The Sorites paradox and its significance for the interpretation of semantic theory

Crispin Wright

An unsolved problem in contemporary philosophy of language concerns how best to construe the relation between a formal theory of meaning for a natural language – whether in the truth-theoretic mould which Davidson has advocated or in some other – and the competence of actual speakers of that language. One influential conception of this relation, favoured for instance by Dummett, is that speakers' competence is subserved by their *knowledge*, in some deep implicit sense, of the contents of such a formal theory: they are to be thought of as deploying the information which such a theory states in the ways mirrored by the deductive articulation of the theory, which is why they are able, for instance, to understand novel utterances which they have never heard before.

There is much to say about the notion of implicit knowledge in general, about the propriety of invoking it in this context in particular, and about how else, if not in terms of implicit knowledge, the relation between speakers and an adequate formal theory of meaning for their language should be conceived. In a paper (Wright 1976) published just over fifteen years ago, I argued that the picture of language masters as people who have, at some deep level, internalised a set of semantic and syntactic rules which subserve their competence – just the rules which a satisfactory formal theory of meaning for their language would make explicit – is subverted by the Sorites paradox. But in order to work to a position from which this point can be appreciated, it is necessary to recognise that there are in fact a variety of species of Sorites, formally similar but individuated by reference to the grounds envisaged for the acceptance of their major premisses (usually, universally quantified conditionals). A second principal concern of

this paper will be with the taxonomy of these species – I shall distinguish three – and with the proposal of an appropriately corresponding variety of solutions.

I should remark at the outset that I assume without argument that it cannot be a defensible response to the Sorites to accept it: to grant that the expressions which are prone to it simply are governed by (*de facto*) incoherent rules.

1 THE TACHOMETER PARADOX

Christopher Peacocke (1981) argues that the paradox will survive even if we abandon the conception that understanding is informed by (implicit) propositional knowledge of semantic rules. Peacocke invites consideration of a predicate, C, which is to apply to an object just in case the linguistic community will agree in calling that object 'red'. Suppose that some small difference, d, in the wavelength of light is not visually discriminable by any member of the community. And let light of wavelength k be definitely red. Then, according to Peacocke, we still have this paradox:

> If a reflects light of wavelength k, then $C(a)$.
> If an object differs in the wavelength of light which it reflects by just d from something that is C, it too is C.
> ∴ All visible objects (reflecting pure light) are C.

Hence, Peacocke claims, 'the paradox seems to arise even if we do not suppose that the use of these expressions is governed by rules'.

The proposal of my earlier papers was, in effect, that the plausibility of the major premisses in Sorites paradoxes is owing to this: that such premisses do no more than reflect certain very general features of the rules which govern the use of the affected expressions, provided only that we suppose that there are indeed such rules and that their nature can be disclosed by certain very intuitive reflections. Such intuitive reflections would concern, for instance, the kind of information which could be got across by the sort of training we typically receive in the use of the expressions in question, the constraints which are imposed on the sort of information we can actually handle by our sensory and intellectual limitations, the criteria we should use to determine whether somebody understood the expressions in question, and our overall conception of the purpose and significance of the classifications which the expressions in question effect. If the content of the

relevant semantic rules is constrained by considerations such as these, I argued, then the rules must be such as to sustain the major premisses. Since, as it seemed to me, there would be no intelligibly representing ourselves as following semantic rules if their content were *not* constrained by reflections such as these, my proposal was that we should work towards abandoning that general conception of linguistic competence. In particular, rather than view our competence with colour vocabulary as informed by implicit knowledge of rules associating its proper use with certain objective phenomenological features, for instance, we might do better to work with the sort of model appropriate to practical skills like balancing, swimming and riding bicycles.

The strength of Peacocke's example is that it prescinds from all the complexities we should certainly encounter in trying to make this recommended shift of perspective concrete. His thought, I take it, is this. Think, if you will, of competence with 'red' as a practical skill, uninformed by propositional knowledge. The operation of the skill involves differential sensitivity to varying visual stimuli, but if – as is so – the stimuli permit variation too slight to be detected by our visual apparatus, the skill cannot involve discriminations exercised over differences of that order of magnitude. So if there is communal consensus that a patch of colour is red, and if all that the participants in the consensus are responding to is the visual stimulus, it may seem quite unintelligible how the response can vary if the situation is changed only by altering the stimulus by an amount insufficient to be picked up by our visual systems.

That this way of looking at the matter really does prescind from the propositional-knowledge conception is testified to by the fact that it now seems incidental to the example that it concerns subjects who are capable of grasping contents or following rules at all. Think instead of a digital tachometer on a car. All such a device goes on, all it is designed to respond to, are variations in electronic impulses. There will be limitations to its sensitivity, so sufficiently slight variations in the incoming impulse will not, presumably, provoke any variation in the reading. And now it is neither more nor less plausible than before to conclude that, provided we are careful at each stage not to vary the impulse too greatly, the reading will *never* vary over a series of steps, no matter how many.

It would, of course, be quite unphilosophical to take comfort in the thought that actual tachometers do not behave this way – it is only *because* they do not so behave that we have the appearance of paradox. So Peacocke should be granted that some apparent paradoxes of the Sorites family are not amenable to the response I

suggested. That, of course, completely disposes of the value of that suggestion only if we assume that the entire family must admit of a uniform solution. The first lesson, I suggest, of Peacocke's example is that that is not so.

How should this paradox be responded to? If it would be unphilosophical to take comfort in the fact, it remains that actual functioning tachometers cannot be stabilised through any number of uni-directional but sufficiently marginal changes in the incoming impulse: sooner or later they jump – and a good one will do so often enough, no matter how marginal the changes, to continue to serve as a practically reliable instrument. Therefore the premisses of the tachometer paradox are not true of actual tachometers; and the paradox must be resolved by explaining how exactly that is so. The obvious and presumably correct thought is nothing very exciting: the major premiss, to the effect that if the instrument gives reading R in response to impulse i, it will also give R to any i' differing from i by no more than some specified amount, will turn out to be a misinterpretation of what is entailed by possession of a sensitivity threshold. Quite what the correct account will be will depend on the design of the particular instrument concerned. But one would expect it to be fairly typical that, within the continuous interval $\langle i_1 \ldots i_n \rangle$, there will be a finite subset of points $\{b_1 \ldots b_k\}$ such that, in response to an impulse of any of these values, b_j, or lying in some margin, d, of b_j, the instrument will always respond with the reading f_j; while responses to impulses with other values in the interval $\langle i_1 \ldots i_n \rangle$ will depend additionally upon the prior state of the instrument. One consequence, of course, is that such an instrument will, on different occasions, give different readings in response to the same impulse, depending on how that impulse is 'approached'; but perhaps that is just what such instruments do.

The salient point is that it has to be consistent with the claims that a tachometer has a sensitivity threshold, and that its response is entirely to stimuli of a certain sort, to suppose that the major premiss in the paradox is not everywhere true. Presumably, then, since communal consensus in the use of 'red' does not extend all the way down to orange objects, the major premiss in the C-paradox will be false on similar grounds. And the explanation of its falsity will be consistent with supposing that we all possess visual sensitivity thresholds and respond only to visual stimuli in our use of 'red'.

