true_n after all.

Version 2:

A sentence which violates levels is semantically defective and so not true; so one can always construct a Strengthened Liar sentence to refute a levels approach to the paradoxes. (Cf. the argument mentioned at the end of section 6.4 above.)

6.6 Self-reference

It is natural to think that something about the self-referential character of Liar paradoxical sentences is the main source of their paradoxical nature. There may be something in this thought, but as it stands it is both incorrect and inadequate.

It is incorrect because a sentence can refer to itself, as for example this very sentence does, without leading to any kind of semantic defect or paradox. So sentential self-reference cannot be the sole source of Liar paradoxes.

It is inadequate because one can construct Liar paradoxes without using any sentence which refers to itself. One example of this phenomenon involves liar cycles like the following.

(A) (said by α on Monday): Everything β will say on Tuesday is true.
(B) (said by β on Tuesday): Nothing α said on Monday is true.

If α and β said nothing other than, respectively, (A) and (B) on, respectively, Monday and Tuesday, we have a paradox of essentially the Liar type. Suppose (B) is true; then (A) is not true, and β will say something not true on Tuesday. Since β only says (B), (B) is not true. So if (B) is true, then it is not true. Suppose (B) is not true; then α said something true on Monday. Since α only said (A), (A) is true, that is, everything β will say on Tuesday is true. This includes (B), so (B) is true. Thus if (B) is not true, it is true.

Neither of the sentences in the story literally refers to itself. Rather, there is a kind of circle, so perhaps we should talk of "circular reference" rather than self-reference. However, as the circularity does not strictly involve reference at all, but rather quantification, it might be safer still just to speak of circularity.

We could expand the story of (A) and (B) by imagining a third utterance:

(C) (said by γ on Tuesday): Nothing α said on Monday is true.

The fact that β and γ use the very same sentence, yet only one of them is circular in the relevant way, shows that circularity is not a property of sentences as such. Being meaningful or meaningless is a property of sentences. Since there is nothing paradoxical about (C), there is no reason to say it is other than meaningful, and since (B) is the same sentence, it follows that the property which circularity prevents is not that of being meaningful. We need a more refined notion, one sensitive to the use to which a sentence is put on a specific occasion. Such a notion emerges naturally from a consideration of indexicality; and the remaining two responses to the Liar which I shall consider both claim to discover, in reasoning related to the Strengthened Liar, some element of

indexicality. One response locates the indexicality in the specific kind of self-reference involved in the Liar; the other, briefly mentioned in the last paragraph of section 6.8, locates it in the predicate "true."

To summarize: if we are to finger self-reference as the villain of the piece, the relevant kind of self-reference must contain an indexical element. However, once indexicality is allowed, we also open the way to a Tarski-like hierarchy of levels, triggered by indexical features of "true."

6.7 Indexicality

For reasons independent of Liar paradoxes, it is necessary to distinguish between sentences, regarded as things which can be uttered by different people and on different occasions, and the things which people can say or express by using sentences. The reason is the "indexicality" of language: the fact that the same words may, without exploiting ambiguity, be used on different occasions to say different things. Indexicality in pronouns provides a familiar example: if you use the sentence "I am hungry" affirmatively, you say one thing, and if I use it affirmatively I say another. The things said are different because it could be that what you say is true whereas what I say is false.

I will use *statement* for what a sentence is used on a specific occasion to say or express. Indexicality shows that it is only statements and not sentences that can properly be said to be true or false. I shall assume that bivalence holds for statements, so that every statement is either true or false. We have already in effect seen that a sentence can be meaningful, yet on a specific occasion be used in such a way as to fail to make a statement ("That elephant is about to charge"). Although sentences can be self-referential, or more generally can have the kind of circularity associated with paradox, it may be that statements cannot. Thus, reverting to the example in section 6.6, we might be able to justify the claim that whereas both β and γ use the same sentence, only γ thereby succeeds in making a statement. The notion of a statement thus seems to have the features we were looking for: it is a function not only of the meaning of a sentence, but also of the use to which it is put on a specific occasion.

The Strengthened Liar needs to be adapted to the distinction between sentence and statement. One way to do so is as follows:

$L_2^\star$: $L_2^\star$ does not express a true statement.

Reconsideration of the reasoning involved in the Strengthened Liar supports the view that some kind of indexicality is at work. We contemplate $L_2^\star$ and regard it as defective. When we come to express this, we may do so by words which are, or entail, $L_2^\star$ itself, for example: "$L_2^\star$ is semantically defective, so (*a fortiori*) it does not express a true statement; only the

use of a non-defective sentence could do that." Intuitively this at first glance seems perfectly sensible (until we realize that we have ourselves re-used the very words we wish to say are defective). This intuition could be vindicated if we could show that the same words, even referring to the same thing, and applying the same predicate to it, may not say the same thing on two occasions of use. We want to say that the first use of $L_2^\star$ is defective, but the second use of those very words is not, since they are then used to express a truth.

The general feasibility of such an approach is suggested by considerations like the following. Suppose that the displayed sentence is the only sentence written on the board in room 101:

The sentence written on the board in room 101 does not express a true statement.

It appears perfectly consistent for me to write on this page that, because of some semantic defect, the sentence written on the board in room 101 does not express a true statement. I use the words which, as written on the board in room 101, are defective, in circumstances in which there is nothing defective about their use. This suggests that the same words, used to refer to the same thing, and applying the same predicate to it, do not necessarily make the same statement. The sentence written on the board in room 101 makes no statement, in its use in room 101; whereas I use those words, on this page, to make a true statement. I did not have to use the same words. Under suitable circumstances I could simply have said *"That* sentence does not express a true statement." That there are special circumstances under which I can reuse the same words is an accident of our use of language, and does not affect the truth of the statement I wish to make.

We are still some way from our goal, for three tasks remain: (i) to provide a more detailed account of what the problematic circularity is; (ii) to give some independent justification for saying that a statement cannot possess it; and (iii) to return to the problems posed by the Strengthened Liar.

6.8 Indexical circularity

For a specification and justification of the relevant kind of circularity, it is worth reverting to an idea of Bertrand Russell's, his so-called Vicious Circle Principle. He gives more than one account of what the principle is, but a fair statement (not a quotation) would be as follows:

> **VCP**: No totality can contain members fully specifiable only in terms of itself.

This is intended as a general metaphysical principle, applicable to classes as well as everything else, and thus also applicable to statements (or as Russell sometimes said, propositions).

The VCP gets no grip on totalities of straightforward material things, for none of these contain members which could be specified only in terms of themselves. We might specify Fred as the tallest man in the regiment, thus specifying him in terms of a totality to which he belongs, but it could not be that this is the *only* way in which he could be specified. The VCP has no tendency to wipe out regiments.

The VCP appears to get no grip on sentences, thought of as marks or shapes, for these, like members of regiments, can be specified in all sorts of independent ways, not all of which involve any kind of totality. In the case of statements, however, it does seem that some can only be fully specified in terms of a totality. Thus my statement that *everything you said in your radio talk was rubbish* can only be fully specified in terms of the totality of your statements. It can be incompletely or indirectly specified in other ways: for example, as the statement which resulted in the end of our friendship, or as the statement whose verbal expression occurs in italics on page 140 of *Paradoxes*. Full specification, however, does seem to involve collective mention of the things you said. In this case, there is no infringement of the VCP, since the totality of your statements does not include mine, and the totality of your statements together with mine is not the totality in terms of which my statement is specified (but is, rather, a larger one).

Earlier in this chapter, I have "specified" a sentence simply by writing it down, and putting a name for it at the left. If we think of a sentence as a purely syntactic thing, there is no room for my failing in this attempt to specify a sentence. But a statement is something more abstract, something which we may hope a sentence expresses, though we cannot always be sure. The possibility of failing in an attempt to use a sentence to specify a statement is just what the present kind of response to the Liar paradoxes seizes on. So there is a substantive question: does just writing what I write in the displayed line below constitute the specification of a totality of statements?

$\mathbf{L}_1{}^{\dagger}$: The statement $L_1{}^{\dagger}$ is false.

If it does succeed, the totality will contain the statement $L_1{}^{\dagger}$ as its only member. But $L_1{}^{\dagger}$ can only be "fully specified" in terms of that member. The VCP rules that there is no such totality: that is, no totality whose only member is $L_1{}^{\dagger}$. Hence, applying the VCP, there is no such thing as the statement $L_1{}^{\dagger}$, since were there such a statement there would be such a totality.

There seems to be some hope that the VCP could be extended to deal with Liar cycles, and thus that we could use it to accomplish our first task, that of specifying the nature of the problematic circularity. The second task is to give an independent motivation for saying that statements cannot possess the relevant kind of circularity. On this point, Russell offers us little by way of argument. The VCP is meant to seem independently acceptable, and is supposed to deliver appropriate restrictions. It is, I think, quite hard to find intuitively plausible considerations tending to show directly that statements cannot be circular in problematic ways. One specific reason for concern is that a kind of circularity which would be precluded by the VCP has shown itself amenable to systematic mathematical treatment, namely, the kind of circularity involved in non-foundational set theory. (See the suggested reading at the end of this chapter for a brief description and references.) So developing the idea that the root of the paradoxes lies with statement self-reference, or more generally statement circularity, requires work on this point.

Supposing that this can be successfully accomplished, will not the Strengthened Liar lie in wait to make the efforts pointless? The problem appears most threatening if we reconsider the sentence $L_2^\star$:

$L_2^\star$: $L_2^\star$ does not express a true statement.

"$L_2^\star$" labels a sentence, and the approach to the paradoxes under discussion wants to claim that it is a sentence which does not express a statement: thanks to circularity, it is semantically defective. Nothing which is semantically defective expresses a true statement, so in particular *$L_2^\star$ does not express a true statement*. The italicized words are just $L_2^\star$ itself, and we seem to have the makings of the familiar paradox.

The theorist must hold that his use of the words (as italicized) does express a statement, and a true one, and thus differs from the original use of these words. He could have made his point by using the non-paradoxical words "The sentence displayed above does not express a true statement." It is an accident of our labeling conventions that the very words to be condemned can be used to condemn them. In its defective use, $L_2^\star$ calls for the impossible, the existence of a self-referential statement; in its non-defective use it does not.

A solution of this kind ought to give rise to at least two kinds of worry. One is specific: could the ingenious opponent not devise Liar sentences for which this response is ruled out (see question 6.10)? The other is more general: is the view consistent with the possibility of genuinely *formal* logic, logic in which the logical relations are mirrored by merely syntactic ones? If there is any looseness of fit, unamenable to theory, between sentences and statements, this project will be doomed.

