Mind & Language, ISSN: 0268-1064
Vol. 13. No. 4 December 1998, pp 457–476.

Conscious Thinking: Language or Elimination?

PETER CARRUTHERS

Abstract: Do we conduct our conscious propositional thinking in natural language? Or is such language only peripherally related to human conscious thought-processes? In this paper I shall present a partial defence of the former view, by arguing that the only real alternative is *eliminativism* about conscious propositional thinking. Following some introductory remarks, I shall state the argument for this conclusion, and show how that conclusion can be true. Thereafter I shall defend each of the three main premises in turn.

1. Introduction

How are language and thought related to one another? While almost everyone allows that language-use, in any full-blooded sense, requires thought, there is considerable dispute about whether thought, in turn, requires or involves natural language. On the one hand are those who espouse what can be called *the cognitive conception* of language, who maintain that language is crucially implicated in thought—as well as being used for purposes of communication, of course (Wittgenstein, 1921, 1953; Vygotsky, 1934; Whorf, 1956; Davidson, 1975, 1982; Dummett, 1981, 1991; Dennett, 1991; as can be seen, they are a varied bunch). And on the other hand are defenders of what can be called *the communicative conception* of language, who maintain that language is *not* essentially implicated in thinking, but rather serves only to facilitate the communication of thought (Russell, 1921; Grice, 1957, 1969; Lewis, 1969; Fodor, 1978; Searle, 1983; Pinker, 1994; again a varied list, though perhaps not quite *so* varied).

While this paper is concerned to defend a form of cognitive conception of language, it is important to see that the version in question is a relatively weak one, in at least two respects. First, the thesis that language is constitutively involved in human thought is here put forward as holding with, at

Versions of this paper were given to Philosophy Department seminars in Sheffield and Coleraine; and to the Cognitive Science Centre in Hamburg. I am grateful to all those who participated in the discussions. I am also grateful to Keith Frankish and Tim Williamson for valuable comments on an earlier draft.
Address for correspondence: Department of Philosophy, University of Sheffield, Sheffield, S10 2TN, UK.
Email: p.carruthers@sheffield.ac.uk.

most, *natural* necessity. It is certainly not claimed to be conceptually necessary. So there is a contrast in this respect with some of the main defenders of the cognitive conception (specifically, Davidson and Dummett), who have claimed that their thesis is an a priori conceptual one. For reasons which I do not intend to go into now, this strikes me as highly implausible (see Carruthers 1996, ch. 1, for some brief discussion). On the contrary, the case for the *conceptual* independence of thought from language is, I believe, a very powerful one. But that leaves open the possibility that there may still be some sort of naturally necessary involvement. Second, the thesis that language is involved in human thought is not here maintained universally, but is restricted to specific kinds of thought, particularly to *conscious propositional* thoughts. I shall make no attempt to show that *all* human thought constitutively involves natural language.

While the version of cognitive conception to be defended in this paper is relatively weak, the question is still of considerable interest and importance. For what is at issue is the overall place of natural language in human cognition. According to the communicative conception, the sole function and purpose of natural language is to facilitate communication (either with other people, in the case of written or spoken language, or with oneself, by means of 'inner speech'—see below). Spoken language thus serves only as the medium, or conduit, through which thoughts may be transmitted from mind to mind, rather than being involved in the process of thought itself. Something like this is now the standard model employed by most of those working in cognitive science, who view language as an isolable, and largely isolated, *module* of the mind, which is both innately structured and specialized for the interpretation and construction of natural language sentences (Fodor, 1978, 1983, 1987; Chomsky, 1988; Levelt, 1989; Pinker, 1994). According to the (weak form of) cognitive conception of language, in contrast, natural language is constitutively involved in some of our conscious thought-processes, at least. So language is not (or *not just*—see Carruthers, 1998) an isolated module of the mind, but is directly implicated in central cognitive processes of believing, desiring, and reasoning.

Now, many of us are inclined to report, on introspective grounds, that at least some of our conscious propositional thinking is conducted in imaged natural language sentences—either spoken (in motor imagery) or heard (in the form of auditory imagery). And certainly, the systematic introspection-sampling studies conducted by Russ Hurlburt (1990, 1993) found that, while proportions varied, *all* subjects reported at least some instances of 'inner speech'. (In these studies subjects wore a modified paging-device, which issued a series of beeps through an earphone at irregular intervals during the course of the day. Subjects were instructed to 'freeze' their introspective consciousness at the precise moment of the beep, and immediately to jot down notes about those contents, to be elaborated in later interviews with the experimenters. Subjects reported finding, in varying proportions, visual imagery, inner speech, emotive feelings, and purely propositional—nonverbal—thoughts.) So the existence of inner speech, itself, is not—or should not

be—in doubt. The question concerns its status, and its explanation. According to the weak form of cognitive conception to be defended here, inner speech is partly constitutive of thinking. According to the communicative conception, in contrast, inner speech is merely *expressive* of thought, perhaps being the medium through which we gain access to our thoughts.