Peacocke is aware of the possibility of this kind of response to his C-paradox, and is dissatisfied with it. He writes:

> any model for an application of vague observational predicates must provide analogues of three things involved in such application: there must be states which are the analogues of having experiences, there must be something analogous to the . . . non-transitivity of non-discriminable difference, and there must be some analogue of the application of an observational predicate upon a particular occasion.

Peacocke believes that it is impossible to provide for all three conditions in the sort of model illustrated by the tachometer. Suppose we assimilate the state of the instrument, induced by the reception of a particular impulse, to the having of an experience; and the reading which it issues to the application of the predicate. What about the non-transitivity of indiscriminability? In the case illustrated, for instance, we can, by choosing a pair of values which are respectively within and just outside the d-margin of some b_j, elicit – at least sometimes – differential responses from the instrument *no matter how small* the difference between the two chosen values may be. How then can it be claimed that sufficiently small differences are indiscriminable for the instrument, so that – since larger ones are not – the relation 'is not discriminable by the instrument from' behaves non-transitively?

But Peacocke is wrong about this. Or rather, he is right to think it essential that some analogue of non-transitive indiscriminability should be a feature of the model, but wrong to suppose that one cannot be provided. The impression to the contrary requires that the instrument's *sometimes* issuing different responses to a pair of stimuli should be regarded as sufficient for its being able to discriminate between them. And the error in that supposition is easily demonstrated if, reverting to a case in which human beings are the 'instruments', we consider Michael Dummett's (1975) example of the slowly moving pointer. You observe a pointer which is initially at rest but then begins to move, too slowly however for you to see that it is moving. You are asked to give a signal – to raise your right hand, say – as soon as it seems to you to occupy a different position to the initial position; and – let us add to Dummett's example – to raise both hands as soon as it seems to you to have moved into another position again. Suppose that you raise one hand after four seconds, and both after eight. Are we to conclude that the position which the pointer seemed to you to occupy after the third second looked different from the position which it seemed to you to occupy after the fourth? The answer, of course, is that you cannot detect any difference between those positions – what you meant to indicate when you first raised one hand was only that the position now seemed to you to be different from the starting-point.

As Dummett brings out, there are ways of describing the phenomenology of the situation which are threatened with incoherence – and perhaps none which are not. But it is legitimate to suppose, for the purposes of the analogy with the tachometer, that your permissible 'signals' are restricted to the raising and lowering of hands. Suppose, then, that the actual positions of the pointer after each second are determined by some appropriately finely calibrated instrument. The situation will be that you sometimes will and sometimes will not respond to the sorts of change involved from one second to the next by varying your signal; but that, *asked* about any single such change, you would report that the positions involved seem to you to be the same. Moreover, if we imagine that, once you have raised both hands, the movement of the pointer is reversed, we would expect to find that, when it regained the position originally occupied after three seconds, you would still have one hand raised. If we – as is permissible – suppose further that your signals for the positions occupied initially, after four seconds and after eight seconds are by and large respectively uniform, then your performance is now in all relevant respects assimilable to that of the hypothetical tachometer. And, crucially, the fact that you always (more or less) have a hand raised when the four-second position is presented, but only sometimes have a hand raised when the three-second position is presented is *no indication that those positions seem different to you*. On the contrary, if they did seem different, you would *always* give a different signal. So far, then, from its being a sufficient reason for regarding the tachometer as able to discriminate between a pair of impulses that it sometimes issues different readings in response to them, the fact that it does not *always* do so should be regarded as decisive for regarding the relation in which they stand to it as an analogue of *indiscriminability*. And the modelling of non-transitivity is then provided by the reflection that, among three distinct impulses, the tachometer may *only sometimes* respond differentially with respect to the first and second, and *only sometimes* respond differentially with respect to the second and third, but will *always* respond differentially with respect to the first and third.

I conclude that this class of Sorites paradox need not detain us. Their major premisses are false, and false in a way which classical two-valued semantics is quite adequate to describe. The details of why they are false will, of course, depend upon the character of the particular 'instruments' concerned; and I am not seriously suggesting that the tachometer provides a satisfactory model of all aspects of human competence with colour vocabulary. My point is only that

the case which Peacocke makes for supposing C to be Sorites-susceptible is perfectly matched by the Tachometer paradox; and the model solution sketched for the latter depends on no feature of the tachometer without an analogue in the 'instrument' constituted by a human being responding to colour. Accordingly, while different things might need to be said about why precisely the major premiss in the C-paradox is false, we can at least understand why we shall not be committed to its truth simply by supposing that our descriptions of colour are nothing but responses to visual stimuli among which our discriminations are not everywhere transitive.

2 HIGHER-ORDER VAGUENESS AND THE NO-SHARP-BOUNDARIES PARADOX

It is widely assumed in the literature that Sorites-susceptibility is deeply connected with vagueness. Since almost all the expressions of typical natural languages are vague, the belief that it is vagueness which generates the paradox brings one uncomfortably close to the thought, advocated by such philosophers as Peter Unger (1980), that natural languages, and the conceptual systems which they reflect, are typically incoherent. The principal point of this section is that, with a qualification on which I shall dwell, the thought that vagueness generates, *per se*, Sorites-susceptibility is a muddled thought.

When spelled out it goes, presumably, something like this. If ϕ is vague, its very vagueness must entail that in a series of appropriately gradually changing objects, ϕ at one end but not at the other, there will be no nth element which is ϕ while the $n + 1$st is not; for if there were, the cut-off between ϕ and not-ϕ would be sharp, contrary to hypothesis. Accordingly, the vagueness of ϕ over such a series must always be reflected in a truth of the form:

$$\sim(\exists x)\,(\phi x \;\&\; \sim\phi x') \qquad (1)$$

where x', of course, is the immediate successor of x. That, of course, is a classical equivalent of the universally quantified conditional which is the major premiss in standard formulations of the Sorites paradox – a thought which has prompted Hilary Putnam to suggest that a shift to intuitionistic logic might be of value in the treatment of the paradox. So indeed it might. But it will not be enough, for intuitionistic logic will yield a paradox direct from the negative existential premiss (as will any logic with the standard $\exists$- and

&-Introduction rules + *reductio ad absurdum*). This form of the paradox – the *No Sharp Boundaries paradox* – thus appears to constitute a proof that (one kind of) vagueness is *eo ipso* a form of semantic incoherence.

Only 'appears' though. What the No Sharp Boundaries paradox brings out is that, when dealing with vague expressions, it is essential to have the expressive resources afforded by an operator expressing *definiteness* or *determinacy*. I have heard it argued that the introduction of such an operator can serve no point since there is no apparent way whereby a statement could be true without being definitely so. That is undeniable, but it is only to say that – in terms of Dummett's distinction – the *content senses* of '*P*' and 'Definitely *P*' coincide; whereas the important thing, for our purposes, is that their *ingredient senses* differ, the vital difference concerning the behaviour of the two statement-forms when embedded in negation. Equipped with an appropriate such operator, we can see that a proper expression of the vagueness of ϕ with respect to the relevant sort of series of objects is not provided by the above negative existential, but rather requires a statement to the effect that no definitely ϕ element is immediately succeeded by one which is definitely not ϕ; formally,

$$\sim(\exists x)\,[Def(\phi x)\ \&\ Def(\sim\phi x')] \tag{2}$$

This principle generates no paradox. The worst we can get from it, with or without classical logic, is the means for proving, for successive x', that

$$Def(\sim\phi x') \rightarrow \sim Def(\phi x) \tag{3}$$

from which nothing untoward follows.