6.10

*Show how the response appears to be inadequate to:
No use of this very sentence expresses a true statement.

The notion of indexicality can be exploited in a different way. We have explored the possibility that indexicality might affect the subject term in sentences such as L_2 and L_2*: whether or not the subject refers in a way demanding the existence of a self-referential statement was held to be a function of the circumstances of its use. One should also consider the possibility that indexicality affects the predicate term "true." Tyler Burge (1979) has developed a suggestion on these lines. There are different levels of truth, and which level is at issue is fixed not by the meaning of the sentence, but by the statement it is used to make on a given occasion. This approach based on indexical levels avoids many of the difficulties associated with Tarski's hierarchy, which attaches to sentences rather than to statements. Burge's construction is complex, and I give no details here because I have the following suspicion: that it is hard or impossible to justify the claim that "true" is indexical, independently of the apparent need for it to be indexical to do justice to the paradoxes, whereas we have independent reason to believe indexicality affects self-reference.

6.9 Comparison: how similar are Russell's paradox and the Liar?

Are the two paradoxes of this chapter totally different, or essentially the same? Is the truth perhaps somewhere between these extremes?

Ramsey urged that the paradoxes are different in kind, and his view has been predominant, at least until relatively recently. He based the distinction on their different subject matter: the logical paradoxes, under which heading he included Russell's paradox, arise from logical notions, like that of class; the semantic paradoxes, under which he included the Liar, arise from semantic notions, like that of truth.

There are also structural dissimilarities, which might be traced to the difference in the concepts involved in the different paradoxes. There is no analog in Russell's paradox of the Strengthened Liar. There is an immediate problem with the idea that *there is no* statement expressed by L_2, namely that it seems to follow that L_2 is not true. No such twist is consequent on the assertion that *there is no* class *R*. For the class paradox, there is quite widespread agreement on what we need to say – that there is

no class *R*. What is unclear is how this can be justified. For the Liar, it is not clear even what ought to be said, let alone how to justify it.

The two paradoxes are also similar in many ways. I enumerate five.

(1) The Class paradox resembles a paradox about properties, and the Property paradox in turn resembles the Liar. Most properties are not true of themselves. For example, the property of being a man is not true of itself, since that property lacks the property of being a man; but the property of being a non-man is true of itself, since the property of being a non-man has the property of being a non-man. The pattern of reasoning used in the Class paradox would lead to the conclusion:

The property of *being not true of itself* is true of itself if and only if it is not true of itself.

There is at least a surface similarity between this contradiction and the contradiction that L_2 truly predicates truth of itself if and only if it does not. Where the Property contradiction uses the notion of *not true of*, a relation that may hold between a property and something else (perhaps also a property), the Liar contradiction uses the notion of *not true*, a property that a sentence or statement may possess.

(2) Both the Class paradox and the Liar involve self-reference, or some similar circularity.

(3) The principles appealed to in the derivation of the two paradoxes (CE: for every intelligible condition *F*, there is a class *x*, such that: for any object *y*, $y \in x$ if and only if *y* satisfies *F*; and T: σ is true iff *p*) are structurally similar, and appear to play similarly constitutive roles with respect to the intuitive notions of *class* and *truth*.

On the side of derivation, the comparison is this. The schema

for any object *y*, $y \in x$ iff *y* is *F*

yields a contradiction when *x* is replaced by a name, say *R*, for the Russell class, and *F* by the condition, expressed in terms of this name, that supposedly defines membership for that class, "$\neg \in R$." Similarly, the schema

σ is true iff *p*

yields a contradiction when *σ* is replaced by a name, say "L_2," for the Liar sentence, and *p* by the condition, expressed in terms of this name, that supposedly defines truth for that sentence, "L_2 is not true."

On the side of roles, the comparison is that just as CE appears constitutive of our pretheoretical notion of a class, so T appears constitutive of our pretheoretical notion of truth. CE determines what it is for a class to exist; T determines what it is for a truth condition to exist.

(4) Hierarchies have been used in response to both kinds of paradox, beginning with one of the earliest systematic treatments, in Russell (1908). It is natural to suppose that we should think of classes as constructed out of non-classes, with each constructional step drawing only upon entities which have already been constructed. Likewise, it is natural to suppose that we should think of statements ascribing truth as constructed out of statements free of the notion of truth, with each constructional step drawing only upon statements which have already been constructed.

(5) Russell's classification of the Class paradox and the Liar as of a common kind is based on the claim that they both alike derive from an infringement of the Vicious Circle Principle:

VCP: No totality can contain members fully specifiable only in terms of itself.

We have already seen how this might be invoked to ban certain kinds of circularity in statements. The way in which it bans circularity in classes is more straightforward. The specification of the class R, of non-self-membered classes, went like this:

For any class x, $x \in R$ iff $\neg x \in x$.

The specification speaks of what Russell would call a totality: the totality of all classes, introduced by the phrase "any class." Russell takes it that this is the only possible specification of R, and intends that the VCP tell us that R cannot belong to the totality introduced by the phrase "any class," as that occurs in the definition of R. For suppose R did belong to that totality; then the totality would contain a member R specifiable only in terms of that totality, which is what the VCP says is impossible. However, if R does not belong to the totality introduced by "any class," then we cannot make the usual move to the contradiction. The usual move goes like this: if the definition holds of *any class*, then in particular it holds of R, so we can infer

$R \in R$ iff $\neg R \in R$.

The VCP has it that the totality introduced by *any class* excludes R, so this reasoning is fallacious. In effect, the upshot of the VCP is that we cannot specify R as we had originally intended – in such a way, that is, that the question could arise about whether it has or lacks its defining property.

A formal vindication of Russell's claim that the paradoxes belong to a significant common kind has been provided by Priest (1994).

The Class and Liar paradoxes, like most things, are similar in some respects, dissimilar in others. So what? Classification matters here because of the constraints it imposes on proper responses to the paradoxes. (Many

other paradoxes need to find their place in a classification: see Priest 1994.) If two paradoxes are *essentially* similar, similar in what really matters, then it is proper to respond in essentially similar ways. For example, if the Class paradox calls for a hierarchy of levels, and the Liar is essentially similar, then it too calls for a hierarchy of levels. If an adequate conception of classes should allow nonwellfounded classes, e.g. the class α whose only member is α, and the Liar is essentially similar, then our response to it should allow for analogous circularity. Russell's allegedly common solution to the two paradoxes, his Ramified Theory of Types, has a somewhat specious uniformity, since some of its more complex features are motivated by matters relating to the Liar rather than the Class paradox.

To infer that paradoxes require a uniform solution, we must show more than that there is *a* kind to which the paradoxes belong. We have also to show that this kind reveals their common essential nature. I doubt, however, whether this can be done quite independently of views about what response is appropriate to each.

Suggested reading

Section 6.1

Russell's earliest published exposition of the Class paradox is in his *Principles of Mathematics* (1903). For a historical overview, see van Heijenoort (1967). Most introductory logic texts have an account. For a fairly non-technical account, with the paradoxes firmly in mind, see Copi (1971). For a good and accessible introduction to set theory see Thomason (1970, chapter 13). A good online resource is Kevin Klement's (2006) encyclopedia article: www.iep.utm.edu/p/par-russ.htm.

Current usage distinguishes between sets and classes: all sets are classes, but not all classes are sets. Intuitively, a set is a well-behaved class. The distinction presupposes a certain approach to the solution of Russell's paradox, and so is not appropriate to the present discussion.

The *power class* of a class is the class consisting of every subclass of the class. Consider, for example, the class consisting of just the three elements *a, b*, and *c*, which we can write $\{a, b, c\}$. A class α is a subclass of a class β if and only if every member of α is a member of β. The class $\{a, b, c\}$ has the following subclasses: (1) $\wedge$ (the null class) – since $\wedge$ has no members, every member of $\wedge$ is a member of $\{a, b, c\}$ (2) $\{a\}$; (3) $\{b\}$; (4) $\{c\}$; (5) $\{a, b\}$; (6) $\{a, c\}$; (7) $\{b, c\}$; and (8) $\{a\ b\ c\}$ – since each member of this class is a member of $\{a, b, c\}$.

There are thus eight subclasses of a class with three members. The class of all subclasses of $\{a, b, c\}$ is more numerous (by five members) than

$\{a, b, c\}$ itself. Cantor's theorem holds obviously for classes with finitely many members. The interest of the theorem consists in the fact that it applies to classes of every cardinality.

A one–one function between two classes, α and β, associates each member of α with exactly one member of β, and each member of β with exactly one member of α. Cantor takes it that two classes have the same number of members (the same cardinality) if and only if there is a one–one function between them.

A good introduction to both kinds of paradoxes discussed in this chapter is Priest (1987): chapter 2 (set theoretic paradoxes) and chapter 1 (semantic paradoxes).

Section 6.2

Good starting points for the Liar: Mackie (1973); Prior (1961); Martin (1984, editor's introduction); Barwise and Etchemendy (1987, chapter 1). An example of more recent work is Smith (2006). A recent overview, with a useful bibliography, is Bradley Dowden's (2007) encyclopedia article at www.iep.utm.edu/p/par-liar.htm.

For discussions of bivalence see Haack (1978) and Burge (1984).

Section 6.3

I borrow the word "grounding" and its cognates from Kripke (1975). I do not claim to have captured what he means by it, for his concept of grounding is embedded in a mathematical theory to which I cannot begin to do justice. However, the early part of his paper is perfectly accessible to the non-mathematician. A classic reference for grounding is Herzberger (1970).

Section 6.4

The expression "Strengthened Liar" originates with van Fraassen (1968), though the problem itself is much older.

Section 6.5

Despite the technical nature of the main body of Tarski's classic paper (1956 [1937]), the first section is non-technical, accessible, and well worth reading. He sets out a condition of adequacy for a formal definition of truth (symbolized "Tr"), called Convention T, as follows:

A formally correct definition of the symbol "Tr" … will be called an … *adequate definition of truth* if it has the following consequences:

(α) all sentences which are obtained from the expression "x ε Tr if and only if p" by substituting for the symbol "x" a structural-descriptive name of any sentence of the language in question and for the symbol "p" the expression which forms the translation of this sentence into the metalanguage;

(β)… (1956, pp. 187–8)

As it is sometimes put, Tarski allows us to draw on our intuitive conception of meaning (translation) in specifying the conditions for a correct definition of truth.

