successful, would show at most that *A cannot rationally* prevent *B* from defecting. It is still possible to maintain that *A can* bring it about that *B* revises her beliefs, for *A* can act irrationally. What is impossible is that *A* act rationally *and* bring about belief revision.

The new line of reasoning offers a promising route for reinstating the BIA, and merits further study.

Some final brief remarks. Part of what makes the iterated prisoner's dilemma so baffling is that there seems to be a significant reward for acting irrationally. If *A* were to act irrationally, she would, let us suppose, force *B* to give up her belief that *A* is rational, and defeat *B*'s reasons for defecting at later stages of the game. Clearly, it is in *A*'s best interest that *B* not defect on future rounds. So in such a case it seems that one ought to act irrationally. But it is contradictory to suppose that it is rational to act irrationally.

The iterated prisoner's dilemma, viewed in this way, is reminiscent of Gideon's paradox. Imagine that you have a choice between \$1,000 (G_1) and \$100 (G_2). You prefer more money to less money. Before you choose, a bystander offers you a reward of \$1,000,000 for acting irrationally. So it seems you ought to act irrationally. That is, the choice you make between G_1 and G_2 will be rational if and only if it is irrational.[23]

One point of similarity between the two situations is that it is tempting in both to come to the conclusion that it is rational to act irrationally. A second common feature is that, in each case, the consequences of an act depend, in part, on whether the act is irrational or not. Normally, however, it is assumed that the rationality of an act is a function of its possible consequences. It is hard to see how to assess rationality if the possible consequences of the alternative acts are not already given.

To sum up, it has been argued in this section that the survival of the belief in rationality fits nicely with other features of the dilemma, indeed follows from them, and that it is irrational for a player to attempt belief revision. Although the original formulation of the BIA requires a belief retention premise asserting the survival of the belief in rationality come what may, there appears to be an alternative line of reasoning available that does not depend on this implausibly strong assumption. The BIA has not been defeated; the paradox of the iterated prisoner's dilemma is with us still.

8 The sorites paradox

The paradox

Suppose that your height is 6'3". Clearly, you are tall. Your friend Tom, however, is only 5'4" and is concerned about his height. You offer him the following philosophical argument. If two people differ in height by only 0.1", you point out, then either both are tall or neither is. A difference in height of 0.1" cannot make the difference between being tall and not being tall. So a person who is 0.1" shorter than you are (6'3" – 0.1") is tall. But then a person who is 0.1" shorter than that (6'3" – 0.2") is also tall. You continue in this way until you reach 5'4", and can reassure Tom that he is indeed tall. Your friend is neither comforted nor amused.

This sort of paradoxical reasoning can be traced back to the logician Eubulides, a contemporary of Aristotle. It is sometimes referred to as "the paradox of the heap", since it is commonly illustrated in terms of a heap of sand. ("Sorites" derives from "soros", the Greek term for "heap".) Suppose you have a heap of sand consisting of 10,000 grains. Removing one grain surely cannot turn a heap into something that is not a heap. So 10,000 – 1 grains constitute a heap. But then so do 10,000 – 2 grains. Continuing with this reasoning will eventually lead to the absurd conclusion that one grain of sand suffices for a heap.

Another classic illustration of the sorites reasoning makes use of the concept of baldness. Suppose that as well as being tall, you are blessed with a luxuriant mop of hair; it is clear that you are not bald. There are, let us suppose, *n* hairs on your head. You reason that the difference between being bald and not being bald cannot

consist in one hair. Thus $n-1$ hairs suffice for not being bald. But by the same reasoning, $n-2$ hairs is sufficient. Finally, you conclude that a man with one hair on his head is not bald.

In each of these examples, the argument hinges on the presumed "tolerance" or "insensitivity" of the key term. That is, a minute difference cannot make a difference, cannot affect the applicability of the predicate. It is not difficult to see that sorites arguments can be constructed for a variety of other predicates: child (consider a difference in age of one minute); red (consider a series of coloured cards ranging in colour from red to yellow, ordered in such a way that any two adjacent cards are indiscriminable in colour). Still other predicates that have been considered susceptible to sorites reasoning, although not always obviously so, include: rich, intelligent, small number, most, moustache, table, tadpole and dead. Sorites reasoning can also involve moral predicates. Grant that it would be morally wrong to have an abortion on (or after) the nth day of pregnancy (or the mth minute, where m is the number of minutes in n days). But if it is wrong at the mth minute, then surely it is also wrong at the $(m-1)$th minute. Thus it is morally wrong, at any point in a pregnancy, to have an abortion; abortion is never morally acceptable.[1]

Sorites reasoning can be cast in two different forms. First:

(A)
(1) A person who is worth \$$M$ is rich.
(2) If a person worth \$$M$ is rich, then so is a person worth \$$(M-1)$.
∴ (3) A person worth \$$(M-1)$ is rich.
(4) If a person worth \$$(M-1)$ is rich, then so is a person worth \$$(M-2)$.
∴ (5) A person worth \$$(M-2)$ is rich.
⋮
(n) A person worth \$1 is rich.

In this form, the argument consists in a sequence of sub-arguments, each of the form *modus ponens*, and each containing a different categorical and conditional premise. The first step of the argument is straightforward and rarely disputed. The sequence of conditionals, which is more likely to draw critical fire, reflects the view that

the predicate in question is insensitive to small differences. An insensitive or tolerant predicate is not fine-grained, not responsive to fine discriminations; but of course, as the argument shows, enough minute differences can add up to a large difference that *is* sufficient to affect the applicability of the predicate.