A believer in *higher-order vagueness* may want to reply that this merely postpones the difficulty. If, for example, the distinction between things which are ϕ and borderline cases of ϕ is itself vague, then assent to

$$\sim(\exists x)\,[Def(\phi x)\ \&\ \sim Def(\phi x')] \tag{4}$$

would seem to be compelled, even if assent to (1) is not. So once again the materials for paradox seem to be at hand, each ingredient move taking the form, for instance, of a transition from $\sim Def(\phi k')$ to $\sim Def(\phi k)$.

I believe that this thought, that higher-order vagueness would be *per se* a source of paradox, may quite possibly be correct. But some complication is needed, for the following is the immediate reply. Of

any pair of concepts, ϕ and ψ, which share a blurred boundary, we shall want to affirm

$$\sim(\exists x)\,[Def(\phi x)\ \&\ Def(\psi x')] \tag{5}$$

when x ranges over the elements of an appropriate series in which the blurred boundary between ϕ and ψ is crossed. The original problem occurred when, with $\sim\phi$ in place of ψ, we overlooked the necessary role of the definiteness operator. And now we are guilty of the same oversight again; it is merely that this time ψ has been replaced by $\sim Def(\phi)$. As soon as the inclusion of the definiteness operator is insisted on, all that emerges is

$$\sim(\exists x)[Def(Def(\phi x))\ \&\ Def(\sim Def(\phi x'))] \tag{6}$$

which yields nothing more than the harmless

$$Def(\sim Def(\phi k') \rightarrow \sim Def(Def(\phi k)) \tag{7}$$

Evidently the strategy will generalise; we need never, it seems, be at a loss for a way of formulating ϕ's possession of vagueness, of whatever order, in a way that avoids paradox.

But this is too quick. We are able to be confident that the sort of formulation illustrated by (7) avoids paradox only because we have so far no semantics for the definiteness operator, and are treating it as logically inert. Without considering what form a semantics for it might take, the crucial question is whether it would be correct to require validation for this principle:

$$\text{(DEF)}\quad \frac{\Gamma \vdash P}{\Gamma \vdash Def(P)}\ ;\ \text{provided } \Gamma \text{ consists of propositions, all of which are prefixed by '}Def\text{'}$$

For, in the presence of DEF, and assuming that the corrected formulation (6) above, of what it is for the borderline between ϕ and its first-order borderline cases to be itself blurred, is itself *definitely* correct, the harmless (7) gives way to

$$Def(\sim Def(\phi k')) \rightarrow Def(\sim Def(\phi k)) \tag{8}$$

a generalisation of which will enable us to prove that ϕ has no definite instances if it has definite borderline cases of the first order. (Proof in Appendix 1.)

DEF says, in effect, that the truth of each of a set of fully definitised propositions ensures that every consequence of that set is likewise definitely true. This gets some spurious plausibility from

conflation with the distinct and indisputable principle that whatever is a consequence of a set of propositions, each of which is definitely true, is itself definitely true. But DEF is plausible in any case, and disclosure of any error in it is, as I have tried to illustrate in Appendix 1, an awkward project to say the least. If DEF is valid, then higher-order vagueness – always a difficult and vertiginous-seeming idea – would indeed be a paradox-generating phenomenon; *ergo*, presumably, a delusory one. The task would then be to explain the delusion. It remains that the idea which has smitten writers such as Unger, that lack of sharp boundaries is *per se* paradoxical, is merely retribution for working with too crude a formulation of what lack of sharp boundaries is.

3 SORITES OF THE THIRD KIND AND THE IMPLICIT-KNOWLEDGE CONCEPTION OF LANGUAGE MASTERY

The third class of Sorites paradoxes consists of those for which support for their major premisses seems to flow just from our accepting that correct use of the expressions in question is determined by rules which competent speakers know, and that the character of these rules can be illuminated by the kind of intuitive reflections which were illustrated earlier: reflection on the nature of the training which (putatively) bestows grasp of them, reflection on the criteria which we should apply for determining whether someone had grasped them, reflection on the manner and circumstances in which competence with the expressions in question is typically exercised, reflection on the constraints imposed on the content of such rules by the intellectual and sensory limitations of those who play by them, and reflection on our understanding of the nature and purposes of the distinctions which the expressions in question enable us to draw.

My earlier paper (1976), and Dummett's (1975), treated of a putative category of *observational* expressions, characterised as those whose application to an object would be determined purely by its appearance. Such an expression would have to apply to both, if to either, of any pair of objects whose appearances are identical. We both assumed that colour predicates, like 'red', would come into this category. If that assumption was correct, we should immediately have the materials to establish a Sorites paradox for 'red' in a series of objects adjacent elements of which were indistinguishable in point of colour, although the end points were

respectively red and orange. The fact is, though, that the sense in which 'red' is observational needs a much more subtle account than that, which here I do not propose to supply. Instead, let us move directly to the more plausibly observational 'looks red' (more plausibly observational, that is, in the above sense of the term). I note in passing that, arguably, the semantic complexity of 'looks red' is only apparent, since an object's looking red is not the same thing as its looking *as if it is* red, which is undoubtedly a semantic complex. (In order to look as if it is red, it may in the circumstances have to look, for example, brown.) Notice, too, that any Sorites paradox for 'looks red' is going to generate one for 'red' as well if, as is plausible, it is necessary and sufficient for its being red that an object look red under certain privileged conditions. Likewise, an object's looking red suffices for its being warrantedly believable that it is red, provided there is no reason to doubt that the conditions are of the privileged sort; so there will be a paradox for 'is warrantedly believable to be red' also.

Can 'looks red' fail to apply to all, if to any, of a range of indistinguishable items? When the sorts of informal consideration canvassed above are allowed, the case for a negative answer almost makes itself. Surely the distinction between what definitely looks red and what does not is one which it must be possible to make salient by ostensive means – which it would not be if it applied selectively among indistinguishable items. Even if we can make trouble for this thought, it is certain that competence with 'looks red' is practised by subjects who have only quite ordinary powers of observation and rely on nothing but casual exercise of those powers, eschewing in particular any use of the kind of external aids – colour charts or whatever – which could compensate for limitations of memory. Any escape route *has* to involve finding a way of avoiding drawing paradoxical conclusions from this undeniable fact. But how – if we say that competence consists in knowing certain rules and their limits? For it then seems inescapable that both every distinction prescribed by the rules, and the distinction between cases where the rules have something to say and cases where they do not, can be based only on contrasts which may be detected by casual exercise of ordinary powers of observation, without reliance on external aids; and, hence, that no such distinction can be exemplified by items which conjointly exhibit no such contrast. Recommended conclusion: competence with expressions of the class which 'looks red' typifies *cannot* consist in knowledge of the requirements of certain rules and of their limits.

An even better argument of this sort for the Sorites-susceptibility of 'looks red' is even simpler. If items are visually indiscriminable,

they look the same. So visual indiscriminability is necessarily a congruence relation for any predicate whose rules of application take account only of how things look. Our conception of the role and purpose of 'looks red', 'looks orange', etc. is, unquestionably, that we use them *purely* to record public appearances. So if we may legitimately base upon this conception conclusions which concern the character of the rules governing such predicates' application, those rules must have the feature, it seems, of relating only to appearance, of prescribing application of such predicates only and purely on the basis of appearance. But then they must prescribe application of such a predicate to both members, if to either, of any pair of items whose appearances are the same.