For each truth predicate in the hierarchy, Tarski accepts every instance of T. The fact that T has an inconsistent instance, if instances can be formed from any grammatically acceptable construction of English – in particular L_2 – was taken in the early paper (1937) to show that there is no such coherent language as English. Replacing the single colloquial language predicate "true" by a hierarchy of truth predicates also involved, in Tarski's eyes, doing away with colloquial language, as normally understood. "In [colloquial] language it seems to be impossible to define the notion of truth or even to use this notion in a consistent manner and in agreement with the laws of logic" (1937, reprinted in 1956: 153).

In later writing, he takes a more gentle line with ordinary language. Referring to his hierarchical view as the "semantic conception," he writes: "I happen to believe that the semantic conception [of truth] does conform to a very considerable extent with the commonsense usage" (1944, p. 360). Tarski (1969) provides a semi-popular exposition.

For a careful statement of Tarski's precise premises, together with a challenge to the full generality of the conclusion Tarski drew, see Gupta (1982, section II). For a recent discussion of Tarskian hierarchies, see Glanzberg (2004).

Section 6.6

For an attack on self-reference, see Jørgensen (1953). For the claim that some kinds of self-reference are innocuous, see Barwise and Etchemendy (1987, esp. pp. 15–16). For a collection on the topic, see Bartlett (1992). The example of the three speakers, A, B, and C is adapted from an example given by Burge (1979, pp. 175–6) and attributed by him to Buridan.

Section 6.7

The classic source of the distinction between sentence and statement, advanced for reasons quite independent of the Liar, is Strawson (1950).

Section 6.8

The formulation of the VCP given on p. 139 is verbally closest to Russell (1908, p. 75). The justification for "only" comes from the formulation at p. 63 of that work; cf. Russell and Whitehead (1910–13, p. 37). I have used "fully specifiable" rather than Russell's "defined."

Hazen (1987) traces the origin of responses of the general kind considered in this section to the fourteenth-century philosopher, Jean Buridan. Hazen defends a version of the view considered here, and the approach has been taken by others, for example Goldstein (1992). For criticisms, see Hinckfuss (1991), Smiley (1993) (though he advances an account which centers on the idea that Liar sentences "malfunction"), and Priest (1993).

To introduce nonfoundational set theory, we think of a set in terms of a diagram. Thus the set consisting just of London and the set whose members are the number 7 and Mount Everest (conventionally written: {London, {7, Everest}}) can be represented by figure 6.1:

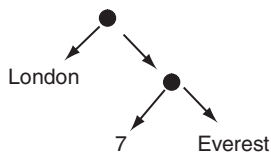


Figure 6.1

Here each blob represents a set, and the branches beneath it represent its members. Figure 6.2 would then represent the set α whose only member is α:

Figure 6.2

The theory of such diagrams is consistent relative to classical set theory. A general account of the evils of circularity, like Russell's VCP, would do well to try to break the link between the diagrams and sets.

Nonfoundational set theory owes a great deal to Aczel (1987). A good account, well adapted to present concerns, is in Barwise and Etchemendy (1987), ch. 3.

Section 6.9

Russell (1908) argues for the common nature of the Class, Liar and other paradoxes. For the distinction between logical and semantic paradoxes, see Ramsey (1925, pp. 171–2). The best exposition of the case for a single family of paradoxes is Priest (1994). Glanzberg (2003) uses dissimilarities between Russell's paradox and the Liar paradox to argue against a "minimalist" view of truth.

7 Are any contradictions acceptable?

In previous chapters, I have at many points argued that if something leads to a contradiction, then either it, or the relevant reasoning, must be rejected. The assumption that one must always reject contradictions has come under attack at various times in the history of philosophy. Recently, the attack has taken a subtle form and has harnessed impressive technical resources.

I shall discuss the following views:

(1) Some contradictions are true.
(2) For some contradictions, it is rational to believe that they are true.

The only version of (1) that I shall consider also holds that every contradiction is false; this is "dialetheism." However, the view that some contradictions are both true and false does not add up to the view that some contradictions are acceptable, for one might go on to insist that anything perceived to be false should be rejected. If so, the assumption of the previous chapters of this book would remain unchallenged: we should reject anything which leads to a contradiction. So in this chapter I will discuss the conjunction of views (1) and (2). I shall call this combination "rational dialetheism."

The rational dialetheist's claim that some contradictions are true is, under natural assumptions, equivalent to the claim that some sentences are both true and false. Any such sentence may be called a *dialetheia*. I shall suggest in section 7.4 that, given natural assumptions, rational dialetheism entails that some sentences which are not contradictions (that is, are not of the form A and not-A) are both true and false.

The main positive case for dialetheism is that there is no better response to various paradoxes, notably, but not only, the Liar and Russell's paradox, than simply to accept the contradictions in question as true. So the main case for dialetheism is also a case for rational dialetheism. My discussion of the paradoxes of classes and of truth will no doubt have persuaded you that every solution faces difficulties; so you can imagine, in

a general way, how one might make the case for simply accepting a contradiction. It is agreed by all parties that accepting a contradiction as true is a last resort. Dialetheists argue that we are forced to it; their opponents, that nothing could force us to it.

Once they have entered their positive case, in terms of the detailed discussion of various possible responses to paradoxes, dialetheists have to ward off objections. These can be divided into those that tell against both dialetheism and rational dialetheism, and those that tell only against the latter. Thus a case for rejecting all falsehoods could count against rational dialetheism without necessarily counting against dialetheism. However, a case against dialetheism is obviously a case against rational dialetheism. With some distress, I come to the conclusion that none of the objections I review ought to force a resourceful rational dialetheist to admit defeat.

7.1 Contradictions entail everything

Classical logic validates the inference rule (sometimes referred to by the scholastic term "ex contradictione quodlibet," more recently called "explosion"): from a contradiction, anything may be derived. This is quite unproblematic for the classicist, since he holds that no contradictions are true. Although arguments with contradictory premises are classically valid, none of them are sound: we could never use them to prove anything, since we could never establish the premises.

As soon as one allows that there is even one true contradiction, explosion ensures that one has a *sound* argument for any arbitrary proposition. One would be committed to holding that everything is true. This conclusion is absurd, and is accepted by dialetheists as absurd. They accordingly reject the classical inference rule.

Whatever one's final judgment upon dialetheism, one can but be impressed by how little need be lost by this modification of classical logic (Priest 1987).

Dialetheists need make no adjustment to the informal conception of validity: a valid argument is one that precludes the possibility of moving from truth to falsehood. In the classical picture, this conception validates explosion, since if a contradiction cannot be true, an argument whose premises include a contradiction is one whose premises cannot all be true, and thus is one which rules out the possibility of the argument leading from truth to falsehood. Dialetheists, by contrast, do not assume that a contradiction cannot be true, and so are not committed to holding that every inference with a contradiction among its premises is valid.

7.2 A sentence which is both true and false could have no intelligible content

The meaning or content of a declarative sentence can be thought of in terms of which states of affairs would make it true, and which states of affairs would make it false. To understand a sentence, we must find out what it rules out, and what would rule it out. There are limiting cases: tautologies rule out nothing, and some sentences, ones which cannot be true, rule out everything. However, were there a sentence which is both true and false, there could be no coherent understanding of it, for there would be no determinate fact concerning what it ruled out, and what ruled it out. One would have to say: there is a state of affairs which the sentence both rules in and rules out. But this is like saying that the state of affairs both is and is not determined by the content of the sentence, which shows that the sentence has no coherent content at all.

These considerations are certainly not decisive. One way to bring this out is to rework them in terms of possible worlds semantics, as follows.

The content of a declarative sentence is given by two disjoint sets: the set of worlds at which it is true, and the set of worlds at which it is false. (This is one way of expressing a version of the view that the meaning of a sentence consists in the conditions under which it is true.) Suppose, for some sentence, the actual world belongs to the set of worlds at which it is true. Then it is not the case that the actual world belongs to the set of worlds at which it is false, since the set of worlds at which it is false is, by hypothesis, disjoint from the set of those at which it is true.

It is very plain that this kind of possible worlds semantics builds in the assumption that no sentence can be true and false, and provides not a shadow of an argument for it. It is not the mere fact of using possible worlds in the semantics which delivers the result. Rather, it is the assumption of disjointness of the set of worlds at which a sentence is true and the set of those at which it is false. There is nothing in the apparatus to prevent a sentence being associated by the semantics with a set of worlds at which it is true which overlaps the set of worlds at which it is false. To argue that no such association should be made requires philosophical work, which is not done merely by alluding to the apparatus.

If we now revert to the earlier formulation, in terms of what a sentence rules out, and what rules it out, the assumption which is hostile to dialetheism is that no meaningful sentence can rule something in which rules it out. This requires justification. It is not as if we cannot use the notions of ruling in and ruling out, just because of this overlap. The fault in the earlier argument is the assumption that if a sentence is both true and false, so that it rules in a state of affairs which rules it out, there is no "determinate" fact about what it rules in and what rules it out. The fact is determinate

enough, but a fact at odds with this anti-dialetheist presupposition: if a state of affairs is ruled out by a sentence, it is not also ruled in by it.

7.3 Three dualities

Assertion is an act, typically linguistic. When sincere, it expresses a mental state of acceptance on the part of the asserter toward what he asserts. Denial is likewise an act. When sincere, it expresses a mental state of rejection on the part of the denier toward what he denies. We have three dualities which, in the best circumstances, seem as if they ought to line up: true/false; assert/deny; accept/reject. The alignment would be as follows: what is true is what is to be asserted and is what is to be accepted; what is false is what is to be denied and what is to be rejected.

According to Priest, if one does not accept *A*, there are two possibilities: to reject *A*, which he characterizes as *refusing* to accept *A*, or to be agnostic about *A*, neither accepting it nor refusing to accept it. Rejection and acceptance are thus not exhaustive; but he claims that they are exclusive: one cannot do both with respect to a single proposition at a single time. Yet, on his view, truth and falsehood are not exclusive: a single proposition may be both true and false at the same time. Are these views cotenable?

They confront a problem, if the following (or a close relative thereof) is true:

F: Anything false should be rejected.

For then all contradictions should be rejected (since they are all false), rejecting them makes it impossible to accept them (since rejection excludes acceptance), and it is hard to see how it could be rational to accept what it is impossible to accept.

F needs modification. It is not rational to reject what is in fact false, if all the available evidence suggests that it is true. We might modify F to "Anything recognized as false should be rejected." Such modifications do not bear on the present discussion, so I shall pass them by.