Sorites arguments can also be expressed in the more compact schema of mathematical induction:

(B)
(1) A person who has m hairs on his head is not bald.
(2) For any x, if a person with x hairs on his head is not bald, then neither is a person with $x-1$ hairs on his head.
∴ (3) A person with one hair on his head is not bald.

Premise (1) is the base step and premise (2) the inductive step. Given suitable values for m, both premises seem clearly true, yet the conclusion is absurd.

As presented in either of these forms, the sorites paradox is a type I falsidical paradox, and is uncontroversial (although there are a few individuals who consider the argument to be sound, as we will see). For the most part, the paradox will be treated below in the form given by (A).

One intriguing feature of the sorites paradox is that, whether expressed as a series of *modus ponens* inferences or as a mathematical induction, it can be stood on its head. Argument (B), for instance, is absurd because a person with only one hair on his head is clearly bald. But if we start with this obvious truth, another sorites argument can be constructed:

(C)
(1) A person with one hair on his head is bald.
(2) For any x, if a person with x hairs on his head is bald, so is a person with $x+1$ hairs on his head.
∴ (3) A person with m hairs on his head is bald.

Supposing that m is the number of hairs on Bill Clinton's head, the conclusions of (C) and (B) are equally absurd. Every sorites argument seems to be reversible in this way: a paradoxical argument can be formulated as a positive sorites argument (in terms of being bald) or as a negative sorites argument (in terms of not being bald). The

positive and negative versions are so closely analogous that it seems they must stand or fall together. On the face of it, then, it is an uphill road for those few commentators who want to accept *only* the negative sorites argument as sound.

The sorites paradox may also be understood as consisting in two opposing arguments such as (B) and (C). Taken this way, it is of course a type II paradox. It is still falsidical, however, for the conclusions cannot both be true, and it is also uncontroversial.

One final, rather unusual, example of a sorites argument may drive home the pervasiveness of the paradoxical reasoning, and the extent of the potential damage. A child of one year, let us suppose, does not have the ability to speak English, although five years later, she is a fluent speaker. Suppose we serially order the child's life by seconds. It seems obvious that there is no second of her life such that:

> She does not speak English at t and she does speak English at $t + 1$.

A difference of a second cannot make the difference between having and not having the ability to speak English. But then starting from the fact that she does not speak English at t, and proceeding by increments of one second, it seems we can conclude that she does not have command of the language five years later.[2]

Vulnerability to sorites reasoning is thus widespread, threatening the coherence of most of our everyday beliefs and language. If a minute difference makes no difference to the applicability of the predicate P, then, it appears, it cannot be consistently maintained that P applies to some items in an ordered series and not to others. The sorites paradox is arguably the most disturbing of the paradoxes considered in this book.

In all sorites arguments, the key concept is a vague predicate: tall, red, rich. Thus much of the attempt to come to terms with the paradox has focused on vagueness, and borderline cases are taken to be essential to vagueness. For instance, Shirley may not be clearly tall, but yet not clearly not tall; and there may be no additional information we can acquire that would help us decide how she should be characterized with respect to height. Even knowing her height to the nearest millimetre would not settle the matter.

Similarly, although we see an object in the best perceptual circumstances imaginable, we may be unable to determine whether it is red or orange. These are borderline cases of tallness and redness, respectively.

Although the notion of a borderline case is critical to understanding vagueness, no definition of "borderline case" will be attempted here. For it is not clear that one can give a crisp definition of the concept that does not prejudge the substantive philosophical theories concerning vagueness. The examples of predicates susceptible to sorites reasoning introduced above, however, should suffice to point the reader in the right direction.[3]

The dominant philosophical theory of vagueness is that in borderline cases, the vague predicate cannot be truly or falsely applied. Our ignorance of whether Shirley is tall, for instance, is simply a result of there being nothing to know; there is no fact of the matter in a borderline case. This is the semantic conception of vagueness: vagueness is a function of the way in which words relate to the world. In borderline cases, vague predicates relate to the world in such a way that perfectly meaningful statements are neither true nor false.

An alternative view is given by the epistemic theory. On this account, vague terms only appear to be insensitive; in fact a vague predicate has a precise boundary determining its correct application. The appearance of insensitivity is a consequence of our irremediable ignorance concerning the correct boundaries. Vagueness is rooted in epistemology.

This gives a preview of some of the philosophical terrain to be covered in this chapter. But it is time to face the paradox head on. Strategies for dealing with the sorites paradox, considered as a type I paradox, can be grouped in three categories. It can be argued that:

- the argument is sound and the paradox is veridical; or
- there is a flaw in the reasoning; or
- at least one of the premises is not true.

Each of these alternatives will be considered in turn.

A veridical paradox

The most radical response to the paradox is offered by Peter Unger.[4] On the basis of sorites reasoning, he maintains that the common-sense view of reality is in error: there are no such things as stones, twigs, tables or swizzle sticks. In short, there are no ordinary things.