One final clarification (and qualification) is in order, before I present the argument for my conclusion. This is that I am perfectly happy to allow that *some* conscious thinking does *not* involve language, and nothing that I say here should be taken to deny this. Specifically, it seems plain that conscious manipulation of visual or other images can constitute a kind of thinking (recall what might run through your mind as you try to pack a number of awkwardly shaped boxes into the trunk of your car, for example), and a kind of thinking which seems clearly independent of language. However, since it also seems likely that such imagistic thoughts are not fully propositional, having contents which can only awkwardly and inaccurately be reported in the form of a that-clause, I can restrict my claim to conscious *propositional* thought. And the standard arguments which refute imagistic theories of meaning, or imagistic theories of thought, can be used to show that there is a space here to be occupied, since imagistic thinking cannot be extended to colonize the whole domain of conscious thought (unless, that is, the images in question are images of *natural language sentences*—see below).

2. *The Argument*

The argument for the claim that conscious propositional thinking is conducted by means of natural language sentences is as follows.

(1) Conscious thinking requires immediate, non-inferential, non-interpretative, access to our own occurrent thoughts, and that access is distinctively different from our access to the thoughts of other people.
(2) Occurrent propositional thoughts either receive articulation in inner speech, or they do not; and if they do, then inner speech is either constitutive of the thought-tokens in question (the cognitive conception), or not (the communicative conception).
(3) If the manipulation of natural language sentences in imagination (in 'inner speech') is *not* constitutive of propositional thinking, then our access to those of our thoughts which receive expression in inner speech is interpretative, and similar to the sort of access which we have to the thoughts of other people, when they speak; and hence such thoughts of ours do not count as conscious (by 1).
(4) The form of access which we have to those of our occurrent propositional thoughts which do *not* receive expression in inner speech also involves self-interpretation, and hence such thoughts, too, fail to count as conscious (by 1).
(5) So if we engage in conscious propositional thinking at all, then natural

language sentences must be constitutively involved in such thinking (from 1, 2, 3, and 4).

(6) But we do sometimes engage in conscious propositional thinking.

(7) So natural language is constitutively involved in conscious thought (from 5 and 6).

The argument itself is plainly valid. Premise 2 presents us with a three-branched disjunction. Two of these disjuncts are then closed off by Premises 3 and 4 respectively (relying upon Premise 1), leaving the third disjunct to be conditionalised in Premise 5. Now, Premise 2 is just a truism. So if Premises 1, 3, and 4 are held firm (supposing that they can be adequately supported by further argument, as I shall try to show respectively in sections 4, 5, and 6 below), then our choices are: to give up Premise 6, hence becoming eliminativists about conscious propositional thought; or to accept 7, thus endorsing a version of the cognitive conception of language.

For many people, I suspect, the correct response to such a dilemma will seem straightforward—if there is a conflict between the communicative conception of language, on the one hand, and the belief that we sometimes entertain conscious propositional thoughts, on the other, then, plainly, it is the former which should be given up. But this would be much too swift. For many of the arguments against eliminativism about the mental in general (Horgan and Woodward, 1985; Fodor, 1987) do not extend to eliminativism about conscious propositional thought in particular. Specifically, the arguments from the indispensability and success of common-sense psychology can cut little ice in this more restricted domain, since it is not at all clear whether, or why, we really *need* the assumption of *conscious* thought for purposes of intentional explanation and prediction. And nor can this sort of restricted eliminativism be immediately refuted by any kind of appeal to introspective phenomenology, since it is, in fact, quite easy to explain how we might come to be under the illusion that we engage in genuinely conscious thinking, as we shall see in sections 5 and 6 below.

(This is not to deny, of course, that the belief that we sometimes entertain conscious propositional thoughts might be very deeply embedded in our common-sense conception of ourselves. Thus Tyler Burge (1996) has recently argued, for example, that the belief that we have non-inferential, non-interpretative, access to our own occurrent thoughts is crucial to our status as *critical reasoners*—and ultimately, perhaps, to our status as persons. But even if this is correct, it does not really provide us with any independent argument against eliminativism, unless we have some further ground for believing that we actually *are* critical reasoners, in Burge's restrictive sense, as opposed to being under the illusion that we are such reasoners.)

Despite what I have just said, it does seem to me somewhat more reasonable to accept Premise 6, than it is to reject the cognitive conception of language (represented by the conclusion in 7). But this involves examining, and disarming, the various arguments which have been put up against the cognitive conception of language, and in favour of the communicative conception.

I shall not attempt to do this here (see Carruthers, 1996, much of which is taken up with consideration and rebuttal of such arguments). For present purposes I shall have to be content to leave it to my readers to decide which way they want to jump. (But note that even a disjunctive conclusion has considerable bite, since there are many people who think that they can combine a belief in Premise 6 with the denial of 7.)

It *is* important to see, however, how the conclusion in 7 can possibly be true, consistent with the other premises; for it is natural to wonder how natural language *can* be constitutive of conscious thinking, if the latter requires non-inferential, non-interpretative, access, as claimed in Premise 1. It is also important to recognize the strength of the empirical commitments which we would take on by accepting 7. So in the next section I shall try to show how the manipulation of natural language sentences in imagination can be partly constitutive of propositional thinking; and if it is, how we would have the right sort of non-inferential access to some of our occurrent thoughts for them to count as conscious ones.