The argument based on F may not be decisive, since perhaps we frail creatures are simply unable to do what an ideally rational agent would do, so what is "hard to see" may none the less be the case. However, there is some pressure upon the rational dialetheist to abandon F. This is the route taken by Priest, who points out that, for the rational dialetheist, rejecting all falsehoods would be incompatible with another ideal, that of accepting all truths. As he puts it: "Truth and falsity come inextricably intermingled … One cannot, therefore, accept all truths and reject all falsehoods" (Priest 1986, p. 106; cf. Priest 1987, p. 124). What one should do is to reject all falsehoods which are not also truths. This is a rule of conduct

with which anti-dialetheists cannot quarrel, since from their perspective it is tantamount to F.

Priest allows that most of us, at one time or another, believe contradictions, that is, for some *A*, we believe that *A* and also believe that not-*A*. It might seem that this will lead to trouble. If, for some *A*, we accept *A* and also accept not-*A*, we could sincerely assert not-*A*. So we could sincerely deny *A*. So we reject *A*. So there is a proposition we both accept and reject.

Priest disputes this reasoning, for he does not allow that one can infer that *A* is to be rejected from the premise that not-*A* is to be accepted. We will examine this issue more closely in the next section. Here I make two observations.

The first is that the exclusive character of acceptance and rejection is the impossibility of a person being in a state of acceptance and state of rejection with respect to a given proposition at a given time. This exclusiveness is not obviously inconsistent with it being demanded by rationality that one both accept and reject a given proposition at a given time. Ought does not imply can. Priest at one juncture appears favorably inclined to this strategy (1993, pp. 40–1). Its wholehearted endorsement would lead to some modifications to his earlier approach, as he himself notes.

The second observation is that Priest does not supply an argument for the exclusive character of acceptance and rejection, and one might be inclined to suppose that there exists an overlooked version of dialetheism, in which the attitudes are held to be compatible, and F is allowed.

The prospects for this version, however, are not bright, if the aim of providing a reasonable response to some of the paradoxes is to be retained. The idea is that when we come to a dialetheia, a proposition both true and false, we can accept it, and thus do not have to criticize the reasoning which led to it. If we were asked in addition to reject the dialetheia, then presumably it would be incumbent upon us to criticize the reasoning: reasoning from premises which must be accepted to a conclusion which must be rejected cannot be good. So adopting this strategy would leave dialetheists with at best a partial response to the paradoxes; they would have to go on, like everyone else, to say what is wrong with the reasoning.

This point also has an impact upon the version of dialetheism considered in the first observation. If we allow that we ought, rationally, to reject dialetheias as well as accept them, then we ought, rationally, to criticize the reasoning which led to them, and we would no longer have a satisfactory response to the paradoxes. In what follows, I shall accordingly treat dialetheists as holding to the exclusiveness of acceptance and rejection in a double sense: it is impossible for a person to be in both states with respect to one proposition at one time; and it is impossible that someone should be rationally required to be in both states with respect to one proposition at one time.

7.4 Negation

The dialetheist needs to say something about how negation relates to truth and falsehood, how it relates to acceptance and rejection, and how it relates to a proper account of our understanding of signs which express negation.

The first issue can be raised by considering an objection:

> Contradictions are conjunctions, having the form "*A* and not-*A*." A conjunction is true iff both conjuncts are. But "not-*A*" is true iff "*A*" is false. Hence we cannot have both "*A*" and "not-*A*" true; so no contradiction can be true.

The dialetheist can accept the premises of this argument (the one about conjunction, and the one about negation) and yet deny the conclusion. Suppose that "*A*" is false (as well as true). Then, by the rule for negation, "not-*A*" is true (as well as false). So, on the supposition, both conjuncts are true (as well as false). So, by the rule for conjunction, the conjunction is true (as well as false).

This shows that dialetheists must take it that some dialetheias are not explicitly of the form "*A* and not-*A*." (Examples are likely to include "Russell's class is a member of itself" and L_1 ("L_1 is false").) If there are any true contradictions, there must be some of the form "*A* and not-*A*" where "*A*" is restricted to non-contradictions. But if "*A*" has at most one truth value, "not-*A*" must have just the other truth value if it has any, so one must be other than true, so the contradiction could not be true. This reasoning depends upon the principles about conjunction and negation which formed the premises of the argument with which this section opened. We can conclude that the essence of dialetheism is that there are dialetheias; that there are true contradictions is, upon natural assumptions, a consequence.

As is plain from the response to the objection, Priest's view of the relation between negation and truth and falsehood coincides with that of the classical logician: not-*A* is true iff *A* is false; not-*A* is false iff *A* is true. However, the classical logician relates negation to acceptance and rejection by the following:

N: Something should be rejected iff its negation should be accepted.

Smiley (1993, p. 20) has claimed that N is in part constitutive of the concept of negation. Yet N cannot be accepted by dialetheists, given that they hold that rejection and acceptance are exclusive. If a contradiction should be accepted, so should both conjuncts, *A* and not-*A*. But if not-*A* should be accepted, it follows from N that *A* should be rejected, and so, given exclusiveness, that *A* should not be accepted.

Priest accepts part of N, namely:

N1: If something should be rejected, its negation should be accepted.

He rejects the other part of N, namely;

N2: Something should be rejected if its negation should be accepted.

The classical logician affirms N2 and the dialetheist denies it. As far as I can see, the position here is a stand-off, neither side being able to produce a non-question-begging argument for their view. Indeed, the position is, at least for a Priestian dialetheist, equivalent to the debate about F ("anything false should be rejected"). For if *A*'s negation is accepted, one is committed to accepting that *A* is false, so, by F, one should reject *A* (thus establishing N2 from F). If *A* is false, then it has a true negation, and so, according to Priest, a negation which should be accepted, and so, by N2, *A* itself should be rejected (thus establishing F from N2).

The questions of how negation relates to truth and falsehood, and to rejection and acceptance, are distinct from the final question about negation to be considered in this section: what is it for a person to understand some expression (e.g. "not" or "it is not the case that") as a sign for negation? The idea is to provide an account of what the person does by virtue of which he can be counted a party to the practice of using the expression in question as a sign of negation.

Here is one such account, which is hostile to dialetheism. To understand a sign for negation (e.g. "it is not the case that") is to know, for any sentence *A*, never to treat both "*A*" and "it is not the case that *A*" as true (Stephanou 1994). If this were a correct account, it would seem that dialetheists would manifest a failure to understand negation. A dialetheist who treats *A* as being both true and false will treat both *A* and "it is not the case that *A*" as true. This is, at a minimum, in tension with knowing never to do this.

Doubtless, dialetheists will dispute the suggestion about what understanding a sign for negation consists in. But they face some questions: can we give an account of understanding in terms of dispositions to treat sentences in certain ways? If so, what should be said about understanding negation? If not, how can we explain what it is to understand negation? It is far from obvious in advance that answers to these questions will be consistent with dialetheism (see question 7.1).

7.1

Might the dialetheist identify understanding a sign for negation with knowing, for any sentence *A*, always to treat one of "*A*" and "it is not the case that *A*" as false?

7.5 Falsehood and untruth

For any predicate φ, we take it that we can form one which holds of just the things of which φ does not hold; or, at least, one which holds of just the things within the significance range of φ of which φ itself does not hold. Let us stipulate that "un" is an operator which has this effect. (We need not worry about whether this operator really surfaces in English, or whether it should count as a negation operator.) Then the set of happy things is disjoint from the set of unhappy things; the set of holy things disjoint from the set of unholy things; and the set of true things disjoint from the set of untrue things.

If we may stipulate thus, then we may form a Liar sentence for which dialetheism provides no adequate response:

L_3: This very sentence is untrue.

We cannot simply accept the reasoning which leads to the conclusion that the sentence is both true and untrue, for these, by hypothesis, are predicates with disjoint extensions. So even a dialetheist must criticize the reasoning. Dialetheism as such does not even address this paradox.

The objection is that even if we cannot disprove dialetheism by proving that "true" and "false" do not overlap, we can prove, because we can stipulate, that "true" and "untrue" do not overlap. This allows us to create a paradox, closely similar to familiar Liar paradoxes (perhaps identical to the Strengthened Liar), which the dialetheist cannot address. This does not show that dialetheism is incorrect. It is consistent with the view that some contradictions are true. But it does show, supposedly, that allowing that some contradictions are true will not do all the work that dialetheists had hoped it would. Moreover, given the similarities among the semantic paradoxes, it suggests that another approach will need to be found to all of them.

The dialetheist might simply accept this objection, stressing that it relates not to the truth of dialetheism, but only to the extent of its application. Alternatively, he might resist the charge. "L_3 is true and L_3 is untrue" is, to be sure, a contradiction; but it is one for which we have powerful arguments, so why should it not be one of the true contradictions? If "L_3 is true" is false, then "L_3 is untrue" can be true; so if, in addition, "L_3 is true" is true, we have what is required for a true contradiction.

This reply in effect commits the dialetheist to the claim that truth and untruth are non-exclusive, as well as exclusive. Priest agrees (1987, pp. 90–1). Arguing that L_3 shows that some sentences are true as well as untrue (thus establishing non-exclusiveness), he also argues that from the

law of excluded middle, which he accepts, one can derive that everything is either true or untrue, which entails that nothing is both true and untrue (thus establishing exclusiveness). Thus he accepts that the machinery he brings to bear upon the nature of contradictions is itself contradictory.

Whether this is tolerable is too large an issue for me to address. However, I will close with one observation on behalf of the dialetheist. I opened this section by imagining a stipulation to the effect that "untrue" would be a predicate with an extension disjoint from "true." I hoped my readers would accept this proposal without a qualm, and thus take it that there could be no question of overlap between "true" and "untrue." My closing remark is simply that confidence of this kind is not in general justified. Wanting our stipulations to be consistent is not enough to ensure that they are, as will be obvious to anyone familiar with legal stipulations. Our use of "true" is governed by certain principles, and there is no *apriori* guarantee that these principles permit a non-contradictory stipulation of a disjoint predicate.

The discussion has been inconclusive: I have not been able to show that dialetheism is incorrect. If it is correct, then it should be considered a candidate in connection with most of the paradoxes discussed in this book.

An inconclusive discussion may, none the less, be worthwhile. As Priest says: "whether or not dialetheism is correct, a discussion of the questions it raises, concerning fundamental notions like negation, truth and rationality … can hardly fail to deepen our understanding of these notions" (Priest 1993, p. 35).