Unger's argument is simple. To begin, consider the concept of a heap. Clearly, one grain of sand is insufficient for a heap. But if there isn't a heap before us, adding one grain will not create a heap. Hence, no (finite) number of grains is sufficient for a heap. There are thus no heaps, and our concept of a heap is incoherent.

Unger considers this a *direct argument* for his conclusion concerning heaps. But we may also begin by supposing that there are heaps, and that a million grains of sand, properly arranged, constitute a heap. Given that we have a heap of a million grains before us, removing one grain of sand will not leave us with no heap. Nor will removing one more. And so on, until finally we reach the conclusion that one grain of sand will suffice for a heap. This, says Unger, is preposterous; our original supposition that there exists a heap is thus reduced to absurdity. This reasoning is considered an *indirect argument* that there are no heaps, and that the concept of heap is incoherent.

To endorse Unger's direct argument is to treat a negative sorites argument as a sound argument, and a veridical paradox. The negative sorites argument is taken to provide an impeccable demonstration that there are no heaps, no tall people, and so on. The corresponding positive sorites argument, in contrast, is regarded as a *reductio ad absurdum* of the first premise. There is clearly an asymmetry here, a lack of congruity, and it is difficult to see how it can be justified, given the obvious parallels between the two arguments.

But even if we grant that Unger has disposed of heaps, how does he arrive at the more sweeping conclusion that there are no ordinary things? Sorites arguments are typically cast in terms of predicates such as being a heap, or being bald, where the unit of increment (or decrement) is obvious: a grain of sand, a hair. Unger, however, maintains that the argument can be extended to ordinary items such as stones, tables and swizzle sticks. Classic sorites reasoning is adapted to, for instance, a stone, by taking a single

atom as the unit of increment. We begin with just a single atom; clearly, that is not a stone. Now the addition of one atom to something that is not a stone cannot create a stone; one atom cannot make the difference between there being a stone and there not being a stone. Thus, no collection of atoms can constitute a stone; there are no stones.

Unger's ultimate conclusion is stated dramatically: there are no stones, tables or swizzle sticks, there are no ordinary things. But this is to be understood as a striking way of expressing what is essentially a semantic thesis: terms such as "stone" and "table" are *incoherent* and therefore have no extension, cannot apply to anything real. Understood semantically, the thesis might be better supported by maintaining, first, that the negative and positive sorites arguments are valid, have entirely plausible premises and incompatible conclusions, and then inferring from this that the key concept is incoherent. This approach would obviate the need for an asymmetry in the treatment of the two arguments.

Unger attempts to soften the blow of his radical conclusion by pointing out that the common-sense view is not entirely devastated, for nothing has been said to rule out the existence of physical objects of various shapes and sizes. Nonetheless, his position flies in the face of common sense and should be considered only as a last resort, only if there is no way to block the paradoxical argument. It is worth noting here that, aside from the conflict with our everyday beliefs, there are also logical difficulties for the view that vague predicates are incoherent. The result of conjoining two incoherent or inconsistent predicates, it would seem, should itself be an incoherent predicate. If vague predicates are all incoherent, then "is an integer slightly more than 104" is incoherent, as is "is an integer slightly less than 106". Hence, the conjunctive predicate "is an integer slightly more than 104 and slightly less than 106" should also be incoherent. But clearly it has a non-empty extension: 105.[5]

A flaw in the reasoning

Close scrutiny of the reasoning in order to find a fallacy may not seem a promising route to resolution of the sorites paradox. It is true that mathematical induction has been the subject of some controversy, both in the context of sorites reasoning and elsewhere;

some have gone so far as to reject it as completely invalid.[6] But the sorites paradox, we have seen, can equally well be formulated in terms of a sequence of arguments of the form *modus ponens*. Surely denying the validity of *modus ponens* can only be seen as a desperate move.

However, invalidity is not the only possible flaw in reasoning. In the chain of *modus ponens* arguments, one might grant that each sub-argument is valid, the initial premise and the conditional premises true, and yet deny that the reasoning is adequate to establish the truth of the final conclusion. Consider a typical sorites argument:

(1) A person who is worth \$1,000,000 is rich.
(2) If a person who is worth \$1,000,000 is rich, then so is a person worth \$1,000,000 – 1.
∴(3) A person worth \$1,000,000 – 1 is rich.
(4) If person worth \$1,000,000 – 1 is rich, then so is a person worth \$1,000,000 – 2.

⋮

∴(*n*) A person worth \$1 is rich.

The objection here is that, given the length of the argument, the degree of confirmation afforded the final conclusion must be low, and it thus cannot be regarded as having been established. For in the first sub-argument, (3) will receive a degree of confirmation equal to that of the conjunction of (1) and (2). Assuming that neither of the two premises has maximal probability of 1, this means that the probability of (3) is less than that of either (1) or (2). Similarly, it can be shown that the probability of (5) is less than that of either (3) or (4). And so on. The essential idea is that, despite valid reasoning, the degree of confirmation will diminish with the length of the argument, and thus it cannot be assumed that step (*n*) has been established.

We need not get entangled in disputes concerning the model of confirmation here.[7] Even if we are entitled to ascribe to (*n*) only a low degree of confirmation on the basis of the sorites argument, the paradox remains. It is no great consolation to be told that the paradoxical argument does not *prove* the truth of (*n*), for the verdict of

common sense is that (*n*) is *false*. But how can this be? If *modus ponens* is valid, and premises (1), (2), (4), (6), (8), ... are true, then (*n*) must also be true. Granting the validity of *modus ponens*, then, (*n*) can be false only if at least one of the premises cited is not true. But each such premise seems to be a transparent truth. The paradox presents a conflict of reasons, which appeal to confirmation theory does nothing to dispel.