Here there is terribly little room for manoeuvre. I see no alternative but to drop the idea that the harmless truism that we use predicates like 'looks red' to record how things appear to us has any bearing on the character of the putative semantic rules which govern their use. Since – so it seems to me – the truism could hardly fail to have the very direct bearing illustrated so long as there *were* governing semantic rules for such expressions at all, the recommendation is that we drop that assumption, and adopt, as the matter was expressed in Wright (1976), a 'more purely behaviouristic' conception of what competence with such expressions involves.

But what does that mean? A way to provide at least the beginnings of a proper account of the contrast is to enquire what connection there is between the obtaining of an instance of the type of state of affairs which confers truth on a token of the sentence '*x* looks red', and our willingness to assent to that token. It should not be controversial, I think, that each is necessary and sufficient for the other so long as certain provisos are met which are distinctive of this class of expression: that is, for no other class of expressions do exactly these provisos subserve such a biconditional dependence. The provisos are that the judging subjects understand 'looks red', that their perceptual faculties are functioning normally, that they are otherwise in good cognitive order, that *x* is presented to them in clear view, and that they are attentive to *x*. Subject to these provisos, assent and truth necessarily coincide. We have, that is,

> (∀S) If *S* understands 'looks red', and enjoys normally functioning perceptual faculties, and is otherwise in good cognitive order, and has *x* presented to him in clear view and is attentive to *x*, then [*S* will judge of *x* that 'looks red' is true of it if and only if 'looks red' is true of it]

Now, two quite different broad perspectives on this principle – the

provisional biconditional – are possible. One, required by the implicit knowledge conception, will hold that the circumstance that 'looks red' applies to *x* is settled, independently of any subject's response to *x*, by the semantic rules which govern the use of 'looks red' in English and by how, objectively, *x* appears. The role of the provisos, on this view, is to ensure the subjects' ability to 'track', or detect, this independent state of affairs: thus, their understanding of 'looks red' and their being in 'good cognitive order' ensures their sensitivity to the requirements of the relevant semantic rules; and their normal perceptual function and attentiveness to *x*, and the clear presentation of *x* to them, ensure their sensitivity to *x*'s objective appearance. On this view, then, appeal to the idea of how a subject will or would respond to the utterance when the provisos are satisfied should play no essential part in an account of its truth-conditions. Such an account need consider nothing but the semantics of the utterance and the characteristics of *x*. What the provisos do is to foreclose on every possible explanation of fracture between the fact of the matter and a subject's response.

As remarked, the implicit-knowledge conception imposes this first perspective. The reason is that, if it is to be *explanatory* of our linguistic competence to suppose that we have knowledge, albeit implicit knowledge, of a framework of rules, then it has to be possible to think of what is involved in following those rules as something which we (implicitly) *recognise* and which is therefore constituted independently, in whatever sense of 'independently' is necessary to give sense to the idea that substantial feats of cognition are involved. The perspective is, moreover, prerequisite for the sort of Sorites-generating thinking which I have been sketching. In order for there to be such a Sorites paradox for 'looks red', our actual classificatory responses have to be out of accord with the requirements of the relevant semantic rules. So to think you have such a paradox, you have to be working with an epistemology of semantic rules which allows firm conclusions to be drawn about their character independently – or anyway, in a manner sufficiently disrespectful – of the shape which those responses are actually disposed to assume. That is: conclusions are, apparently, *not* to be drawn by reference to the character of our response to a predication of 'looks red' when, *by ordinary criteria*, all the provisos are fulfilled. In order for this exclusion to be legitimate, the dictates of the semantic rules for 'looks red' have to be thought of as constituted somehow independently of such responses.

The first perspective is thus both enjoined by the implicit-knowledge conception and essential to the thinking that makes the

'looks red' paradox seemingly so powerful. And to recognise that the perspective is wrong is both to solve the paradox and to learn that the implicit-knowledge conception is in error. It is wrong, if the second, alternative perspective is correct. This perspective turns everything around. Now the provisional biconditional is seen, instead, as itself supplying the canonical form of a statement of the truth-conditions of '*x* looks red'. The provisos are no longer seen as serving to describe the conditions under which a subject succeeds in tracking an independent fact; rather, for *x* to look red just *is* for subjects to be willing to assent to that judgement when, by ordinary criteria, the provisos are met. Whatever else we want to say about the content of expressions like 'looks red', or about the epistemology of their content, it is all answerable to this point. Hence there is no possibility of such an epistemology teaching us that the provisos are in fact not satisfied in circumstances where, by normal criteria, we should have been satisfied that they are. It is the other way about: if we are tempted to opinions about the meanings of such expressions which force us to draw such a conclusion, it is those opinions which are at fault. There is no ulterior fact which meeting the provisos ensures that we can detect, so no possibility – by reference to independent criteria for the existence or non-existence of such a fact – of surprising conclusions about when the provisos really are not met, notwithstanding the satisfaction of ordinary criteria for saying that they are.

To spell out what this distinction does for us, suppose a group of subjects are agreed that 'looks red' applies to the first colour patch in a Sorites-series in circumstances when, by ordinary criteria, the provisos are met. As the series is run from apparently red to apparently orange patches, a point will be reached where, despite its indiscriminability from the immediately preceding patch, a consensus in subjects' responses breaks down for the first time. That is undeniable. It is also undeniable that, when that happens, no single subject in the group need, by ordinary criteria, have fallen out of accord with any of the provisos. (Borderline cases are exactly cases about which competent subjects are allowed to differ.) The second perspective on the provisional biconditional gives us the right to treat such a circumstance as raising a doubt about the predicability of 'looks red'. The first perspective, by contrast, leads to the cancellation of that right.

Someone who succumbs to the arresting, apparently simple thought that 'looks red' must be applicable to both, if to either, of any pair of indiscriminable items is likely to be taking it in one of two ways. Taken in one way, it is the central move in the Tachometer paradox: how can a signalling device, even a properly

functioning signalling device, discriminate among stimuli whose difference is smaller than its sensitivity threshold? We now know, I hope, the outline of a satisfactory response to that question. But the second and, I think, more natural way of taking the thought conceals, in effect, a presupposition of the first perspective on the provisional biconditional. You have to forget that we do not, or would not so apply the predicate in every case and fall in, instead, with the idea that something can be discerned about its *proper* conditions of application just by intuitive reflection upon the kind of content which it overtly seems to have. If there were facts about the proper application of such predicates which were constituted independently of our best responses – the responses we would have when, by ordinary criteria, all the provisos were met – what else could they be but the offspring of rules which correlated their proper use with *appearances*? And how, when following such rules, could it ever be justifiable, in consequence, to assign to identical appearances distinct responses? The reply should be that competence with such predicates is nothing to do with the capacity to fit one's usage to the dictates of rules of that kind. Indeed, it is not a matter of compliance with rules at all, if that is taken to imply the propriety of a 'tracking' or detective direction of interpretation of the provisional biconditional. What it is correct to say using such a predicate is a function only of what we are actually inclined to say when there is no available reason to doubt that the provisos are met.