3. The Possible Truth of 7: On Thinking Linguistically

Suppose that imaged sentences have the causal roles distinctive of occurrent thoughts. That is, suppose that it is *because* I entertain, in judgemental mode, the sentence 'The world is getting warmer, so I must use less fuel' that I thereafter form an intention to walk rather than to drive to work. If the imaged natural language sentence is a crucial part of the process which causes the formation of an intention, and is thus, ultimately, part of what causes my later action, then this might seem sufficient for it to be constitutive of the occurrent thought. This would, however, be too swift. For a defender of the communicative conception can allow that there are some chains of reasoning which cannot occur in the absence of an imaged natural language sentence (Clark, 1998). If it is, for example, by virtue of our thoughts causing the production of imaged natural language sentences that we gain access to their contents and occurrences, then any chain of reasoning which requires us to have such access will constitutively involve an imaged sentence. But, by hypothesis, the imaged sentence is not *itself* the thought, but is merely what gives us *access to* the thought. So, rather more needs to be done to get at the intended idea behind (this version of) the cognitive conception of language.

The obvious thing to say, in fact, is that an imaged sentence will occupy the causal role of a thought if it has the distinctive causes and effects of that thought, but *without* these being mediated by events which themselves carry the same (or a sufficiently similar) content. So the sentence 'The world is getting warmer' will count as constitutive of my conscious thought if it (together with my other beliefs and desires) causes my intention to walk to work, but *not* by virtue of first being translated into a non-imagistic event which carries the content, [that the world is getting warmer]. But is it even

possible for an imaged sentence to occupy the causal role of an occurrent thought? The answer, surely, is 'Yes'—in at least three different ways.

First, it may well be that our *non*-conscious thinking does not involve natural language sentences, but rather consists in manipulations of sentences of Mentalese (or, alternatively, of activations in a connectionist network, as it might be). These non-conscious thoughts may also be used to generate imaged natural language sentences, which are then processed in some sort of meta-representational executive system, in such a way that we can say, not merely that the imaged sentence gives us *access* to the underlying thought, but that it *constitutes* a distinct (conscious) token of the *same* thought. (This is the weaker of the two hypotheses I explore in Carruthers, 1996, ch. 8, where natural language sentences are constitutive, not of propositional thought types, but of the conscious *tokens* of those types.) Such a description will be appropriate provided that the effects of the imaged sentence-token within the executive are not mediated by the equivalent sentence of Mentalese. This will be the case if, for example (and as Dennett, 1991, has argued), there are certain kinds of inference which one can learn to make amongst thoughts, which one can make only when those thoughts are tokened in the form of a natural language sentence. (For evidence that there are tasks which can be solved only with concurrent verbalisation, see Berry and Broadbent, 1984, 1987.)

Second, it may be that all propositional thoughts involve natural language representations of one sort or another (or, at least, that some significant subset of propositional thought-types do). Conscious thoughts might, as above, be constituted by imaged natural language sentences, which interact with one another in the manner distinctive of occurrent thoughts. But non-conscious tokens of (some of) those same thought-types, too, might be constituted by some non-phonological natural language representation, say a sentence of Chomsky's 'Logical Form' (LF), as it might be, in which sentences are regimented in such a way as to resolve scope-ambiguities and the like (May, 1985; Chomsky, 1995). On this picture, then, human cognition will involve computations on two sorts of natural language representation—computations on phonological entities, in consciousness, and non-conscious computations on LF representations with the same contents. (This is the stronger of the two hypotheses explored in Carruthers, 1996, ch. 8, according to which some propositional thoughts, as *types*, constitutively require natural language representations.)

Third, there is the possibility currently being developed by Keith Frankish, which builds on some early work of Dan Dennett's on the difference between belief and opinion (Dennett, 1978; Frankish, 1998). The idea, here, is not (as above) that imaged sentences are manipulated by the sub-personal mechanisms operative in some central executive system, but rather that they are objects of personal (rationally motivated) *decision*. On this model, when the sentence 'The world is getting warmer, so I must use less fuel' figures in consciousness, I may *decide* to accept it, thereby deciding to adopt a policy of using that sentence thereafter as a premise in my theoretical and practical

reasoning. Since such decisions may then have many of the same effects as would the corresponding belief, they may be thought of as constituting a kind of *virtual belief*. And here, as before, the sentence-token in question is partly constitutive of the opinion thereby adopted. (Note that this possibility will then admit of both weaker and stronger variants corresponding to those canvassed above, depending upon whether ground-level, non-virtual, beliefs are carried by sentences of Mentalese—or perhaps patterns of activation in a connectionist network—on the one hand, or rather by natural language sentences—i.e. non-imagistic, LF representations—on the other.)