Degrees of truth

At this point, the only option left is to abandon one of the premises of the sorites paradox. Attempts to defuse the paradox by rejecting a premise have focused on the conditional premises, since the first premise of a sorites argument is generally regarded as beyond question.

It has been argued that within the framework of a many-valued logic, one can find independent reason to reject certain of the conditional premises. Vague predicates admit borderline cases, cases in which it is, in some sense, impossible to know whether the attribution of the predicate results in a true or a false statement. The best explanation of this unavoidable ignorance, it is maintained, is that the statement is *neither* true nor false. Of course, to accept this explanation it is necessary to reject a fundamental principle of classical logic:

Bivalence: Every statement is either true or false.

The theories examined in this and the next section, both of which locate the flaw in the sorites argument in a conditional premise, agree in rejecting bivalence, and recognizing more than two truth-values.

The degrees of truth theory, the topic of this section, maintains that statements about borderline cases, although neither true nor false, have *degrees of truth or falsity*. Borderline cases can often be precisely compared: *A* may be clearly taller than *B*, although both are borderline cases of "tall". The statement that *A* is tall thus seems closer to the truth than the statement that *B* is tall, or, put differently, to have more truth in it. From this, it seems natural to infer that "*A* is tall" has a *greater degree of truth* than "*B* is tall".

So we arrive at the notion of degrees of truth. Since the sorites paradox presents us with a continuum of cases, it seems appropriate to have a continuum of truth-values to assign to statements about borderline cases. The convention standardly adopted is that degrees of truth are represented by the real numbers $0 \leq x \leq 1$. The value 1 is assigned to strictly true statements, 0 to strictly false statements. Intermediate truth-values are assigned to attributions of vague predicates in borderline cases according to their approximation to the truth. The initial attribution of a vague predicate in a sorites series will have value 1; but values will gradually decline as one proceeds down the series, ultimately reaching 0.

Although degrees of truth theory departs from classical logic in its rejection of bivalence, the system of logic that emerges from the theory is an attempt to minimize the discrepancy. The logical constants are usually treated as truth-functional: the degree of truth of a compound statement is a function of the degree of truth of its components. Where the degree of truth of P is $[P]$, the favoured proposal is:

$$[\sim P] = 1 - [P]$$

$$[P \,\&\, Q] = \min \{[P], [Q]\}$$

$$[P \vee Q] = \max \{[P], [Q]\}$$

$$[P \supset Q] = 1 \text{ when } [Q] \geq [P]$$
$$= 1 - \{[P] - [Q]\} \text{ when } [P] > [Q]$$

The rationale for this last assignment is that since a conditional is clearly false when its antecedent is true and its consequent false, if the antecedent has a higher degree of truth than the consequent, the conditional cannot be entirely true. Where $[P]$ is greater than $[Q]$, the degree of truth of the conditional decreases as the gap between $[P]$ and $[Q]$ grows, since there is greater decay of truth from antecedent to consequent.

How does the degrees of truth theory dispel the sorites paradox? Consider a typical conditional in a sorites argument where the individuals referred to are borderline cases:

$$\Phi x_n \supset \Phi x_{n+1}$$

If the antecedent is true and the consequent false, the conditional is of course false. But in borderline cases, the components of the conditional have a truth-value less than 1 and greater than 0. Further, Φx_n is closer to the truth, has a higher degree of truth, than Φx_{n+1}. The truth-value assigned to the conditional is thus less than 1; that is, the conditional is not true. Hence the sub-argument in which the conditional is a premise is not sound, and the sorites argument is overturned.

Still, the conditional premises in the chain of arguments are, at the least, very nearly true, while the final conclusion is strictly false. How is this possible? Consider the following argument, which deals only with borderline cases:

$$\begin{array}{ll} & \Phi x_n \\ & \Phi x_n \supset \Phi x_{n+1} \\ \therefore & \Phi x_{n+1} \end{array}$$

Suppose $[\Phi x_n]$ is 0.9, and $[\Phi x_{n+1}]$ is 0.8. Then the value of the first premise is 0.9, as is the value of the second. But the conclusion is just the consequent of the conditional and thus its value is 0.8, which is lower than either of the premises. The conditional of the next sub-argument has an antecedent whose value is 0.8, and a consequent with value of, say, 0.7. So, in the next sub-argument, the conditional premise again has a value of 0.9, but the value of the conclusion is 0.7. In general, the value of the conclusion of each sub-argument will continue to decline as the argument progresses. However, since the antecedent and consequent will presumably continue to differ *by the same amount*, say 0.1, the value of each conditional will be high.

The degrees of truth theory may have the virtue that it can derail the sorites argument, but it is beset with internal difficulties. Of particular concern is the revised system of logic which arises as an offshoot of the theory. Attempting to mirror classical logic, degree theory typically defines a valid argument as one in which the conclusion must have a degree of truth not less than that of the least true premise. The difficulty is that on this account, *modus ponens* is not valid, as the previous paragraph illustrates. Perhaps it would be

more perspicuous to say that the conclusion of a valid argument must have a degree of truth no less than the conjunction of its premises.[8] Such a definition, of course, would yield different results from the first only if coupled with an account of conjunction distinct from the one usually proposed by degree theory. But, in any case, the standard treatment of the logical constants in degree theory has its own counter-intuitive implications.