Suggested reading

Graham Priest's contribution to the *Stanford Encyclopedia of Philosophy* is probably the best place to start: plato.stanford.edu/entries/dialetheism/.

For the early history of dialetheism, see the editors' introduction to Priest *et al.* (1989). The word was coined by Priest and Routley in the early 1980s. The most important reference is Priest (1987). His defense of dialetheism involves some excellent discussions of many paradoxes relating to classes and truth, since motivating his own view requires showing that alternative responses to them are inadequate. See also his symposium with Smiley (1993), which I have used as the basis of much of this chapter; and Rescher and Brandom (1980). The state of the art of these topics is represented by Priest *et al.* (2004).

The expression "rational dialetheism" was suggested to me by Priest (personal communication).

Its being rational to hold dialetheism does not entail rational dialetheism, any more than its being rational to believe that Tom loves someone entails, concerning some person, that it is rational to believe that Tom loves that person.

Horn (1989) suggests that there are different kinds of negation. This idea would need to be considered in a full account of the relation between negation and rejection.

Appendix I: Some more paradoxes

(An asterisk before a title indicates that there is an observation on the entry in Appendix II.)

THE GALLOWS

The law of a certain land is that all who wish to enter the city are asked to state their business there. Those who reply truly are allowed to enter and depart in peace. Those who reply falsely are hanged. What should happen to the traveler who, when asked his business, replies, "I have come to be hanged"?

*BURIDAN'S EIGHTH SOPHISM

Socrates in Troy says, "What Plato is now saying in Athens is false." At the same time, Plato in Athens says, "What Socrates is now saying in Troy is false." (Cf. Buridan, in Hughes 1982, pp. 73–9).

THE LAWYER

Protagoras, teacher of lawyers, has this contract with pupils: "Pay me a fee if and only if you win your first case." One of his pupils, Euathlus, sues him for free tuition, arguing as follows: "If I win the case, then I win free tuition, as that is what I am suing for. If I lose, then my tuition is free anyway, since this is my first case."

Protagoras, in court, responds as follows: "If you give judgment for Euathlus, then he will owe me a fee, since it is his first case and that was our agreement; if you give judgment for me, then he will owe me a fee, since that is the content of the judgment."

THE DESIGNATED STUDENT

Five students are told by the teacher that all of them are to have a star pinned on their backs; that just one of the stars is gold – the recipient of this is the "designated student"; and that the designated student will not know he or

she is designated. The students are lined up so that the fifth can see the backs of the other four, the fourth the backs of the other three, and so on.

They argue that what the teacher said cannot be true, for the following reasons:

> The fifth student can infer that he cannot be unknowingly designated, since, if he were designated, he could see, from the non-gold nature of the stars on the other students, that none of them is designated, and thus could infer that he himself is designated.
>
> The fourth student can infer that (a) the fifth student cannot be unknowingly designated and (b) the fourth cannot be either, since, given that the fifth has not been designated, the fourth would be able to infer, from the non-gold nature of the three visible backs, that he was designated, if he was … and so on.

Is this a genuine paradox? Is it a version of the Unexpected Examination? (See Sorensen 1982.)

*THE GRID

The following paradox has been said to be structurally like the Unexpected Examination. Is it? Does it contain a serious paradox?

In the Grid game, you are blindfolded and placed on a grid with numbered squares as shown table AI.1:

Table AI.1

1	2	3
4	5	6
7	8	9

The double outer line represents a wall. You are allowed to move only horizontally or vertically, and you may attempt only two moves from your initial position. Your aim is to determine which square you are on. You might be lucky. For example, if you attempt a move right and feel the wall, then attempt a move down and feel the wall, you can infer that you are placed on square 9. However, you might be unlucky. For example, if you moved left twice without reaching the wall, you could not tell whether your initial position was 3, 6, or 9.

Suppose I claim that I can put you in an initial position not discoverable in two moves. However, you reason as follows: I cannot be put in any of the corner squares, since there are two-move sequences (like the one mentioned for square 9) that would tell me where I am; but if 1, 3, 7, and 9 can be eliminated, so can 2, 4, 6, and 8, since, for example, a move

up into the wall would tell me that I was at 2, given the elimination of 1 and 3 as possibilities. Hence my initial position must be 5, and so I *can* discover my initial position – and in zero moves! (See Sorensen 1982.)

*THE STONE

Can an omnipotent being make a stone so heavy that he cannot lift it? He can, because, being omnipotent, he can do everything. He also cannot, since, if he did make it, there would be something he could *not* do – that is, lift it. Reading: Savage (1967); Schrader (1979); and, for a variant, Mele and Smith (1988).

HETEROLOGICAL

Call an expression "heterological" if and only if it does not describe itself. Thus "long" is heterological because

"long" is long

is false; but "short" is not heterological since

"short" is short

is true. Is "heterological" heterological or not?

We could shorten the definition to the following schema, abbreviating "heterological" as "het":

het("φ") iff ¬φ("φ").

The contradiction follows immediately by taking "het" as the replacement for the schematic φ. See Russell (1908), Quine (1966, esp. pp. 4ff., "Grelling's Paradox").

*THE LOTTERY

Suppose there are a thousand tickets in a lottery and only one prize. It is rational to believe of each ticket that it is very unlikely to win. Hence it must be rational to believe that it is very unlikely that any of the thousand tickets will win – that is, rational to believe that it is very unlikely that there will be a winning ticket.

THE PREFACE

Knowing one's frailties as one does, it is rational to suppose that one's book contains errors, and it is not unknown for authors to say as much in their prefaces. However, a sincere author will believe everything asserted in the text. Rationality, plus modesty, thus forces such an author to a contradiction. (Cf. Makinson 1965.)

THE PREFACE AGAIN

Suppose an author's preface consists solely in this remark: "At least one statement in this book is false." Then the body of the book must contain at least one false statement. For suppose it does not. Then if the preface is true, it is false, and if it is false, it is true; which is impossible. (Cf. Prior 1961, pp. 85–6.)

THE INFALLIBLE SEDUCER

An unsuccessful wooer was advised to ask his beloved the following two questions:

(1) Will you answer this question in the same way that you will answer the next?
(2) Will you sleep with me?

If she keeps her word, she must answer Yes to the second question whatever she has answered to the first.

This paradox is amusingly generalized in Storer (1961).

BURIDAN'S TENTH SOPHISM

Suppose that:

A is thinking that $2 + 2 = 4$.
B is thinking that dogs are reptiles.
C is thinking that an odd number of the current thoughts of A, B, and C are true.

Is what C thinks true or not? See Buridan, in Hughes (1982, p. 85), Prior (1961), Burge (1978, p. 28).

FORRESTER'S PARADOX

Suppose Smith is going to murder Jones. It is obligatory that if he murders Jones, he should do so gently. This appears to imply that if Smith murders Jones, it is obligatory that he do so gently. However, he cannot murder Jones gently without murdering him. Hence, given that Smith is going to murder Jones, it is obligatory that he do so. (See Forrester 1984.)

THE CHOOSER

Someone you trust implicitly presents you with a choice: you can take either or both of box A or box B. Whatever happens, there is $100 in

box *B*; moreover, there will be in addition $10,000 in box *A* if and only if you choose irrationally. What should you do? Cf. Gaifman (1983).

BERTRAND'S PARADOX

What is the probability that a random chord of a circle exceeds the side of an inscribed equilateral triangle? It is longer if its midpoint lies on the inner half of the radius bisecting it; so, since the midpoint may lie anywhere on this radius, the probability is one-half. It is also longer if its midpoint lies within a concentric circle with half the original radius; so, since the area of this inner circle is a quarter that of the original circle, the probability is one-quarter.

*THIS IS NONSENSE

Line 1: The sentence written on Line 1 is nonsense.
Line 2: The sentence written on Line 1 is nonsense.

For a suitable interpretation of "nonsense," we incline to believe that the sentence on Line 2 is true: the sentence it refers to is viciously self-referential, deserves to fall in the truth-value gap, or whatever. Yet the sentence on Line 2 is the very sentence it so justly criticizes.

The example comes from Gaifman (1983).

*THE PENNY GAME

A game for two players. Playing in turn, either player may take one or two pennies from a pile before them. Any penny taken belongs to the player who takes it. Any pennies left in the pile when the game is over vanish. The game stops either when there are no pennies left in the pile, or after a player takes two pennies in one turn.

The paradoxical conclusion is that if both players are rational, and know it (of themselves and of each other) the first player will take two pennies, thus ending the game. The conclusion is paradoxical because one would think that two rational people ought to end up splitting the pot, for then both would be better off. (And if rationality does not help you satisfy your desires, what is the point of it?)

An argument for the conclusion goes as follows. Suppose there are just two pennies left at your turn. If you take one penny, then the other player will take the last penny, and the game will end, with you gaining just one penny. This is less good than if you had taken two pennies. The game would end then, but you would have gained one penny rather than two.

Now suppose that there are three pennies left. If you take just one penny, you will leave the other player with two pennies. You know that the other player is rational, and so will do the rational thing, and you have just worked out that the rational thing, faced with two pennies, is to take two. So if, faced with three pennies, you take just one, you know that the game will end after the next move, and you will gain only one penny. By contrast, if you take two pennies now, you will gain two pennies rather than one.

The argument iterates backwards, regardless of how many pennies there are in the pile. See Hollis and Sugden (1993).

MONTY HALL PARADOX

In a game show, the contestant knows that, of the three closed doors in front of him, *A*, *B*, and *C*, there is a car behind one and nothing behind the other two. He is invited to choose one of the doors, knowing that he gets to keep anything that lies behind the selected door. He wants to choose the door with the car.

After he has made his selection – let us say he chooses door *A* – the host opens one of the two remaining doors, say door *B*. There is nothing behind it. The host then asks the contestant whether he would like to switch the choice he has made, from *A* to *C*. Should he switch?

There is no point, since nothing that has happened has given him any new information. He now has a 50 percent chance of winning, whether he switches or not, whereas before he had only a one-third chance. Switching cannot affect the odds, so there is no point.

He should switch, since doing so will lead to a win two-thirds of the time. His original choice was either of the door with the car, or of one of the two doors without. That gives three possibilities. In the first, switching leads to a loss. But in *both* the other two cases, switching leads to a win. Switching doubles the chance of winning. See: en.wikipedia.org/wiki/Monty_Hall_problem.

BERRY'S PARADOX

Russell (1908, pp. 222–4) lists a number of paradoxes, including one he attributes to "Mr. G. G. Berry of the Bodleian Library":

> "'the least integer not nameable in fewer than nineteen syllables' is itself a name consisting of eighteen syllables; hence the least integer not nameable in fewer than nineteen syllables can be named in eighteen syllables, which is a contradiction." (1908, p. 223).