The only question remaining, then, in order for us to demonstrate the possible truth of 7, is whether our access to our own mental images has the kind of non-inferential character necessary for those images to count as conscious. And the answer, surely, is that it has. Our access to our own visual and auditory images, for example, seems to be part of the very paradigm of immediate, non-inferential, non-self-interpretative awareness. And similarly, then, with inner speech: we have immediate access to a particular phonological representation, together with its interpretation. The latter point is worth stressing: when I form an image of a natural language sentence in a language which I understand, just as when I hear someone else speak in that language, what figures in consciousness is *not* just a phonological object standing in need of interpretation. Rather, what figures there is *already* interpreted—I *hear meaning* in the words, just as I hear the speaker's (or my own imagined) tone of voice. (Note, however, that this claim about the phenomenology of inner speech is not sufficient, by itself, to establish the truth of 7. If the phenomenally-immediate content of a tokened sentence is to count as a conscious *thought*, then it must itself occupy the causal role distinctive of such a thought. So we would also need to endorse one or other of the three possibilities sketched in the paragraphs above.)

It might be questioned how the *content* of an imaged sentence *can* be an object of immediate awareness. For suppose that the contents of my sentences are determined, at least in part, by my inferential dispositions—perhaps by my dispositions to find certain inferences to and from the sentences in question *primitively compelling* (Peacocke, 1992). Then how could these dispositions be objects of immediate, non-inferential, awareness? There are really two issues here, however: first, how could I know *that* the sentence has a content for me? and second, how could I know *what* content the sentence has for me? And thus divided, the problem is easily conquered. I can know *that* the sentence is contentful by a kind of feeling of familiarity (or, more plausibly perhaps, by the absence of any feeling of *un*familiarity)—by means of a well-grounded confidence that I should know how to go on from it, for example. And I can know *what* content the sentence has, simply by *embedding* it in a content-report. Given that I have just entertained the sentence 'The world is getting warmer' and that it is contentful, I can then immediately and non-inferentially report that I have just thought *that the world is getting warmer*. The content of the initial sentence is automatically

made available within the content of the embedding sentence which reports on that content.

It is worth noting that the claimed immediacy of our access to the forms and contents of our own mental images is (despite initial appearances to the contrary) fully consistent with recent neuro-psychological work on the generation of images. In the model developed at length by Stephen Kosslyn (1994), for example, the same backwards-projecting neural pathways which are used in normal perception to direct visual search, are used in imagination to induce an appropriate stimulus in the primary visual cortex (area V1), which is then processed by the visual system in the normal way, just as if it were a percept. So, on this account, the generation of imagery will involve at least sub-personal computations and inferences in the visual system, just as perception does. But that does *not* mean that the image is only available to us by means of such inferences. It is the self-induced pattern of stimulation in V1 which has to be interpreted by the visual system, on this account, *not* the image. Rather, the image is the *result of* such a process of interpretation. The image itself is the *output* of the visual system to central cognition, not the input. So it is entirely consistent with Kosslyn's account to say that our access to our own conscious images is wholly non-inferential (that is, that it does not even involve sub-personal computations of any sort).

4. In Defence of Premise 1: On Conscious Thinking

This first premise of the argument laid out in section 2 above stated that our mode of access to our own occurrent thoughts must be non-interpretative in character, if those thoughts are to count as conscious ones. This claim is by no means wholly uncontroversial; but it is accessible to, and defensible from, a number of different perspectives on the nature of consciousness.

For example, many philosophers, especially those writing within a broadly Wittgensteinian tradition, are apt to emphasize that we are *authoritative* about our own conscious mental states, in a way that we cannot be authoritative about the mental states of others (Malcolm, 1984; Shoemaker, 1988, 1990; Heal, 1994; see also Burge, 1996). If I sincerely claim to be in a particular mental state, then this provides sufficient grounds for others to say of me—and to say with justification—that I *am* in that state, in the absence of direct evidence to the contrary. Put otherwise, a sincere claim that I am in a particular mental state is self-licensing—perhaps because such claims are thought to be *constitutive* of the states thereby ascribed—in a way that sincere claims about the mental states of others are not. Now, it is very hard indeed to see how we could possess this kind of epistemic authority in respect of our own occurrent thoughts, if those thoughts were known of on the basis of some kind of self-interpretation. For there is nothing privileged about my standpoint as an interpreter of myself. Others, arguably, have essentially the same kind of interpretative access to my mental states as I do. (Of course I shall, standardly, have available a good deal more data to interpret in my own

case, and I can also generate further data at will, in a way that I cannot in connection with others—but this is a mere quantitative, rather than a qualitative difference.) So believers in first-person authority should also accept Premise 1, and maintain that our access to our own occurrent thoughts, when conscious, is of a non-inferential, non-interpretative, sort.