Classical logic takes $P \vee \sim P$ to be a logical truth, an instance of the law of excluded middle. However, in degree theory, if both disjuncts have a value of 0.5 (both concern borderline cases), $[P \vee \sim P] = 0.5$; the disjunction is not true. Similarly, $[P \,\&\, \sim P] = 0.5$; contradictions need not be false. Nor are such counter-intuitive results limited to cases where there is a logical relation between the components. Suppose x, y and z are balls of different colours and sizes and that:

[x is red] = 1	[x is small] = 0.5
[y is red] = 0.5	[y is small] = 0.5
[z is red] = 0.5	[z is small] = 0

Intuitively, the value of "x is red & x is small" should be higher than "y is red & y is small", but on standard degree theory they are the same. Similarly, "y is red $\vee$ y is small" has the same value as "z is red $\vee$ z is small", although intuitively the former should be higher.[9] Possibly degree theory can develop a more credible account by abandoning the assumption that the logical connectives are truth-functional, that is, that the value of a compound is determined by the value of the components.[10] But clearly there is difficult terrain here for degree theory to navigate.

A final source of concern has to do with higher-order vagueness. Imagine a sorites series for the predicate "tall". At one end, the attribution of tallness to the individual has the value 1 (the individual is tall); at some point, moving along the series, such an attribution has a value of less than 1. Where, one might ask, is the last statement with value 1? If there is such a last statement, then there is a sharp boundary between clear cases of being tall and borderline cases. But this seems contrary to our understanding of vague predicates. Just as there is no precise transition point between being tall and not being tall, so, it seems, there is none between clear cases of being tall and borderline cases. Alternatively, it may be that "has truth-value of 1" is itself vague, and there is thus no last statement with value 1. If so, then the meta-language in which the theory is expressed must itself be vague, and yet must be so constructed that sorites reasoning at the meta-level is not sound.

Truth-value gaps: supervaluationism

Adopting the framework of degrees of truth makes it possible to reject some of the conditional premises of the sorites argument. But an alternative route to the same end is provided by the theory of supervaluationism, which posits only truth-value gaps: statements that are neither true nor false.[11]

Vague predicates may be thought of as having positive extension (things to which the term clearly applies), negative extension (things to which the term clearly does not apply), and penumbral extension (things to which the term neither clearly applies nor clearly fails to apply). For the supervaluationist, statements attributing a vague predicate to a penumbral object are neither true nor false. That is, vague predicates permit truth-value gaps; there simply is no fact of the matter as to whether or not a vague predicate applies to an object in its penumbra.

There is, however, a variety of ways to convert a vague predicate into a precise one, only some of which are acceptable. For instance, it might be useful, for certain purposes, to make the term "morning" more precise. One acceptable way to do this might be to take 6am as the cut-off between night and morning. Another might be to have 5am as the transition point. In any context, however, 11am is unacceptable, for it is ruled out by our understanding of the term "morning". A *sharpening* of a predicate, in order to be *acceptable*, must not be precluded by the meaning of the predicate; objects originally in the positive extension must remain so and those originally in the negative extension must remain there. In effect, any acceptable sharpening must draw a line either through the penumbra or on one extreme edge of it.

The key idea of supervaluationism is that a statement with a vague predicate is true just in case it is true in all sharpenings, and false just in case it is false in all sharpenings. ("Sharpening" now has the sense of "acceptable sharpening".) Otherwise the statement is

neither true nor false. Attributing a vague predicate to a penumbral object will yield a true statement under some sharpenings, and a false statement under others; consequently such statements are neither true nor false.

What does this logical machinery accomplish? First, it successfully blocks sorites reasoning by disqualifying at least one of the conditional premises. Suppose that x and y are immediately adjacent individuals in a sorites series for "tall", that both are borderline cases, and that x is taller than y. Consider the conditional:

x is tall $\supset$ y is tall

There are sharpenings of "tall" in which both antecedent and consequent are true, as well as sharpenings in which both are false. But there is, in particular, a sharpening that draws the line precisely between x and y, in which case the antecedent is true, the consequent false, and the conditional therefore false. So it is not the case that the conditional is true in all sharpenings; therefore it is not (simply) true. Any sorites argument that uses it as a premise is thus derailed. (Note that we cannot say that the conditional is false either; for, assuming it is a material conditional, we can see from the above that there are sharpenings in which it is true.)

A second virtue of supervaluationism is that, although bivalence is rejected, the law of excluded middle still holds, as does the law of non-contradiction. Where x is in the penumbra for "tall", consider:

x is tall $\vee$ $\sim x$ is tall

For any sharpening, exactly one of the two disjuncts will be true, for either x will be in the expanded positive extension (in which case the first disjunct is true), or x will be in the expanded negative extension (in which case the second disjunct is true). So the disjunction is true in all sharpenings, and is thus simply true. What is peculiar is that the disjunction is true without either disjunct being true: "x is tall" is not true in all sharpenings, and nor is "x is not tall".