The distinction between the two perspectives needs further work, of course, preparatory for a demonstration, if one can be given, that the second, non-detective reading is indeed the right one. My suggestion is that the detective reading, which, on pain of incoherence – or so I have been arguing – we must jettison for predicates like 'looks red' is incorrect everywhere: that this kind of Sorites paradox forces us to recognise local examples of what is in fact a global error. But we should not stampede to the conclusion that philosophy of language can make no legitimate play with the notions of semantic rule and implicit knowledge. What we have to jettison is the myth of linguistic competence as constituted by the ongoing exercise of capacities of semantic *detection*. Whatever notions of semantic rule and of implicit knowledge can continue to serve us have to be consonant with this. In opposition, I would like to put something like the picture of content as actively and ceaselessly determined by our best responses, the judgements we make about content when, by ordinary criteria, appropriate provisos are fulfilled. But that is to begin to paddle in a familiar, yet very badly charted, Wittgensteinian sea.

Appendix 1
DEF and the No Sharp Boundaries paradox for higher-order vagueness

DEF transforms the harmless 7 (p. 143) into the noxious 8 because it sanctions the inference from $Def(\phi k)$ to $Def(Def(\phi k))$, and hence the biconditional theorem:

$$(\text{IT}) \vdash Def(A) \leftrightarrow Def(Def(A))$$

(which is uncontroversial right to left). Now if, as was supposed, A and $Def(A)$ may have differing ingredient senses, this may seem obviously unacceptable. But matters are not so straightforward. The question is whether the difference between them is one to which the context '*Def*( )'is, like 'Not( )', itself sensitive. Suppose, for instance, we had a semantics which accounted $Def(A)$ *false* when A was anything other than definitely true, and Not-A as borderline when A was borderline. Not-$Def(A)$ would then, presumably, be true when A was borderline, diverging, as it intuitively should, from Not-A; but $Def(A)$ and $Def(Def(A))$ would both be false. An approach having these features would not, I suppose, be a non-starter – the idea has some appeal that, if A is on any sort of borderline, the claim that it is anything else is false. But unless we found more to say, there would be no evident objection to DEF.

There *is* more to say, of course. To take higher-order vagueness seriously is just to allow that cases may arise where it is indeterminate whether a statement is true or borderline. To say that its definitisation was *false* in such a case would be, in effect, to rule that the original was borderline – to ignore its leanings, as it were, towards truth. So the sort of semantics just prefigured, which promises to validate DEF, is anyway guilty of failing to take higher-order vagueness seriously: to repeat, taking high-order

vagueness seriously involves allowing that $Def(A)$ may itself, on occasion, be borderline.

So to allow, however, will make no difference as far as IT is concerned unless, when $Def(A)$ is borderline, $Def(Def(A))$ may either be false *or*, at any rate, of an inferior degree of truth to that of $Def(A)$. But the idea that $Def(Def(A))$ may be false when $Def(A)$ is borderline can hardly be separated from the corresponding claim about $Def(A)$ and A respectively – the claim just accused of failing to take higher-order vagueness seriously. So a defender of higher-order vagueness should prefer the second type of proposal: when a statement is borderline, so should its definitisation be, but not in such a way as to sustain IT, left to right.

How might this proposal be elaborated? Let us pretend for a moment that we really do understand the idea of an indefinite hierarchy of orders of vagueness, along the lines:

$$F \ldots -2, -1, 0, 1, 2, \ldots T$$

where level 0 comprises statements which are neither definitely true nor definitely false, level 1 comprises statements on the borderline between those at level 0 and those which are definitely true, level 2 comprises those on the borderline between those at level 1 and those which are definitely true, and so on; likewise, level -1 comprises those on the borderline between those at level 0 and those which are definitely false, level -2 comprises those on the borderline between those which occupy level -1 and those which are definitely false, and so on. (Exercise: the justification for the suspicion that understanding of this idea is fictional will rapidly emerge if you try to characterise as it were primitively – using just the notions of truth, falsity, negation and the definiteness operator – what it is for a statement to occupy some specific level in the hierarchy other than 0, F and T.) Still, if we can draw these distinctions, then a relatively straightforward proposal would be that, for any statement of level $k \neq T$ or F, its definitisation is of level $k-1$, but otherwise is also of level k. So if A is (definitely) true, or false, so is $Def(A)$; but if A is on some sort of borderline, $Def(A)$ lies on the immediately inferior borderline. And $A \leftrightarrow B$ will hold only if A and B are of the same level.

But the problem with this is evident enough. Any statement, A, of level $k > 0$, $< T$, ought to be partially characterisable as one which is not definitely true. So the statement of which that part-characterisation is the negation ought to be *false* – and that statement ought to be $Def(A)$. It is accordingly impossible to marry

the sense given to '*Def*' by the proposal with our intuitive understanding of 'definitely'.

The point is a general one. We cannot intelligibly characterise higher orders of vagueness in terms which invoke statements' failure to be definitely true, yet simultaneously require definitisation to generate falsity only when applied to false statements. We could, of course, drop the latter requirement without reverting to the original idea that $Def(A)$ *always* polarises to truth or falsity. A (somewhat messy) compromise would be that $Def(A)$ continues to drop a level on A just in case A occupies some negative level $>F$, but polarises to falsity if A occupies 0 or any positive level $<T$. But, as will be apparent, this compromise does nothing to obstruct the relevant version of the No Sharp Boundaries paradox.

1 (1) $Def{\sim}(\exists x)\,[Def(Def(\phi x))\ \&\ Def({\sim}Def(\phi x'))]$; Ass.
2 (2) $Def({\sim}Def(\phi x'))$; Ass.
3 (3) $Def(\phi x)$; Ass.
3 (4) $Def(Def(\phi x))$; IT.
2,3 (5) $(\exists x)\,[Def(Def(\phi x))\ \&\ Def({\sim}Def(\phi x'))]$: 2,4, $\exists$-intro.
1 (6) ${\sim}(\exists x)[Def(Def(\phi x))\ \&\ Def({\sim}Def(\phi x'))]$; 1, *Def*-elim.
1,2 (7) ${\sim}Def(\phi x)$; 3,5,6, RAA.
1,2 (8) $Def({\sim}Def(\phi x))$; 7, DEF.
1 (9) $Def({\sim}Def(\phi x')) \rightarrow Def({\sim}Def(\phi x))$; 2,8 CP.

Clearly IT and DEF do not survive the mooted compromise without restrictions: $Def(A)$ will fail of equivalence to $Def(Def(A))$ for any A of negative level $>F$; and $Def(A)$ will be of level lower than A for A of any level, negative or positive, $\neq T$ or F. But remember that each ϕx or $\phi x'$ with which we are concerned in this version of the paradox is of level 0 or greater – we start with an x' which is a definite borderline case of ϕ and 'work left', so to speak. So the compromise poses no obstacle to the application of IT at line 4: $Def(\phi x)$ and $Def(Def(\phi x))$ will both be *false for every germane* ϕx. Similarly, the compromise offers no objection to the application of DEF at line 8: if ϕx is of level 0 or greater level $<T$, ${\sim}Def(\phi x)$ and $Def\,{\sim}Def(\phi x))$ will both be true, and if ϕx is of level T, ${\sim}Def(\phi x)$ and $Def\,{\sim}(Def(\phi x))$ will both be false. And, to stress: every ϕx at which we arrive by 'working left' is in one of those two cases.