Premise 1 can also be adequately motivated from a variety of more cognitivist perspectives. On the sort of approach that I favour, a mental state becomes conscious when it is made available to a faculty of thought which has the power, not only to entertain thoughts about the *content* of that state (e.g. about an item in the world, perceptually represented), but also to entertain thoughts about the *occurrence* of that state (see Carruthers, 1996, ch. 7). When I perceive a ripe tomato, for example, my perceptual state occurs in such a way as to make its content available to conceptual thought about the tomato, where some of those concepts may be deployed recognitionally (e.g. *red*, or *tomato*). That state is then a conscious one, if it also occurs in such a way as to be available to thoughts about *itself* (e.g. 'It looks to me like there is a tomato there', or 'I am now experiencing red')—where here, too, some of the concepts may be deployed recognitionally, so that I can judge, straight off, that I am experiencing red, say. On this account, then, a conscious thought will be one which is available to thoughts about the occurrence of that thought (e.g. 'Why did I think *that*?'), where the sense of *availability* in question is supposed to be non-inferential, but rather recognitional, or at least quasi-recognitional.

It is worth noting that this account is fully consistent with so-called 'theory-theory' approaches to our understanding of mental states and events (which I endorse). On such a view, our various mental concepts (*perceive*, *judge*, *fear*, *feel*, and so on) get their life and significance from their embedding in a substantive, more-or-less explicit, *theory* of the causal structure of the mind (Lewis, 1966; Churchland, 1981; Stich, 1983; Fodor, 1987). So to grasp the concept *percept of red*, for example, one has to know enough about the role of the corresponding state in our overall mental economy, such as that it tends to be caused by the presence of something red in one's line of vision, and tends to cause one to believe, in turn, that one is confronted with something red, and so on. It is perfectly consistent with such a view, that these theoretical concepts should also admit of recognitional applications, in certain circumstances. And then one way of endorsing Premise 1, is to say that a mental state counts as conscious only if it is available to a recognitional application of some corresponding mental concept.

Amongst those who should *deny* Premise 1 will be some (but by no means all) of those who think of introspection on the model of outer perception (the difference, here, will then turn on how perception itself is conceived of, as will shortly emerge). For suppose that consciousness is mediated by the operation of some sort of internal self-scanning mechanism (Armstrong, 1968, 1984)—in that case it might seem obvious that our access to our own mental states is not crucially different from our access to the mental states of other people, and that such access must at least be partly inferential, con-

trary to what is claimed in Premise 1. This conclusion would be too hasty, however. For it is important to distinguish between personal-level inference, on the one hand, and sub-personal computation, on the other. The sense of 'inference' which figures in Premise 1 is not, of course, restricted to *conscious* inference; but it *is* restricted to person-level inference, in which a cognitive transition or process draws on the subject's current beliefs on the matters in hand. The claim is that we take our access to our conscious thoughts to be immediate, not necessarily in the sense that it depends upon no sub-personal computations, but rather in the sense that it does not depend for its operation upon any other particular *beliefs* of ours. In which case it will be possible for someone to think of the self-scanning mechanism as an isolated *module*, in the sense of Fodor (1983), which may well effect computations on its inputs, but which does so in a manner which is mandatory and hard-wired, and which is encapsulated from changes in background belief.

So, if someone conceives of introspection on the model of outer perception, then much may depend, for our purposes, on whether or not they have a *modular* conception of the latter. If so, then they should be happy to endorse Premise 1, since the self-scanning mechanism which produces awareness of our conscious mental states will operate independently of background belief (even if it embodies, itself, some sort of implicit theory of the mind and its operations), and so will not be *inferential* or *self-interpretative* in the intended sense. If, on the other hand, they think that all perception is theory-laden, in the sense that it is partly determined by the subject's explicit beliefs and changes of belief, then they *may* be committed to a denial of Premise 1, depending upon which kinds of belief are in question. Certainly, rejection of Premise 1 is not necessarily mandatory for such a person. For as we saw earlier, theory-theory accounts of our conception of the mental are consistent with Premise 1, provided that the theory-imbued concepts can also be deployed recognitionally. So one could claim that our quasi-perceptual access to our own mental states is theory-laden, while also maintaining that the access in question is non-inferential in character.

What really *is* inconsistent with Premise 1 is a view of our relation to our own mental states which makes the latter dependent upon our *particular* beliefs about our current environment or circumstances, or about our recently prior thoughts or other mental states. If my awareness that I am in some particular mental state depends, not just on recognitional deployment of theoretically embedded concepts, but also on inferences which draw upon my beliefs about the current physical or cognitive environment, then introspection really *will* be inferential, in a manner which conflicts with Premise 1. But it is, I claim, a presupposition of our common-sense conception of consciousness that our access to our conscious mental states is *not* inferential in this sense. Those who disagree can stop reading this paper here, since I shall say nothing further to persuade them of the falsity of their view—and yet the remainder of the argument depends upon it being false. Or better (or better for me): they should read what follows, but read it as having the

form of a conditional, to see what they *would* be committed to if they did *not* hold their particular theory of the nature of introspection.

Some people might allege that I have subtly begged the question in favour of my conclusion by writing as if consciousness were a unitary phenomenon. They may, for example, be keen to stress the difference between *phenomenal* consciousness and *reflexive* (or higher-order) consciousness, claiming that some mental states are conscious in the former sense and some only in the latter (e.g. Davies, 1993; Block, 1995). (States are phenomenally conscious which have phenomenal properties, or *feels*, like pains and experiences of red. States are reflexively conscious which are *available* or *accessible* to be thought about by the subject.) And then they may assert that the sort of immediacy of access to our conscious thoughts, which is claimed in Premise 1, is really only appropriate in connection with phenomenally conscious states. In which case it is being taken for granted that conscious thoughts must be imagistically expressed, and the only remotely plausible candidates for the images in question will be imaged natural language sentences, or 'inner speech'. So Premise 1, it may be said, just *assumes* that conscious propositional thought is conducted in natural language.