Similarly, an instance of the law of non-contradiction

$\sim$(x is tall & $\sim x$ is tall)

is true in all sharpenings, and is thus true. Thus far, the principles of classical logic are preserved by supervaluationism.

These consequences of the theory are generally considered to be positive, and weigh in its favour. Other implications of supervaluationism, however, are problematic. Consider a statement that seems to assert a sharp boundary between the tall and the not tall:

(S) ($\exists x$)(A person x millimetres in height is tall and a person $x - 1$ millimetres in height is not tall)

For any sharpening, there is a precise division between the tall and the not tall. That boundary determines the value of x that makes the existential claim true in the sharpening. That is, in every sharpening, there is a value of x that makes (S) true. So (S) is true in every sharpening, and is thus simply true.

This is a highly counter-intuitive result. The theory of supervaluations maintains that there is no number that constitutes the cut-off for "tall". (S) is true, according to the theory, even though there is *no particular number* that satisfies the existential generalization. This means that the behaviour of quantifiers in vague contexts constitutes a radical departure from classical logic. Similarly, we saw earlier that, in the logic of supervaluationism, a disjunction may be true even though neither of its disjunctions is true. Given such disparities, it can hardly be maintained that the theory manages to preserve classical logic in its entirety.

Another area of concern is supervaluationism's account of truth. For many, a necessary condition of any such account being acceptable is that it satisfy Tarski's schema (T), according to which biconditionals of the form

"Snow is white" is true if and only if snow is white

are true. However, (T) ensures that any statement that is not true is false, and thus that there are no truth-value gaps. The reasoning that establishes this connection is:[12]

(1) Suppose "Snow is white" is not true.
(2) "Snow is white" is true if and only if snow is white. (T)

(3) ~Snow is white. from 1 and 2
(4) "~Snow is white" is true if and only if ~snow is white. (T)
(5) "~Snow is white" is true. from 3 and 4
(6) "Snow is white" is false. from 5

The supervaluationist, it is argued, has no reason to reject any step in this reasoning. In order to make room for the truth-value gaps that are essential to his theory, he can only reject (T). For many, this means that the supervaluationist conception of truth is fundamentally misguided.

Finally, supervaluationism, like the degrees of truth theory, faces problems with higher-order vagueness. The concept of an admissible sharpening presupposes that there are precise boundaries between the positive extension, penumbra and negative extension of a vague predicate. But this does not seem to fit with our understanding of vague predicates. A statement such as "Tom is clearly tall" may itself be vague. There is unresolvable uncertainty, it seems, about where the penumbral area begins and where it ends. Some borderline cases are clearly borderline. But it may be uncertain whether an individual falls within the positive extension of "tall", or the penumbra. That person would be a borderline clear case, and a borderline borderline case of tallness. If there is indeed such higher-order vagueness, then the notion of an admissible sharpening will itself be vague. This raises the prospect of sorites reasoning breaking out again at the meta-level.

The epistemic theory

To provide a basis for rejecting one of the premises of the sorites argument, the theories of the previous two sections require complex systems of logic and semantics. In contrast, the epistemic theory is simplicity itself. It licenses the rejection of a premise without any need to tamper with classical logic or semantics. The essence of the theory is that vague terms, contrary to standard conception, do indeed have sharp boundaries; and statements containing vague terms are either true or false. The distinctive feature of a borderline case of a vague predicate, what makes it a borderline case, is that we are *unavoidably ignorant* of the truth-value of the ascription of the predicate to the individual. Vagueness is essentially a matter of ignorance.[13]

If vague terms have precise cut-offs, as the epistemic theory maintains, the sorites argument is easily defused. A typical conditional premise in a sorites series has the form:

$$\Phi x_n \supset \Phi x_{n+1}$$

Given that there is a sharp boundary dividing Φs from non-Φs, it is apparent that exactly one of these conditionals is false, and the argument is therefore unsound. For the same reason, if the argument is cast in the form of a mathematical induction, the inductive step will fail.

The chief virtue of the epistemic theory is conservativeness. The principle of bivalence is preserved, which in turn permits the retention of Tarski's schema for truth. Classical logic and semantics stand without need for revision. This is powerful inducement to regard the theory favourably; yet the standard reaction to the epistemic theory ranges from scepticism to hostility.

What prompts this negative response to the epistemic theory? Timothy Williamson distinguishes four considerations as follows.[14]

Meaning and use

The first objection to the theory contends that words have the meaning we give them; the meaning of a term is determined by how we use it. However, if *b* is a borderline case for a vague term, we typically refrain from either affirming or denying the statement that applies the term to *b*. How can a sharp cut-off possibly be determined by this pattern of use? How can our use determine the location of a boundary when we do not seem to draw a line? To suppose that there is a fact of the matter in borderline cases is to suppose that the meaning of a vague term draws a line where competent speakers of the language do not.

The first response to this objection on behalf of the epistemic theory equates the thesis that meaning is determined by use with the view that meaning *supervenes* on use. The latter is then understood to say that the same use implies the same meaning, that is, that there can be no difference in meaning without a difference in use. But, the defender of the epistemic theory points out, it is entirely consistent with the theory to suppose that there can be no

difference in the meaning of vague terms without a difference in use.