I am under no illusions that these sketchy remarks constitute a treatment of the topic, of course. But they do begin to suggest something of the nature of the obstacles confronting one who would wish to hold that higher-order vagueness is a genuine but unparadoxical semantic characteristic.

Appendix 2
On some criticisms of Sainsbury's

Just before this volume was due to go to press, I had the opportunity to see Mark Sainsbury's interesting 'Is there higher-order vagueness?'[1] Professor Sainsbury's paper (1991) raises a number of fundamental issues which I can make no attempt to comment on in detail here.[2] One of the most radical is his contention that it is a mistake to seek to capture vagueness, of whatever 'order', in the form of a characteristic sentence: 'a sentence schema, containing a schematic predicate position, such that the sentence resulting by replacing the schematic element by a predicate is true iff that substitute is a vague predicate' (section 3). The impression to the contrary derives, in Sainsbury's view, from the lingering influence of what he styles the 'classical conception' of vagueness: the idea that what defines a vague predicate is that it effects a tripartite division – into positive, negative and borderline cases, respectively – where a precise predicate determines a merely bipartite one. The difficulty is then to say something coherent about how the tripartite division can itself be blurred at the edges – so that a merely tripartite distinction is seemingly not enough. Obviously my

$$ {\sim}(\exists x)\,[Def(Def(\phi x))\ \&\ Def({\sim}Def(\phi x'))] \qquad (6) $$

was exactly an attempt to produce the execrated sort of characteristic sentence for *second-order* vagueness: it tries to say what it is for the distinction between the definite ϕs and the definite borderline cases to be itself vague – for the transition between the two kinds of case not to occur at an abrupt threshold. According to Sainsbury, this attempt is misguided in principle. Rather we should recognise that:

> The right way to characterise the vagueness of a predicate is by the fact

> that it classifies without drawing boundaries: it is *boundaryless*. A boundaryless predicate allows for borderline cases, but this is not its defining feature. A boundaryless predicate draws no boundary between its positive and negative cases, between its positive cases and its borderline cases, between its positive cases and those which are borderline cases of borderline cases. The phenomena which, from a classical viewpoint, lead to notions of 'higher order vagueness' are accounted for by boundarylessness . . . To convince you that boundaryless classification is possible, I would ask you to think of the colour spectrum. It contains bands but no boundaries. The different colours stand out clearly, as distinct and exclusive, yet close inspection shows that there is no boundary between them. The spectrum provides a *paradigm* of classification, yet it is boundaryless . . . We must shift away from the classical perspective. We are carried away by images which make us find boundarylessness problematic. We think of a system of classification as like a grid, a system of pigeon-holes, a way of drawing a line, dividing a field. In this way of thinking, Frege's idea that a boundaryless concept is no concept at all seems irresistible. But we should shift images. Classification is better likened to providing magnetic poles around which some objects cluster more or less closely and from which others are more or less repelled; some fall between a number of poles, drawn by more than one but especially close to none.

I think there is something importantly correct in the adjustment to much contemporary thought about vagueness which, in the discussion around these remarks, Sainsbury is trying to teach. But it is another question whether the adjustment will bestow an understanding of vagueness which is both intuitively satisfying and paradox-free. Even if the 'classical conception' mislocates what *defines* vagueness, it is quite another matter whether it altogether misdescribes it. Sainsbury himself acknowledges that 'a boundaryless predicate allows for borderline cases' – so vagueness *is* associated with a tripartite division, or taxonomy of relevant cases, even if this is 'not its defining feature'. But now, why should it matter whether a feature is a defining feature or not, provided it *is* a feature? How could the classical conception have been led to avoidable paradox by correct characterisation of *non-defining* features, even if it mistakenly took them to be defining ones? Above all, how is the conception of vagueness as boundarylessness fundamentally at odds with the characteristic-sentence approach – why should there not be 'a sentence schema, containing a schematic predicate position, such that the sentence resulting by replacing the schematic element by a predicate is true iff that substitute is a *boundaryless* predicate'? And why in particular is the approach illustrated by (6) not suitable to generate such a sentence-schema? Perhaps these questions somehow miss Sainsbury's point. But I cannot attempt to take matters further here.

I shall briefly comment on two other aspects of Sainsbury's discussion. Given that he wishes to turn thought about vagueness away from the 'characteristic-sentence' approach, the strongest possible support for his view – provided it can indeed avoid sustaining characteristic sentences of its own – would be if vagueness is unavoidably represented as paradoxical by what *would* be a characteristic sentence, *if* there were any such thing. He thus has a dialectical interest in sustaining the suspicions about higher-order vagueness and the No Sharp Boundaries paradox which I have been attempting to develop. So it is surprising to find him placing almost as much weight on two reservations he has about the detail of that development as on what he perceives as the general misdirectedness of the characteristic-sentence approach.

His first reservation concerns the particular form of characteristic sentence I have been working with, (6) above, which he transposes into its classical equivalent

$$Def(Def(\phi x)) \rightarrow \sim Def(\sim Def(\phi x'))^3 \qquad (6)^c$$

The motivation for this, recall, was that if x' is a borderline case of ϕ, it will at least be true that it is not definitely ϕ; and that if it is a *definite* borderline case, then the same will be definitely true. Thus (6) or, if you will, $(6)^c$ says that no (definitely) definite ϕ thing is succeeded by a definite borderline case – that the distinction between the ϕs and the definite borderline cases is not one with an abrupt threshold: not a sharp one. Is that not just what second-order vagueness ought to be?[4]

Sainsbury's reservation is that there are other candidates for the characteristic sentence, from which one cannot 'by a proof of the same general structure as Wright's, derive anything paradoxical . . . So their entitlement to represent vagueness needs to be undermined before any conclusion antithetical to vagueness can be drawn from Wright's proof' (section 4). What candidates? Presumably, when vagueness is at issue, we shall be working with some notion, however intuitive, of truth-value gaps, or of truth-values other than truth and falsity. Either will set up the possibility of a fracture within the notion of negation. One notion, the *strong* negation of A, may be defined as true just in case A is false. The *weak* negation of A, by contrast, will be true just in case A is other than true. In these terms, it is natural to characterise a borderline case of ϕ as something such that neither the claim that it is ϕ nor the strong negation of that claim is true. Writing 'Neg A' for the strong

negation of A and 'Not A' for the weak, x is thus a borderline case of ϕ if

$$\text{Not } \phi x \ \& \ \text{Not Neg } \phi x$$

This gives us something else to play the role of '*Def*($\phi x'$)' in $(6)^c$. If we also reflect that no intuition is offended by replacing the occurrence of '*Def*(*Def*(ϕx))' by one of '*Def*(ϕx)', we arrive at Sainsbury's first alternative:

$$Def(\phi x) \rightarrow \text{Not } Def(\text{Not } \phi x' \ \& \ \text{Not Neg } \phi x') \qquad (6)^s$$

Alternatively, reflect that, plausibly, '*Def*($\phi x'$)' is *false* when x is borderline; so

$$\text{Neg } Def(\phi x')$$

is true. The thought that no x', next to a definite ϕ, is a borderline case of ϕ is thus equally plausibly expressible by

$$Def(\phi x) \rightarrow \text{Not Neg } Def(\phi x')^5 \qquad (6)^{ss}$$

Sainsbury's claim, then, is that neither $(6)^s$ nor $(6)^{ss}$ generates a paradox in the fashion of (6) and $(6)^c$; but that their credentials as characteristic-sentences for higher-order vagueness are no less plausible.