A NEW ZENOIAN PARADOX

This comes from Bernadete (1964), as quoted by Priest (1999, p. 1):

> A man decides to walk one mile from A to B. A god waits in readiness to throw up a wall blocking the man's further advance when the man has traveled half a mile. A second god (unbeknown to the first) waits in readiness to throw up a wall of his own blocking the man's further advance when the man has traveled a quarter of a mile. A third god … etc. ad infinitum. It is clear that this infinite sequence of mere intentions (assuming the contrary-to-fact conditional that each god would succeed in executing his intentions if given the opportunity) logically entails the consequence that the man will be arrested at point A; he will not be able to pass beyond it, even though not a single wall will in fact be thrown down in his path. The … [effect] will be described by the man as a strange field of force blocking his passage forward.

FITCH'S "PARADOX" ABOUT KNOWABILITY

Verificationists believe that anything that is so can in principle be known to be so. You may not agree with them, but it seems not to be a silly view, easily refutable. If that is right, the following refutation should at least be surprising, whether or not strictly paradoxical.

We shall use "K(p)" for "it is known that p" and "$\Diamond(p)$" for "it's possible that p." The verificationist view is then

V: if p then $\Diamond$ K(p).

Even verificationists do not suppose we are omniscient, so they will accept that there is at least one unknown truth, say A:

$A \,\&\, \neg$ K(A).

V is meant to be entirely general, and so should apply to this, yielding

$\star$ $\quad \Diamond$ K($A \,\&\, \neg$ K(A)).

But it seems this cannot be true: it entails the possibility that A is known even though it is not known. Showing in more detail that ($\star$) is false requires two principles about knowledge: that whatever is known is true (if K(p) then p) and that knowledge distributes over conjunction (if K(p & q) then K(p) & K(q)). Given these principles, together with two principles about necessity noted below at (5) and (6), a proof of the falsehood of ($\star$) is:

(1)	K(A & ¬ K(A))	[assumed for reductio]
(2)	K(A) & K¬ K(A)	[distribution of K]
(3)	K(A) & ¬ K(A)	[applying "whatever is known is true" to right conjunct]

(4) $\neg$ K(A & $\neg$ K(A)) [(1) led to a contradiction at (3)]
(5) $\Box\neg$ K(A & $\neg$ K(A)) [Necessitation applied to (4): what is proved is necessary]
(6) $\neg \Diamond$ K(A & $\neg$ K(A)) [by definition, "necessarily not" means "not possibly"]

Fitch's original statement of the crucial claim is given as theorem 5: "If there is some true proposition which nobody knows (or has known or will know) to be true, then there is a true proposition which nobody can know to be true" (1963, p. 139). See Berit Brogaard and Joe Salerno's (2004) encyclopedia entry at plato.stanford.edu/entries/fitch-paradox/.

QUINE'S PARADOX

Does "Yields a falsehood when appended to its own quotation" yield a falsehood when appended to its own quotation (Quine 1966, p. 7)? How is this related to "heterological"? Does it involve self-reference? Does it involve indexicality?

THE CABLE GUY PARADOX

The Cable Guy is coming to install your new cable service sometime between 8 a.m. and 4 p.m. (Annoyingly, the company said it could not specify the time more narrowly.) To while away the time, you consider whether it would be best to bet on his coming before noon or after.

It is an equal bet, just like tossing a coin. You have no information to base a preference for the morning or the afternoon.

You should bet on the Guy coming in the afternoon. Suppose you bet on his coming in the morning. It is very likely he will not be there by 8.15. By then, it would clearly be best to switch your bet (were this allowed) to the afternoon, since the afternoon time stretch is now bigger, and given what you know, any stretches of time of the same duration are equally likely to be times at which the Guy comes. It is irrational to make a bet which it is highly likely you will almost immediately regret, wishing you had made the other. See Hájek (2005).

Appendix II: Remarks on some text questions and appended paradoxes

Remark numbers refer to question numbers in the given chapter.

1. ZENO'S PARADOXES: SPACE, TIME, AND MOTION

1.2 One possible argument is this. If there were as many as two things, say α and β, then we could consider the whole W formed by these two things. Then W has α and β as its parts. So if nothing has parts, there are not as many as two things, i.e. there is at most one thing.

1.6 Yes, it does mean that the button will travel faster than the speed of light. Whether or not this is a logical possibility could be disputed, but it is fairly uncontroversial that it is not impossible *apriori;* that is, reasoning alone, unaided by experiment, cannot establish that nothing can travel faster than light.

1.8 "Going out of existence at $Z^{\star}$" might mean "$Z^{\star}$ was the last point occupied" or, alternatively, "$Z^{\star}$ was the first point not occupied." The latter serves Benacerraf's cause against the objection.

3. VAGUENESS: THE PARADOX OF THE HEAP

3.11 He should deny the first premise. Since he believes that there are no heaps, he will think that it is false that a 10,000-grained collection is a heap. This shows that one should treat with care the taxonomy of responses described in the text. Unger's position, though in a sense unified, must be regarded as accepting the conclusion of some soritical arguments and denying the premises of others.

4. ACTING RATIONALLY

4.3 If a person's utilities can be measured in cash value terms, then his or her utilities are "commensurable": of any two possible circumstances, either one has more utility than the other, or else they are of equal utility. However, if we think of very disparate "utilities," it may be not

merely that we do not *know* how they compare, but that there is no such thing as how they compare. For an early defense of commensurability see Rashdall (1907, ch. 2). For a more recent discussion see Nussbaum (1986, esp. pp. 107ff.).

4.4 It would seem that it could not as it stands, for the following reason. The outcome of a gamble, in cash terms, does not register the fact that it is obtained by gambling. A given sum has the utility it has, whether earned or won. So what would be needed, to register a dislike of gambling, would be a "higher order" conception of utility. The expected utilities delivered by the MEU as it stands would be subjected to a weighting, which would augment expected utilities whose probability component is high, and diminish those whose probability component is low.

4.5 The example shows that the dominance principle does not issue in guidance about which actions are rational, unless it is combined with instructions about how the possible outcomes should be divided up.

4.17 Each reasons as follows: "In the *last* game it will be best for me to confess, since any loss of trust this induces will be irrelevant, as we are not going to 'play' again." However, the other can work this out and will adopt the same strategy: "Since he will confess in the *last* game, I should confess in the penultimate game, since any loss of trust this may induce will be irrelevant" (and so on).

5. BELIEVING RATIONALLY

5.4 It might be that the conditions for evidence being good could not be known to obtain.

5.6 "All emeralds are green" and "All emeralds are grue" are not, strictly, inconsistent. If there were no unexamined emeralds, both generalizations would be true; whereas it is impossible for genuinely inconsistent propositions both to be true.

A given body of evidence can "point both ways", i.e. can provide grounds for two propositions that are genuinely inconsistent. In such cases, the evidence can normally be divided into evidence that supports the one proposition, and evidence that supports the other, without overlap (at least, without total overlap). What is paradoxical about the grue case is that no such division is possible: divide the evidence as fine as you like, every bit that confirms "All emeralds are green" also confirms "All emeralds are grue."

5.16 No. EB1 is simply silent about what beliefs you have concerning what you don't believe. (David Lewis pointed this out in correspondence.)

6. CLASSES AND TRUTH

6.3 Suppose L_1 is not false, that is, suppose not-L_1. (Principle: since L_1="L_1 is false" we may substitute one for the other. For an attack on the use of this principle, see Skyrms (1982).) Then not-L_1 is, by supposition, true. So L_1 is false. (Principle: a sentence is false if its negation is true.) So if L_1 is not false, it is false. So it is false. But since this is what L_1 says it is, it is true. So it is both true and false.

6.5 Suppose L_1 is false; then it tells it the way it is, so it is true, and hence *L_1 is not false*. (Principle: Anything true is not false.) Cf. Martin (1984, pp. 2–3).

6.10 Call the sentence "No use of this very sentence expresses a true statement" S. Consider an arbitrary use U_1 of S. If U_1 succeeds in making a true statement, then it does not make a true statement (since it says that *no* use of S expresses a truth). So U_1 does not make a true statement. But if this were generally true, of all uses of S, U_1 would be true after all. So there must be a U_2 which does make a true statement. But this leads to a contradiction. Cf. Hazen (1987); for a criticism, see Hinckfuss (1991); see also Smiley (1993).

APPENDIX I

BURIDAN'S EIGHTH SOPHISM

Roy Sorensen (2001, p. 180) has pointed out that earlier printings of the second edition (and late printings of the first edition) reported Buridan's sophism incorrectly, representing Plato as saying that what Socrates is saying is *true*. This makes for a familiar Liar paradox. Buridan's Eighth Sophism is paradoxical in a different way: it seems that we can establish that Socrates and Plato said things with opposite truth values, one true and the other false, but there appears to be no fact about which one is true and which one false.

THE GRID

The alleged paradox seems to turn on an equivocation between whether there is a sequence of moves that fixes one's position, and whether every possible sequence does so. There is no corresponding equivocation in the Unexpected Examination. (But see Sorensen 1982.)

THE STONE

No, an omnipotent being cannot make a stone so heavy that he cannot lift it. He will still be omnipotent with respect to making and lifting stones, however, if for any weight of stone (in grams, megatons, or whatever) he

can make one of that weight and lift one of that weight. However, see Mele and Smith (1988). Suppose one power possessed by any omnipotent being is the power to shed her omnipotence. Does that suggest a different solution to the paradox?

THE LOTTERY

One suggestion is that this shows that one may have good reason for believing that *A* and good reason for believing that *B*, yet not have good reason for believing that *A* and *B*. This suggestion would need to be supplemented by a treatment of sorites-style reasoning based on: "If one has good reason to believe a proposition with probability *n*, then one has good reason to believe a proposition with probability minutely smaller than *n*."

THIS IS NONSENSE

The example supports the view that two sentence tokens of a single sentence type can differ in truth value (one true, the other not) even if both refer to the same thing and predicate the same property of it.

THE PENNY GAME

This has the structure of a multiple Prisoner's Dilemma which is known by the parties to have finitely many iterations. Compare question 4.17 (p. 88), and the comment above (p. 169).

Bibliography

Aczel, Peter (1987) *Lectures on Nonwellfounded Sets*. Stanford, CA, CSLI Lecture Notes, no. 9.