Let us grant the distinction between phenomenal consciousness and reflexive consciousness. Still, it would surely be a mistake to claim that the thesis of *immediacy*, expressed in Premise 1, applies only to the former. If reflexive consciousness is genuinely to be a form of *consciousness*, indeed, then the sort of access in question must be non-inferential and non-interpretative. We surely believe, for example, that there is all the difference in the world between entertaining, *consciously*, a jealous thought (even allowing that such a thought may lack phenomenal properties), and realizing, by interpretation of one's current behaviour, that one is acting out of jealousy. So, in insisting that we must have immediate knowledge of thoughts, too, if they are conscious, I am certainly not *assuming* that such thoughts must, of their very nature, be phenomenally conscious.

Moreover, there is no good reason to think that if we do have self-knowledge of mental states which are reflexively conscious without being phenomenally conscious, then such knowledge would have to be inferential or interpretative. For it is easy to imagine a possible mechanism, which would underpin the kind of immediate access which we take ourselves to have to our conscious occurrent thoughts, but without presupposing any sort of phenomenal consciousness. In particular, suppose that *thinking that P* were constituted by entertaining, in appropriate mode (that is: judgement, supposition, expression of desire, etc.), some Mentalese sentence 'S' which means that P. Then you could imagine a mechanism which operated by semantic ascent, in such a way that the occurrence of 'S' in the belief mode would automatically cause one to be disposed to entertain, in judgemental mode, the Mentalese equivalent of, 'I have just thought that S' (where this would, by hypothesis, be the same as thinking that one has just entertained the thought that P). But this would happen without our having any awareness of, or mode of access to, the fact that the sentence 'S' was used in the

expression of the original belief. That sentence would be *used* over again, embedded in the higher-order sentence which carried the content of the higher-order thought, but without the subject having any knowledge that it is so used. Such a mechanism would give us reliable non-inferential access to our own occurrent thoughts, without any sentences (let alone natural language sentences) having to figure as objects of phenomenal awareness.

5. In Defence of Premise 3: On Interpreting One's Own Inner Speech

Suppose, then, that inner speech is not constitutive of, but rather expressive of, propositional thinking. In that case the picture would be this: first a thought is entertained, in a manner which does *not* constitutively involve natural language (carried by a sentence of Mentalese, as it may be); and then that thought is encoded into a natural language sentence with the same (or sufficiently similar) content, to be displayed in auditory or motor imagination. By virtue of the conscious status of the latter, we thereby gain access to the underlying thought. But this access is not, I claim, of the kind necessary for that thought to count as a conscious one.

One argument for this conclusion is that the imaged natural language sentence can only give us access to the thought which caused it, through a process of interpretation. In order for me to know *which* thought it is that I have just entertained, when the sentence 'I just have time to get to the bank' figures in auditory imagination, that sentence needs to be interpreted, relying on cues provided by the context. These cues can presumably be both cognitive and non-cognitive in nature—what enables me to disambiguate the sentence may be other recent thoughts, or current goals, of mine; or it may be background knowledge of the circumstances, such as that there is no river nearby. Not that this process of interpretation is characteristically a conscious one, of course; quite the contrary. In general, as emphasized in section 3, the sentences entertained in inner speech certainly do not *strike* one as standing in need of interpretation; their phenomenology is rather that they are objects which are *already* interpreted. But interpretation there must surely have been, nevertheless. An imaged sentence, just as much as a heard sentence, needs to be parsed and disambiguated in order to be understood. So what figures in consciousness, on this account, is not a thought, but rather a *representation of* a thought; and a representation constructed through a process of interpretation and inference. This is, I claim, sufficient to debar the thought represented from counting as a conscious one.

It might be objected that one does not *need* to interpret the imaged sentence, in order to disambiguate it, since one *already* knows what one *meant* or *intended*. But this objection presupposes, surely, that we have non-inferential access to our own conscious, but non-verbalized, intentions. For if my only access to my own meaning-intentions were itself inferential, then it is very hard to see how their existence could help in any way to demonstrate that I have *non*-inferential access to the thoughts which I verbally articulate in

inner speech. But as we shall see in section 6, there is every reason to think that our only access to our occurrent unverbalized thoughts *is* inferential, by means of swift self-interpretation.