Sainsbury himself criticises both $(6)^s$ and $(6)^{ss}$. About the latter, he remarks that if '*Def A*' always polarises to true or false, whatever the truth-status of A ('and who at this stage is to say it does not?'), then 'Neg *Def*($\phi x'$)' will collapse into 'Not *Def*($\phi x'$)', whence – since 'Not' sustains double negation elimination[6] – $(6)^{ss}$ will collapse into the familiar villain

$$Def(\phi x) \rightarrow Def(\phi x')$$

About $(6)^s$ he notes that, since 'Not' generates a sentence with a polar truth-value from any sentence, instances of

$$\text{Not } \phi x' \ \& \ \text{Not Neg } \phi x'$$

will always be polar in truth-value, so that it is unclear what role '*Def*' is playing in the consequent of $(6)^s$.

These observations are correct; but they seem to me to be, respectively, the wrong one to make and not to go far enough. The right observation to make about $(6)^{ss}$ is that, when we put it under pressure, we either recover the means to generate a Sorites paradox, or wind up committed to the non-existence of the higher-order vagueness of which it was meant to serve as a coherent characterisation. When '*Def*(ϕx)' is true, it tells us, it is always untrue that '*Def*($\phi x'$)' is false. So how do things stand with

'*Def*($\phi x'$)'? We must not allow that it is always true or we get a Sorites. But if double negation elimination holds for 'Not', then so will the corresponding version of the law of excluded middle:

$$A \text{ or Not } A^7$$

We are consequently powerless to avoid a dilemma: for each '*Def*($\phi x'$)', either it or its weak negation will hold. The cost of blocking the Sorites which will result if we accept each '*Def*($\phi x'$)' is thus that we must at some point weakly deny one – *tertium non datur*. And as soon as we do, we have established a sharp boundary to definite ϕ-ness, when the whole point was to describe what is involved in there being none.

Things are no better with $(6)^s$. Sainsbury's thought was that, if not in general, then at least when A is polar in truth-value, '*Def A*' will necessarily have the same truth-value as A – the definitisation of a true or false sentence must likewise be, respectively, true or false. And in that case, granted that

$$\text{Not } \phi x' \ \& \ \text{Not Neg } \phi x'$$

will always be polar, the truth-value of $(6)^s$ must coincide with that of

$$Def(\phi x) \rightarrow \text{Not}(\text{Not } \phi x' \ \& \ \text{Not Neg } \phi x')$$

but that is equivalent to

$$Def(\phi x) \rightarrow \text{Not Not } \phi x' \vee \text{Not Not Neg } \phi x'$$

which, in the presence of double negation elimination, yields the catastrophic

$$Def(\phi x) \rightarrow \phi x' \vee \text{Neg } \phi x'$$

It is true that there is no directly generating a Sorites from this. But it is obviously completely hopeless as what it was meant to be – a characteristic sentence for higher-order vagueness – and it *will* generate a paradox if we append the thought, apparently accepted by Sainsbury, that the definitisations of true sentences are true. For then, starting with a definitely ϕ case, we can proceed once again by dilemma, forcing acceptance of each successive x' either that it is ϕ, so definitely ϕ, or – contrary to the hypothesis that ϕ is even *first-order* vague! – that its immediate successor is a negative instance of ϕ.

In the main part of the paper I remarked that, when dealing with vague expressions, it is essential to have the expressive resources afforded by *Def*. I suspect that one moral of the foregoing is that it is

essential to *lack* the expressive resources of the sort of weak negation operator introduced by Sainsbury.[8] However that may, and prescinding from the reasons for dissatisfaction with $(6)^s$ and $(6)^{ss}$ which Sainsbury himself offers, I cannot see that they contribute to any case at all that the connection between higher-order vagueness and paradox which I suggested was merely an artefact of reliance on (6) or $(6)^c$ as characteristic schemata. On the contrary: the suggestion emerges, if anything, then the stronger, since – on natural assumptions, made by Sainsbury himself – both $(6)^s$ and $(6)^{ss}$ are, as we have just seen, paradoxical in their own right.

There is, however, a more basic point concerning Sainsbury's strategy. Part of his purpose in introducing $(6)^s$ and $(6)^{ss}$ was to provide at least *prima facie* coherent characteristic schemata for vague predicates whose 'entitlement to represent vagueness needs to be undermined before any conclusion antithetical to vagueness can be drawn from Wright's proof'. This thought seems to me to have matters precisely backwards. Rather it is the entitlement of (6) or $(6)^c$ to represent vagueness which needs to be undermined before any *comforting* conclusion about the status of vagueness could be drawn from the availability – had we but seen it – of paradox-free characterisations. If a notion has an intuitively acceptable characterisation which generates paradox, it is no progress towards a resolution of matters merely to devise other seemingly acceptable characterisations which, so far as one can see, avoid paradox. So long as nothing is done to disarm the intuitive credentials of the villain, they – the credentials – merely transfer into grounds for thinking that the apparently innocent characterisations either fail to do justice to the intended notion or are not really innocent. There is really no substitute in this context, for one who believes a paradox-free characteristic sentence should be formulable and has no other objection to my 'proof', to disclosing how (6) or $(6)^c$ involves determinate misrepresentation of the idea they are supposed to characterise, and to disarming or otherwise speaking to the motivation for them. That is what Sainsbury, in that part of his paper where, without his 'no characteristic-sentence' hat on, he tried to make use of $(6)^s$ and $(6)^{ss}$ against my 'proof', should have been trying to do instead.

Sainsbury's other principal reservation concerns the role of the inference rule, DEF. He writes:

> one cannot feel happy with the introduction of the undefined *'Def'* followed immediately by an assumption about its logic which leads to paradox. It would seem a clear possibility that there should be a conception of *'Def'* upon which it demands progressively higher standards. Such a conception would fail to validate IT
>
> – the iterativity principle for *Def*: see Appendix 1 –
>
> or the definitisation rule (DEF), and would need to be argued against if higher order vagueness is to be shown paradoxical by the argument.

I largely agree with these reactions, which betray some misunderstanding of my original intent. One thing I regard as definite progress, in an area where it is exceedingly hard to make any, is the modest insight that the No Sharp Boundaries paradox may be defused by appropriate use of an operator of definiteness. What I sought to show was that this point, which ought to extend to a coherent characterisation of vagueness of higher order, if that notion is coherent at all, will not so extend *unless* DEF fails in some relevant way. The alternatives are thus, prescinding from any residual doubt about (6), to disclose a relevant failing or face the consequence that higher-order vagueness is *per se* paradoxical. Sainsbury reacts as if I had claimed to establish the latter disjunct, when my aim was the disjunction.

Nevertheless, I believe he underestimates the problem of disclosing a relevant failing in DEF. It is perfectly true that if, as Sainsbury puts it, *'Def'* demands progressively higher standards – if, in other words, in order for *'Def A'* to reach a certain level of acceptability, *A* must in general surpass it – then IT may fail for a range of cases in which the truth-value of *A* is not polar. But there are difficulties involved in the attempt to put this idea to work against the paradox, some of which were reviewed in Appendix 1. In particular, if *'Def A'* is not always to be polar – in which case, of course, IT will hold – there is the problem of how to state the intuitive point that all borderline cases of *A* should be cases where that statement is *not definitely true*. Since the operator of negation which figures in that thought is wanted to generate truths when applied to non-polar statements, it would seem that it has to be the sort of weak negation which we have already had cause to consider. But that notion sustained the law of excluded middle, and will thus impose the sort of paradox by dilemma reviewed in discussion of $(6)^s$ and $(6)^{ss}$ above.