Anand, Paul (1993) *Foundations of Rational Choice Under Risk*. Clarendon Press, Oxford.

Aristotle (1970) *Physics*, trans. W. Charlton. Oxford University Press, Oxford.

Asher, Nicholas M. and Kamp, Johan A. W. (1986) "The knower's paradox and representational theories of the attitudes." In J. Halpern (ed.), *Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufman, New York, pp. 131–48.

Axelrod, R. (1984) *The Evolution of Cooperation*. Basic Books, New York.

Bar-Hillel, Maya and Margalit, Avishai (1972) "Newcomb's paradox revisited." *British Journal for the Philosophy of Science* 23: 295–304.

(1983) "Expecting the unexpected." *Philosophia* 13: 263–88.

(1985) "Gideon's paradox – a paradox of rationality." *Synthese* 63: 139–55.

Bartlett, Steven J. (1992) *Reflexivity: A Source-Book in Self-Reference*. North-Holland, Amsterdam and London.

Barwise, Jon and Etchemendy, John (1987) *The Liar: An Essay in Truth and Circularity*. Oxford University Press, New York and Oxford.

Benacerraf, Paul (1962) "Tasks, super-tasks, and the modern Eleatics." *Journal of Philosophy* 59: 765–84. Reprinted in Salmon (1970), pp. 103–29.

Benditt, T. M. and Ross, David J. (1976) "Newcomb's paradox." *British Journal for the Philosophy of Science* 27: 161–4.

Bernadete, J. (1964) *Infinity: An Essay in Metaphysics*. Clarendon Press, Oxford.

Black, Max (1937) "Vagueness: an exercise in logical analysis." *Philosophy of Science* 4: 427–55. Reprinted in his *Language and Philosophy*. Cornell University Press, Ithaca, NY, 1949, pp. 25–58.

Brogaard, Berit and Joe Salerno (2004) "Fitch's paradox of knowability." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/fitch-paradox/.

Burge, Tyler (1978) "Buridan and epistemic paradox." *Philosophical Studies* 34: 21–35.

(1979) "Semantical paradox." *Journal of Philosophy* 76: 169–98. Reprinted in Martin (1984), pp. 83–117.

(1984) "Epistemic paradox." *Journal of Philosophy* 81: 5–29.

Buridan, John *Sophismata*. In Hughes (1982).

Campbell, Richmond and Sowden, Lanning (eds.) (1985) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. University of British Columbia Press, Vancouver.

Cargile, J. (1965) "The sorites paradox." *British Journal for the Philosophy of Science* 20: 193–202.
Copeland, J. (1994) "On vague identity, fuzzy objects and fractal boundaries." *Southern Journal of Philosophy* 33: 83–96.
Copi, Irving M. (1971) *The Theory of Logical Types*. Routledge and Kegan Paul, London.
Dowden, Bradley (2007) "Liar paradox." In *The Internet Encyclopedia of Philosophy*, available at: www.iep.utm.edu/p/par-liar.htm.
Dummett, Michael (1975) "Wang's paradox." *Synthese* 30: 301–24. Reprinted in his *Truth and Other Enigmas*. Duckworth, London. 1978, pp. 248–68.
Edgington, Dorothy (1992) "Validity, uncertainty and vagueness." *Analysis* 52.4: 193–204.
Edwards, Paul (1967) *The Encyclopedia of Philosophy*, 8 vols. Collier-Macmillan and Free Press, New York.
Evans, Gareth (1978) "Can there be vague objects?" *Analysis* 38: 208. Reprinted in his *Collected Papers*. Oxford University Press, Oxford. 1985, pp. 176–7.
Fine, Kit (1975) "Vagueness, truth and logic." *Synthese* 30: 265–300.
Fitch, F. (1952) *Symbolic Logic*. Ronald Press Company, New York.
(1963) "A logical analysis of some value concepts." *Journal of Symbolic Logic* 28: 135–42.
Forrester, James William (1984) "Gentle murder and the adverbial Samaritan." *Journal of Philosophy* 81: 193–7.
Foster, John (1983) "Induction, explanation and natural necessity." *Proceedings of the Aristotelian Society* 83: 87–101.
Foster, Marguerite and Martin, Michael L. (eds.) (1966) *Probability, Confirmation and Simplicity*. Odyssey Press, New York.
Gaifman, Haim (1983) "Paradoxes of infinity and self-application, I." *Erkenntnis* 20: 131–55.
Gale, Richard M. (ed.) (1968) *The Philosophy of Time*. Macmillan, London.
Gensler, Harry J. (1986) "A Kantian argument against abortion." *Philosophical Studies* 49: 83–98.
Gibbard, A. and Harper, W. L. (1978) "Counterfactuals and two kinds of expected utility." In C. A. Hooker, J. J. Leach, and E. F. McClennon (eds.), *Foundations and Applications of Decision Theory*, vol. I. Reidel, Dordrecht, pp. 125–62. Reprinted (abridged) in Campbell and Sowden (1985), pp. 133–58.
Glanzberg, M. (2003) "Minimalism and paradoxes." *Synthese* 135: 13–36.
(2004) "Truth, reflection, and hierarchies." *Synthese* 142: 289–315.
Goguen, J. A. (1969) "The logic of inexact concepts." *Synthese* 19: 325–78.
Goldstein, Laurence (1992) "'This statement is not true' is not true." *Analysis* 52.1: 1–5.
Goodman, Nelson (1955) *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, MA; 2nd edn, Bobbs-Merrill, Indianapolis, IN, 1965.
(1978) *Ways of Worldmaking*. Hackett, Indianapolis, IN.
Gowans, Christopher W. (ed.) (1987) *Moral Dilemmas*, Oxford University Press, New York.
Graff-Fara, D. G. (2000). "Shifting sands: an interest-relative theory of vagueness." *Philosophical Topics* 28: 45–81. Originally published under the name "Delia Graff."

Grünbaum, Adolf (1967) *Modern Science and Zeno's Paradoxes*. Wesleyan University Press, Middletown, CT.

Gupta, Anil (1982) "Truth and paradox." *Journal of Philosophical Logic* 11: 1–60. Reprinted in Martin (1984), pp. 175–235.

Haack, Susan (1978) *The Philosophy of Logics*. Cambridge University Press, Cambridge.

Hájek, Alan (2005). "The Cable Guy paradox." *Analysis* 65: 112–19.

Hazen, Allen (1987) "Contra Buridanum." *Canadian Journal of Philosophy* 17: 875–80.

Hempel, Carl (1945) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York, reprinted 1965. The material on the Ravens Paradox is reprinted from *Mind* 54: 1–26, 97–121 (1945).

Herzberger, Hans A. (1970) "Paradoxes of grounding in semantics." *Journal of Philosophy* 67: 145–67.

Hinkfuss, Ian (1991) "Pro Buridano; contra Hazenum." *Canadian Journal of Philosophy* 21.3: 389–98.

Hollis, Martin and Sugden, Robert (1993) "Rationality in action." *Mind* 103: 1–35.

Horn, L. R. (1989) *A Natural History of Negation*. University of Chicago Press, Chicago.

Huggett, Nick (2004) "Zeno's paradoxes." In E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/paradox-zeno/.

Hughes, G. E. (ed. and trans.) (1982) *John Buridan on Self-Reference*. Cambridge University Press, Cambridge and New York. This work by Buridan was originally entitled *Sophismata*.

Hume, David (2000 [1738]) *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Hyde, Dominic (1994) "Why higher-order vagueness is a pseudo-problem." *Mind* 103: 35–41.

(2005) "Sorites paradox." In E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/sorites-paradox/.

Jackson, Frank (1975) "Grue." *Journal of Philosophy* 72: 113–31.

Janaway, Christopher (1989) "Knowing about surprises: a supposed antinomy revisited." *Mind* 98: 391–409.

Jeffrey, Richard C. (1965) *The Logic of Decision*. McGraw-Hill, New York.

Jørgensen, Jørgen (1953) "Some reflections on reflexivity." *Mind* 62: 289–300. Reprinted in Bartlett (1992), pp. 63–74.

Kamp, Hans (1975). "Two theories about adjectives." In E. Keenan (ed.), *Formal Semantics of Natural Language*. Cambridge, Cambridge University Press, pp. 123–55.

(1981) "The paradox of the heap." In U. Monnich (ed.), *Aspects of Philosophical Logic*. Reidel, Dordrecht, pp. 225–77.

Kavka, Gregory (1987) *Moral Paradoxes of Nuclear Deterrence*. Cambridge, Cambridge University Press.

Keefe, R. (2000). *Theories of Vagueness*. Cambridge, Cambridge University Press.

Klement, Kevin (2006) "Russell's paradox." *The Internet Encyclopedia of Philosophy*, available at: www.iep.utm.edu/p/par-russ.htm.

Koons, Robert (1992) *Paradoxes of Rationality*. Cambridge, Cambridge University Press.

Kripke, Saul (1975) "Outline of a theory of truth." *Journal of Philosophy* 72: 690–716. Reprinted in Martin (1984), pp. 53–81.

(1982) *Wittgenstein on Rules and Private Language*. Blackwell, Oxford.

Kuhn, Steven (2007): "Prisoner's dilemma." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/prisoner-dilemma/.

Kyburg, Henry (1961) *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, CT.

Levi, Isaac (1967) *Gambling with Truth*. Routledge and Kegan Paul, London.

Lewis, David (1979) "Prisoner's Dilemma is a Newcomb problem." *Philosophy and Public Affairs* 8: 235–40. Reprinted in Campbell and Sowden (1985), pp. 251–5.

(1986) *On the Plurality of Worlds*. Blackwell, Malden, MA.

(1988) "Vague identity: Evans misunderstood." *Analysis* 48: 128–30.

Ludwig, K. and Ray, G. (2002). "Vagueness and the sorites paradox." *Philosophical Perspectives* 16: 419–61.

McConnell, Terrance (2006) "Moral dilemmas." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/archives/sum2006/entries/moral-dilemmas/.

Mackie, J. L. (1973) *Truth, Probability and Paradox*. Oxford University Press, Oxford.

(1977) "Newcomb's paradox and the direction of causation." *Canadian Philosophical Review* 7: 213–25. Reprinted in his *Collected Papers II*. Oxford: Oxford University Press, 1986.

MacFarlane, John (2007). "Fuzzy epistemicism." Available at: http://johnmacfarlane.net/fuzzy-epistemicism.pdf.

Makinson, D. C. (1965) "The paradox of the preface." *Analysis* 25: 205–7.