Another, related, argument for the conclusion that verbalized thoughts are not really conscious, if the thoughts themselves are distinct from their verbalizations, is this: in that case we would have essentially the same sort of access to our own occurrent thoughts as we have to the thoughts of other people, when we hear them speak. In both cases the communicative conception of language would have it that an occurrent thought causes the production of a natural language sentence, which is then represented and interpreted by the consumer of that sentence (another person, in the case of overt speech; the same person, in the case of inner speech). But in both cases the resulting representation of the content of the sentence (on this account, the underlying thought) strikes us, normally, with phenomenological immediacy. It is true that we do sometimes have to pause to reflect, before settling on an interpretation of someone else's utterance, in a way that we don't have to reflect to understand our own inner speech. This is presumably because the cognitive factors necessary to cue the disambiguation of another person's utterance are often not available to guide the initial process of interpretation. And it is also true that there is scope for *mishearing* in connection with the utterances of another, in a way that appears to lack any analogue in the case of inner speech. But these differences appear insufficient to constitute a difference in the *kind* of access achieved in the two cases.

It might be objected against the line being argued here that, if sound, it must also undermine the position defended in section 3—the position, namely, that if imaged sentences occupy the causal roles of occurrent thoughts (and hence are constitutive of thinking), then those thoughts can count as conscious ones. For it may be said that interpretation will need to take place in any case. Whether an imaged sentence is constitutive of an occurrent thought of ours, or caused by the occurrence of a thought existing independently of it, that sentence must still be subject to a process of interpretation. And I agree. But the difference lies in whether or not the process of interpretation occurs *upstream* or *downstream* of the event which occupies the causal role of the thought. According to the communicative conception, the process of interpretation occurs downstream of the thought—first a thought is tokened, which is then used to generate a natural language sentence in imagination which is interpreted; but the causal role of the initial thought, sufficient for it to qualify *as* that thought, is independent of what happens after it gets tokened. According to the cognitive conception, in contrast, it is quite the reverse. Here the hypothesis is that the causal role of the token thought in question is dependent upon its figuring as an interpreted image. It is the imaged (and interpreted) natural language sentence itself which causes the further cognitive effects distinctive of entertaining the thought in question.

It might also be objected that all the arguments of this section share a common assumption: namely, that the mechanisms which generate a mean-

ingful sentence of inner speech will involve some process of disambiguation and interpretation. But this assumption can be challenged. For why should not the content of the sentence of inner speech be determined, *non*-inferentially, by the content of the thought which causes its production? Why should not the sentence just drag its own interpretation with it, as it were—acquiring its content, not through any sort of process of inference, but simply by virtue of its causal connection with the underlying thought which it serves to express? One sort of answer to this challenge is to point out that this does not seem to be how imagination, in general, works—at least if we take visual imagination as our model. As was pointed out in section 3, our best current theories of visual imagery would have it that images are generated by the production of an input to the visual system, which is then interpreted by that system in the normal way. I am not aware that any similar work has been done on the generation of inner speech. But if similar mechanisms are involved, then one would expect that inner speech operates by one's producing an input to the language system, which is then interpreted by that system (in a manner which involves processes of parsing and disambiguation) in exactly the way that it would set about interpreting the speech of another person.

A more powerful answer to the above challenge is also available, however. For it is doubtful whether the assignment of content to sentences of inner speech *can*, even in principle, be determined in any other way than by means of a process of interpretation and disambiguation, drawing on the thinker's current beliefs. This is because the systems which produce, and which consume, such sentences must be distinct. Of course it is true that the inner sentence in question *has* content, independently of any act of interpretation, by virtue of its causal connection with the thought which produced it—just as my utterances have content whether or not you succeed in interpreting them. But this is no good at all to the system (or person) who has to make use of the generated sentence, or who has to draw inferences from it. For, by hypothesis, the consumer system for the sentence (in the sense of Millikan, 1984) lacks access to the thought which caused the production of that sentence. (If it did have such access, then it wouldn't need inner speech in order to gain access to the underlying thought.) The idea of a sentence 'dragging its own interpretation with it' is surely incoherent, in fact. If the mere fact of having been caused by a certain thought were sufficient to confer an interpretation on it, from the perspective of the consumer system, then one might just as well say that the mere fact that my *spoken* utterance is caused by a particular thought of *mine* is sufficient for *you* to interpret it. But that would be absurd. So, in conclusion: if the sentences of inner speech are distinct items from the thoughts to which they give us access, then it must follow that the sort of access in question does *not* have the kind of non-inferential immediacy necessary for those thoughts to count as conscious ones.

6. In Defence of Premise 4: Against Purely Propositional Conscious Thinking

Supposing, then, that 'inner speech' is neither constitutive of conscious thinking, nor gives us the kind of non-inferential access to the underlying thoughts which is necessary to render the latter conscious: is there any *other* way in which we might nevertheless have non-inferential access to our own occurrent thoughts? Many people believe so. They believe that we can entertain a thought and just know, immediately, that we have entertained it, without any sentence (or image) figuring in consciousness. Thus the introspection-sampling data provided by Hurlburt (1990, 1993) contains many reports of *purely propositional* thought, where subjects say that at the time of the beep they were thinking *that P*, but non-linguistically, without any sort of inner voice or inner speech. (And recall that we saw in section 4 how there are conceivable, Mentalese-involving, mechanisms which would make such purely propositional conscious thinking possible.)