Still, the critic might persist, there is no account offered of *how* the truth-conditions of a vague term may be derived from its use. To this, the counter is that even for precise terms, there is no recipe for calculating the meaning of a term from its use. In particular, neither universal assent nor majority assent guarantees correct application of a term, or truth. The inability to show how the meaning of a vague expression is derived from its use hardly impugns the epistemic theory if there is no known formula, outside the context of vagueness, for extracting meaning from use.

The question remains: does the epistemic theory make it *impossible in principle* to give an account of the connection between meaning and use for vague terms?

Vague properties supervene on precise properties

If two objects are identical with regard to precise properties, then they are identical with regard to vague properties. That is, vague properties supervene on precise properties. Suppose Bernard's height is h, and he is a borderline case of tallness. According to the epistemic theory, either he is tall or he is not tall. This means that being tall (or not tall, as the case may be) is a necessary consequence of having height h. So if I determine Bernard's exact height, I should be able to determine for certain whether or not he is tall. In reality, however, I can know all the precise facts on which the vague fact supervenes without being able to determine the vague fact. I can know Bernard's exact height without knowing whether he is tall. How can this be squared with the epistemic theory?

Williamson's response to this objection is straightforward. Suppose that

(H) Anyone with height h is tall

is a necessary truth given the meaning of "tall". Still, the fact that (H) is necessary does not guarantee that it can be known a priori. Williamson maintains, as he must, that (H) is unknowable by us all; this is what explains how I can know that Bernard's height is h without being able to determine whether or not he is tall.

But this, of course, raises a crucial question: how can the epistemic theory possibly provide a convincing explanation of why simple necessary statements such as (H) should be unknowable by us all?

We cannot know what we mean

The objection here is that since the meaning of "tall" determines a sharp cut-off, and since I cannot locate that cut-off, I use the term without completely understanding it. Worse, no one in the community of competent speakers fully understands the term. What we mean goes beyond what we can know.

Here the answer to the previous objection bears repeating: there is no guarantee that (H) can be known, even if it is a necessary truth. Williamson also points out that to know what a word means is not to know a complete set of necessary truths. Rather, to understand the meaning of a term is to participate in a practice that does determine the meaning, that is, to acquire a set of dispositions that at least roughly matches those of other competent speakers.

Again, the question remains: why it is that (H) is unknowable by us all? This is clearly a critical issue for the epistemic theory.

Why are we ignorant of the cut-off?

What blocks the progress of the sorites argument is the existence of an unknowable but true statement giving the location of a sharp boundary. Where $x_1, \ldots, x_n$ is a sorites series for a predicate F, there is a true but unknowable statement of the form:

$$Fx_i \;\&\; \sim Fx_{i+1}$$

But what exactly is the obstacle to our knowing the relevant conjunctions given that they are true? In general, why it is impossible to have knowledge of borderline cases? Unless we can find an *epistemic account* of why such statements are unknowable, it may be more plausible to suppose simply that they have no truth-value.

Responding to the challenge, Williamson constructs an analysis within the framework of a margin for error. In order for a given true belief to qualify as knowledge, he maintains, it must be

reliable; insofar as a belief's being true is just a matter of luck or accident, it cannot have the status of knowledge. In the particular context of vagueness, reliability requires a *margin for error*. Consider the sorites series $x_1, \ldots, x_n$ for the predicate F. You have a margin for error in believing Fx_i only if all individuals sufficiently close to x_i are also F. For suppose that you truly believe Fx_i, and that Fx_{i+1} is false. Then if things had been *very slightly different*, Fx_i might also have been false, and yet you might still have believed Fx_i. For instance, suppose you truly believe that Sam is not bald, Sam has m hairs on his head, and baldness begins with $m-1$ hairs. Given that you cannot perceptually distinguish between m and $m-1$ hairs, you might have believed that Sam was not bald even if he had one fewer hairs. So if things had been slightly different, your belief would have been false. Your belief is thus not reliable; your getting it right is a matter of luck.[15]

In the context of vagueness, knowledge requires a margin for error; and there is a margin for error in believing Fx_i only if both Fx_{i-1} and Fx_{i+1} are true. So we have:

(M) If it is known that Fx_i, then Fx_{i+1}.

This principle provides the basis for explaining our ignorance of sharp cut-offs. To know the location of a boundary is to know a statement of the form:

$$Fx_i \,\&\, \sim Fx_{i+1}$$

Clearly, knowledge of a conjunction requires knowledge of each conjunct. By (M), knowledge of the first conjunct implies the truth of Fx_{i+1}. So if the first conjunct is known, then it is impossible to know $\sim Fx_{i+1}$, for the simple reason that it is false.[16]

Williamson's derivation of the margin for error principle for vagueness is an adroit and perceptive application of the traditional view that knowledge cannot be a matter of luck, and provides the basis for a credible explanation of our ignorance of sharp boundaries. But the victory is at best partial. For not only is it impossible to have knowledge of the location of a cut-off point, but it is also impossible to have *reasonable belief* on the matter. *Any* belief concerning exactly what height is necessary and sufficient to make a person tall is surely unreasonable. There is more here for the epistemic theory to explain than just lack of knowledge.

The impossibility of reasonable belief concerning the location of a cut-off point appears to present a more intractable problem for the epistemic theory. For one thing, reasonable belief that P, unlike knowledge, has no immediate implications for the truth of P. It should also be noted that justified belief can be right as a matter of luck or accident. Indeed, this is the basis of the Gettier counter-examples to the traditional definition of knowledge. Suppose Sam is justified in believing both:

(S) Smith will be promoted to associate vice-president.
(T) Smith has ten coins in his pocket.