In any case, although DEF entails IT, so that shortcomings in the latter must rebound to the discredit of the former when it is taken in full generality, such shortcomings are not really to the point. For we might be satisfied that DEF is not generally valid, yet still have no fault to find with the kind of application of it essentially involved in the reasoning to the paradox. Reflect that IT was utilised only to

collapse the iterated occurrences of '*Def*' in (6) and (6)^c. But, as remarked above, there is really no good intuitive cause to prefer (6)^c as formulated to

$$Def(\phi x) \rightarrow \sim Def(\sim Def(\phi x'))$$

as a characterisation of higher-order vagueness. So we can dispense with IT and derive the paradox using just the definitisation of the above and DEF. The crucial question is therefore whether fault can be found with the specific type of application of DEF which is needed.

Consider how it goes. We proceed as follows:

(1)	(1) $Def[Def(\phi x) \rightarrow \sim Def(\sim Def(\phi x'))]$	Assumption
(1)	(2) $Def(\phi x) \rightarrow \sim Def(\sim Def(\phi x'))$	From (1)
(3)	(3) $Def(\sim Def(\phi x'))$	Assumption
(1)(3)	(4) $\sim Def(\phi x)$	From (2) and (3)
(1)(3)	(5) $Def(\sim Def(\phi x))$	From (4) by DEF
(1)	(6) $Def(\sim Def(\phi x')) \rightarrow Def(\sim Def(\phi x))$	From (5) by conditional proof

Sainsbury's thought was in effect that DEF might fail for cases in which the conclusion of its premiss-sequent was not polar, when the definitisation of that conclusion might drop crucially further in truth-value, as it were. Obviously, such a case can arise only if (some of) the assumptions of the premiss-sequent are already themselves non-polar – otherwise the premiss-sequent would not be a valid entailment in the first place. And in that case the thought is beside the point. For (1), we are taking it, is *true*; and we presumably so select our starting-point for the Sorites that (3) is also. More simply, if (1) and (3) are true claims, then so is (4). So if the transition to (5) is to fail, definitisation must be capable of producing a drop in truth-value even when applied to true premisses. Not only does Sainsbury produce no reason for thinking that it can, but also, as noted above, the belief that it cannot was active in the reservation about (6)^s which he himself canvassed.

I conclude that, as far as Sainsbury's two specific reservations – about DEF and (6)/(6)^c – are concerned, the case for thinking that higher-order vagueness may be intrinsically Sorites-generating remains essentially intact. But that does not speak to his more general proposal about boundarylessness which, notwithstanding the doubts I expressed earlier, remains for further development and discussion.

NOTES

1. Sainsbury's paper is in substantial measure a reaction to parts of my 'Further reflections on the Sorites paradox' in *Philosophical Topics*, vol. 15, spring 1987, pp. 227–90, from which my contribution to the present volume is largely derived. My thanks to the Editor of *Philosophical Topics*, Christopher Hill, for permission to reprint material from 'Further reflections'.
2. In particular, I shall not attempt to respond to his criticisms of the argument, in Section 4 of 'Further reflections', pp. 244–5, that 'red' is not higher-order vague.
3. Like Sainsbury, I now omit the initial quantifier.
4. There is a slight infelicity here, in so far as '$\sim Def(\phi x)$' is not actually definitive of x's being a borderline case of ϕ, but will also be true if x is a negative case. But nothing important hangs on this. The thought in the text is restored – if good at all – by restricting the range of x to positive and borderline cases of ϕ. Alternatively, the reader may prefer to treat '$\sim Def(\phi x)$' as characterising the agglomerate of borderline and negative cases together: (6) and (6)^c then plausibly capture what it is for the distinction between ϕ's and this agglomerate to be vague – which is just what it is for ϕ to be second-order vague.
5. Sainsbury does not quite motivate (6)^{ss} this way. I have simplified his discussion slightly, which had a generalised version of (6)^s characterising vagueness of order $n+1$ as

   $$Def^n(\phi x) \rightarrow \text{Not } Def^n(\text{Not } \phi x' \ \& \ \text{Not Neg } \phi x')$$

   where 'Def^n' expressed n iterations of '*Def*'. As he remarks, this works out only for $n > 0$ – hence his shift to (6)^{ss}. Note that the latter shares the harmless infelicity of (6) and (6)^c: see note 4 above.
6. The thought is presumably something like this. Suppose P is something other than true. Then 'Not P' is true; so the claim that it is something other than true is *false* – so of truth-value equal or inferior to that of P. Hence

   $$\text{Not Not } P \rightarrow P$$

   will go through on the standard sort of many-valued semantics for '$\rightarrow$'. (Double negation *introduction*, by contrast, will fail for this type of weak negation, as the reader will swiftly verify.)
7. The standard natural-deductive proof will involve, besides double negation elimination, only the ordinary rules of *reductio ad absurdum* – valid, presumably, for weak negation – and vel-introduction. I know of no reason to regard either as suspect in the presence of vagueness.
8. See the ante-penultimate paragraph of this appendix.

REFERENCES

Dummett, M. (1975) 'Wang's paradox', *Synthèse*, vol. 30, pp. 301–24; reprinted in *Truth and Other Enigmas*, London: Duckworth, 1978.
Peacocke, C. (1981) 'Are vague predicates incoherent?', *Synthèse*, vol. 46, pp. 121–41.
Sainsbury, M. (1991) 'Is there higher-order vagueness?', presented to the Anglo-French conference on 'Language, concepts and communication', University of Sussex, 6–8 April 1990; *The Philosophical Quarterly*, vol. 41, pp. 167–8.
Unger, P. (1980) 'There are no ordinary things', *Synthèse*, vol. 41, pp. 117–54.
Wright, C. (1976) 'Language mastery and the Sorites paradox', in Evans, G. and McDowell, J. (eds), *Truth and Meaning*, Oxford: Clarendon Press.

Knowledge representation versus meaning representation*

Daniel Kayser

1 INTRODUCTION

I am convinced that the problems which are discussed in this paper are relevant to philosophy. However, as my background in philosophy is virtually zero, I expect to have difficulties in making myself understood, as I often have difficulties in understanding the point of several philosophical papers. So in order to make my position clear, I need to make a few somewhat oversimplified preliminary remarks, hoping that, by so doing, I shall minimise the risk of my using words which already have a specialised philosophical employment.

I consider my field, artificial intelligence (henceforth AI), as a science, not as a technique; our problem is to get a better understanding of what intelligence really is, not to build machines giving the impression of reacting intelligently, no matter how.[1] An important subfield of AI is the design of devices accepting natural-language texts, and able to give responses to a wide range of questions concerning these texts. The problem is only superficially different if, instead of texts, the device is given speech, visual scenes or whatever.

For the task considered – namely, having machines which are to give intelligent responses – it is obvious that a large quantity of

* I am grateful to the referees; their critical comments helped me greatly in rephrasing, and sometimes in rethinking, some obscure points of the first version. Of course, I am alone responsible for the remaining weaknesses of the paper.