Marcus, Ruth Barcan (1980) "Moral dilemmas and consistency." *Journal of Philosophy* 77: 121–36.

Martin, Robert L. (ed.) (1984) *Recent Essays on Truth and the Liar Paradox*. Oxford University Press, Oxford.

Mele, Alfred R. and Smith, M. P. (1988) "The new paradox of the stone." *Faith and Philosophy* 5: 283–90.

Mellor, D. H. (1971) *The Matter of Chance*. Cambridge University Press, Cambridge and New York.

Montague, Richard and Kaplan, David (1960) "A paradox regained." *Notre Dame Journal of Formal Logic* 1: 79–90. Reprinted in Richmond Thomason (ed.), *Formal Philosophy*. Yale University Press, New Haven, CT, 1974, pp. 271–85.

Nozick, R. (1969) "Newcomb's problem and two principles of choice." In Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Reidel, Dordrecht. Abridged version reprinted in Campbell and Sowden (1985), pp. 107–33.

(1993) *The Nature of Rationality*. Princeton University Press, Princeton, NJ.

Nussbaum, Martha C. (1986) *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press, Cambridge and New York.

Parfit, Derek (1984) *Reasons and Persons*. Oxford University Press, Oxford.
Peacocke, C. A. B. (1981) "Are vague predicates incoherent?" *Synthese* 46: 121–41.
Peirce, C. (1935) *The Collected Papers of Charles Sanders Peirce*, ed. Charles Hartshorne and Paul Weiss, 8 vols. Harvard University Press, Cambridge, MA.
Priest, Graham (1986) "Contradiction, belief and rationality." *Proceedings of the Aristotelian Society* 86: 99–116.
(1987) *In Contradiction*. Nijhof, Dordrecht.
(1993) "Can contradictions be true?" *Supplementary Proceedings of the Aristotelian Society* 67: 35–54.
(1994) "The structure of the paradoxes of self-reference." *Mind* 103: 25–34.
(1999) "On a version of one of Zeno's paradoxes." *Analysis* 59: 1–2.
(2004) "Dialetheism." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/dialetheism/.
Priest, Graham, Routley, Richard and Norman, Jean (eds.) (1989) *Paraconsistent Logic: Essays on the Inconsistent* Philosophia Verlag, Munich.
Priest, Graham, Beall, J. C. and Armour-Garb, B. (eds.) (2004) *New Essays on the Law of Non-Contradiction*, Oxford University Press, Oxford.
Prior, Arthur N. (1961) "On a family of paradoxes." *Notre Dame Journal of Formal Logic* 2: 16–32.
Quine, Willard van O. (1953) "On a so-called paradox." *Mind* 62: 65–7. Reprinted in Quine (1966), pp. 19–21.
(1966) *Ways of Paradox and Other Essays*. New York, Random House. (The title essay is on pp. 1–18.)
Raffman, D. (1994). "Vagueness without paradox." *Philosophical Review* 103: 41–74.
Ramsey, Frank P. (1925) "The foundations of mathematics." Reprinted in D. H. Mellor (ed.), *Foundations*. Humanities Press, Atlantic Highlands, NJ, 1978, pp. 152–212.
(1926) "Truth and probability." Reprinted in D. H. Mellor (ed.), *Foundations*. Humanities Press, Atlantic Highlands, NJ, 1978, pp. 58–100.
Rashdall, Hastings (1907) *The Theory of Good and Evil*, vol. II. Clarendon Press, Oxford.
Rescher, N. (2001) *Paradoxes: Their Roots, Range and Resolution*. Open Court, Chicago.
Rescher, N. and Brandom, R. (1980) *The Logic of Inconsistency*. Blackwell, Oxford.
Russell, Bertrand (1903) *The Principles of Mathematics*. Cambridge University Press, Cambridge and New York.
(1908) "Mathematical logic as based on the theory of types." *American Journal of Mathematics* 30: 222–62. Reprinted in R. C. Marsh (ed.), *Logic and Knowledge*. Allen and Unwin, London. 1956, pp. 59–102.
(1936) "The limits of empiricism." *Proceedings of the Aristotelian Society* 36: 131–50.
Russell, Bertrand and Whitehead, Alfred North (1910–13) *Principia Mathematica*, 3 vols. Cambridge University Press, Cambridge.
Sainsbury, R. M. and Williamson, Timothy (1995) "Sorites." In Bob Hale and Crispin Wright (eds.), *Blackwell Companion to the Philosophy of Language*. Blackwell, Oxford.

Salmon, Nathan U. (1982) *Reference and Essence*. Blackwell, Oxford.
Salmon, Wesley C. (1970) *Zeno's Paradoxes*. Bobbs-Merrill, Indianapolis, IN.
(1980) *Space, Time and Motion: A Philosophical Introduction*, 2nd edn. University of Minnesota Press, Minneapolis.
Sanford, David (1975) "Borderline logic." *American Philosophical Quarterly* 12: 29–39.
(1976) "Competing semantics of vagueness: many values versus super-truth." *Synthese* 33: 195–210.
Savage, C. Wade (1967) "The paradox of the stone." *Philosophical Review* 76: 74–9.
Schiffer, Stephen (2003) *The Things We Mean*. Clarendon Press, Oxford.
Schlesinger, G. (1974a) *Confirmation and Confirmability*. Oxford University Press, New York.
(1974b) "The unpredictability of free choice." *British Journal for the Philosophy of Science* 25: 209–21.
Schrader, David E. (1979) "A solution to the stone paradox." *Synthese* 42: 255–64.
Scriven, Michael (1951) "Paradoxical announcements." *Mind* 60: 403–7.
Selton, Reinhard (1978) "The chain store paradox." *Theory and Decision* 9: 127–59.
Skyrms, Brian (1982) "Intensional aspects of semantical self-reference." In Martin (1984), pp. 119–31.
Smilansky, Saul (2007) *Ten Moral Paradoxes*. Blackwell, Malden, MA.
Smiley, Timothy (1993) "Can contradictions be true?" *Supplementary Proceedings of the Aristotelian Society* 67: 17–33.
Smith, Nicholas J. J. (2006) "Semantic regularity and the Liar paradox." *The Monist* 89: 178–202.
Sorensen, Roy A. (1982) "Recalcitrant variations of the prediction paradox." *Australasian Journal of Philosophy* 60: 355–62.
(1988) *Blindspots*. Clarendon Press, Oxford.
(2001) *Vagueness and Contradiction*. Clarendon Press, Oxford.
(2003) *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind*. Oxford University Press, Oxford.
(2006a) "Vagueness." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/vagueness/.
(2006b) "Epistemic paradoxes." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/epistemic-paradoxes/.
Stephanou, Ioannis (1994) "The meaning of the logical constants and the justification of logical laws." University of London PhD thesis.
Storer, Thomas (1961) "MINIAC: world's smallest electronic brain." *Analysis* 22: 151–2.
Strawson, Peter (1950) "On referring." *Mind* 59: 269–86. Reprinted in his *Logico-Linguistic Papers*. Methuen, London. 1971, pp. 1–27.
Tarski, Alfred (1944) "The semantic conception of truth: and the foundations of semantics." *Philosophy and Phenomenological Research* 4: 341–76.
(1956 [1937]) "The concept of truth in formalized languages." In his *Logic, Semantics, Metamathematics*. Clarendon Press, New York, pp. 152–278; first published in Polish in 1933.
(1969) "Truth and proof." *Scientific American* 194: 63–77.

Thomason, Richmond H. (1970) *Symbolic Logic: An Introduction*. Macmillan, Toronto.
Thomson, James F. (1954) "Tasks and super-tasks." *Analysis* 15: 1–13. Reprinted in Gale (1968), pp. 406–21, and in Salmon (1970), pp. 89–102.
Tye, Michael (1990) "Vague objects." *Mind* 99: 535–57.
(1994a) "Sorites paradoxes and the semantics of vagueness." In James Tomberlin (ed.), *Philosophical Perspectives: Logic and Language*. Ridgeview, Atascadero, CA.
(1994b) "Why the vague need not be higher-order vague." *Mind* 103: 43–5.
Unger, Peter (1979a) "There are no ordinary things." *Synthese* 41: 117–54.
(1979b) "I do not exist." In Graham MacDonald (ed.), *Perception and Identity: Essays Presented to A. J. Ayer*. Macmillan, London, pp. 235–51.
Van Fraassen, B. (1966) "Singular terms, truth-value gaps, and free logic." *Journal of Philosophy* 53: 481–5.
(1968) "Presupposition, implication and self-reference." *Journal of Philosophy* 65: 136–52.
Van Heijenoort, John (1967) "Logical paradoxes." In Edwards (1967), vol. V, pp. 44–51.
Vickers, James (2006) "The problem of induction." In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: plato.stanford.edu/entries/induction-problem/.
Vlastos, Gregory (1967) "Zeno of Elea." In Edwards (1967), vol. VIII, pp. 369–79.
Wiggins, David (1986) "On singling out an object determinately." In Philip Pettit and John McDowell (eds.), *Subject, Thought, and Context*. Oxford University Press, Oxford, pp. 169–80.
Williams, Bernard (1966) "Ethical consistency." *Supplementary Proceedings of the Aristotelian Society* 40: 1–22.
Williamson, Timothy (1992a) "Inexact knowledge." *Mind* 101: 218–42.
(1992b) "Vagueness and ignorance." *Supplementary Proceedings of the Aristotelian Society* 66: 145–62.
(1994) *Vagueness*. Routledge, London.
(1999) "On the structure of higher-order vagueness." *Mind* 108: 127–43.
(2003) "Vagueness in reality." In M. Loux and D. W. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*. Oxford University Press, Oxford, pp. 690–715.
Wittgenstein, Ludwig (1953) *Philosophical Investigations*. Basil Blackwell, Oxford.
Wright, Crispin (1975) "On the coherence of vague predicates." *Synthese* 30: 325–65.
(1987) "Further reflections on the sorites paradox." *Philosophical Topics* 15: 227–90.
(1992) "Is higher order vagueness coherent?" *Analysis* 53.3: 129–39.
(1995) "The epistemic conception of vagueness." *Southern Journal of Philosophy* (Supplement) 33: 133–59.
(2001) "On being in a quandary. Relativism vagueness logical revisionism." *Mind* 110: 45–98.
Wright, Crispin and Sudbury, Aidan (1977) "The paradox of the unexpected examination." *Australasian Journal of Philosophy* 60: 41–58.
Zadeh, L. A. (1965) "Fuzzy sets." *Information and Control* 8: 338–53.

Index