From this Sam infers, and is therefore also justified in believing:

(J) The man who will get the job of associate vice-president has ten coins in his pocket.

But (J) is true, let us suppose, because Brown, whom Sam does not know, will get the job and, as it happens, Brown has ten coins in his pocket. This is just the sort of story that persuades us that knowledge cannot be a matter of luck. The judgement that Sam is warranted in believing (J), on the other hand, is unaffected by our knowledge that it is Brown who will get the promotion. Once it is granted that justified belief may be false, it seems clear that it can also be true as a matter of accident or luck.

Williamson concedes that reasonable belief does not satisfy a margin for error principle. In general, there is a margin for error for a belief P in situation s if P is true in all cases sufficiently similar to s. If reasonable belief that P in situation s required a margin for error, then, since s is sufficiently similar to itself, P would have to be true in s. But it is clear that reasonable belief can be false. How, then, can the epistemic theory account for the impossibility of reasonable belief about the location of boundaries for vague predicates? In tackling this issue, Williamson provides the outline of an intricate analysis that rests on an *externalist* theory of evidence and justification. There are two central theses. First, one's evidence is restricted to what one knows, and thus to true statements.

Secondly, a belief *P* is justified just in case *P* is highly probable relative to the evidence. Suppose now that you are in situation *s*, and you know just those statements that leave a margin for error d. Then, says Williamson, what you know is that your situation is within d of *s*. So the belief *P* will be highly probable conditional on what you know if *P* is true in *most* situations within *d* of *s*.[17] From this, Williamson derives:

(R) A belief is reasonable in a situation *s* just in case it is true in most worlds within *d* of *s*.[18]

This principle forges the critical link between justification and truth, thus providing the key for Williamson's explanation of the impossibility of reasonable belief concerning sharp boundaries. Briefly, the analysis runs as follows. Consider the belief that having height *h* is sufficient for being tall, but a height of $h-1$ is not. Suppose that you are in situation *s*, and that *h* and $h-1$ are indistinguishable to the unaided eye. Since there is only a minute, imperceptible difference between *h* and $h-1$, in some of the worlds closest to *s*, $h-1$ is the minimum height which qualifies a person as tall; in some it is $h+1$; and so on. Williamson goes on to argue that the statement that height *h* makes a person tall, while $h-1$ does not, cannot be true in most worlds sufficiently close to *s*. Given (R), it follows that the belief in a cut-off point cannot be justified, even if true.

The argument is intricate, but at its core is the externalist conception of evidence and justification, which limits evidence to the known, and thus to the true. The implications of this externalism, however, run counter to our strong intuitions in many situations, including the Gettier scenario just outlined. There it seems clear that Sam is justified in believing (J), despite the fact that he infers (J) from (S) and (T), and (S) is false. It should also be noted that the denial that (J) is reasonable conflicts with the deductive closure principle for justified belief. Nevertheless, Williamson must deny that it is reasonable for Sam to believe (J), since (S) cannot count as evidence given that it is false. He has the company of those commentators who have taken this line as a way of evading the Gettier counter-examples, and protecting the traditional definition of knowledge.

Let us try to clarify the issues here by reflecting on what is to be understood by "justified belief". First, it is clearly epistemic

justification that is at issue here. Secondly, justification, as thus understood, is closely linked to rationality. If you believe only what you are justified in believing, you cannot be charged with irrationality *qua* believer; you have done all that can be demanded of you. Put differently, to be justified in your beliefs means that you are epistemically faultless: not biased, not dogmatic or close-minded, not credulous, not gullible, not overly sceptical or incredulous. In short, to be justified in your beliefs is to be immune to criticism on purely intellectual grounds, on grounds of any type of irrationality.

Call this "subjective justification". It is possible that there are other concepts of justification that play a different role. Some (not Williamson) would maintain that there is a concept of justification that suffices to convert true belief to knowledge. For the sake of argument, let us make a concession to Williamson and grant that there is *a* concept of justification that is externalist, and that (J) is accordingly not justified in that sense. Still, it surely must be acknowledged that the belief in (J) has *some positive epistemic status*: surely (J) is subjectively justified. Sam is not irrational in believing (J), he is not at fault *qua* believer. Subjective justification for *P* evidently does not impose truth conditions either on *P*, or on the evidence for *P*.

To return to vagueness and the epistemic theory: There is a type of justification, subjective justification, which (J) has, but which any belief concerning a cut-off point for a vague predicate lacks. Any belief concerning the exact number of grains of sand necessary and sufficient for there to be a heap seems irrational, or subjectively unjustified. What Williamson's theory of vagueness has yet to come to grips with is how the impossibility of subjectively justified beliefs concerning borderline cases may be explained. This may well prove an intractable problem for the epistemic theory.

We have canvassed the most prominent theories of vagueness, none of which seems, at the moment, to provide a wholly satisfactory solution to the sorites paradox. Yet the sorites may well be the most troubling of the paradoxes studied in this work. Our everyday language and concepts are riddled with vagueness; if the argument stands, most of our ordinary concepts and beliefs are incoherent. But the prospect of yielding our most fundamental beliefs to the inexorable steps of the sorites is surely unthinkable.