I propose to argue that we are subject to a systematic illusion here. What we take to be non-inferential access to purely propositional thought is, in reality, the result of a swift bit of self-interpretation, which takes place so smoothly and quickly that we do not know that that is what we are doing. This has been convincingly demonstrated by a rich body of data coming out of the social psychology literature, where it has been found that there is a wide variety of circumstances in which subjects will confabulate self-explanations which are manifestly false, but without realizing that this is what they are doing (Nisbett and Wilson, 1977; Nisbett and Ross, 1980; Wilson et al., 1981; Wilson, 1985; Wilson and Stone, 1985). What follows are just a few salient examples.

First: when asked to select from a range of identical items (shirts, say), identically presented, people show a marked preference for items on the right-hand side of the display; but their explanations of their own choices never advert to position, but rather mention superior quality, appearance, colour, and so on. These explanations are plainly confabulated. (Remember, there is really no difference at all between the items.) And note that people's explanations, here, can be offered within seconds of the original choice. So the problem is unlikely to be one of memory (contrary to the suggestion made by Ericsson and Simon, 1980). Moreover, although the explanations are in fact elicited by experimenter questioning, there is every reason to think that they could equally well have been spontaneously offered, had the circumstances required.

Second: people who have been paid to play with some sort of puzzle or game report less intrinsic interest in it than those who do so purely voluntarily; but these reports do not correlate with the extent to which they are observed to play with it in their free time. It seems that people assume that they must enjoy the puzzle less, since, knowing the facts of the situation, they interpret their behaviour as motivated by payment rather than by enjoyment. But their explanations do not match their actual behaviour.

Third: people are also very poor at knowing which factors in a situation influence their evaluations or decisions, such as which aspects of someone's behaviour influenced their evaluation of his physical characteristics (appearance, etc.), or which aspects of a job-applicant's portfolio influenced their decision to call her for interview; and, interestingly, observers merely *told about* these studies make exactly the same erroneous judgements as do the subjects *in* them. Moreover, both groups (participants and observers) tend to make correct judgements when, and only when, the influential factor is recognized as such within common-sense psychology.

The best explanation of these data (and the explanation offered by Nisbett and Wilson), is that subjects in such cases lack any form of conscious access to their true thought-processes. (See also Gopnik, 1993, for a range of developmental data which are used to argue for the same conclusion.) Rather, lacking immediate access to their reasons, what they do is engage in a swift bit of retrospective self-interpretation, attributing to themselves the thoughts and feelings which they think they *should* have in the circumstances, or in such a way as to make sense of their own behaviour. And note, too, that in all but the first of the above examples, at least, the subjects do seem to act for reasons. So it is not open to us to say that subjects will confabulate self-explanations when, and only when, there are not any reasons for their actions for them to have access *to*. Who knows what cognitive mechanisms produce the bias for right-hand items in a display? But the other cases cited seem plainly to be caused by the subject's occurrent thoughts. And, indeed, in the third type of case it is possible to *discover* (or at least make reasonable inferences about) what those thoughts were, by correlating factors in the situation with subjects' choices across a range of studies.

It is important to emphasize that it does not *feel* to subjects in any of the above experiments as if they are engaging in self-interpretation. On the contrary, their self-attributions strike them with just the sort of phenomenological immediacy one might expect of a conscious thought (that is, one which is immediately and non-inferentially available, but unverbalized). But then nor does it *feel as if* we are interpreting *other* agents much of the time, either—rather, we just *see* much of their behaviour as intentional, and as imbued with certain thoughts. Indeed, our theory-of-mind faculty appears to have a number of the properties of a Fodorean module (Fodor, 1983): besides being at least partly innately specified (Baron-Cohen, 1995), its operation is both mandatory and fast. We often just *cannot help* seeing the behaviour of an actor on the stage as displaying anger, or fear, or whatever, despite the fact that we know him to be acting. And much of the time we are not aware of ourselves as having to interpret his behaviour, either, as being deceitful, or conciliatory, or whatever; rather, we just *see it that way*, immediately. So it is only to be expected that, when people engage in self-interpretation, this will often take place extremely swiftly, and without self-awareness of what they are doing.

Now, it does not follow from the psychological data that there is no such thing as purely propositional conscious thinking. From the fact that we *some-*

times engage in unwitting self-interpretation, in attributing unverbalized thoughts to ourselves, it does not follow that we *always* do. But there is, surely, a sound inference to the best explanation to this conclusion. For, rather than believe that those of our occurrent propositional thoughts which are not expressed in inner speech fall into two very different classes (which are nevertheless indistinguishable to introspection)—namely, those to which the subject has immediate non-inferential access, and those which are self-ascribed on the basis of swift self-interpretation—it is simpler to suppose that *all* such thoughts are only available through interpretation. Indeed, in the absence of some definite proposal about what might be the distinguishing characteristics of the two kinds of circumstances in which we form beliefs about our unverbalized thoughts, the hypothesis that we sometimes have non-inferential access to such thoughts, while sometimes engaging in self-interpretation, is too unformed to constitute a real competitor.

In fact, looking across the full range of the experimental data available, the one factor which seems to stand out as being common to all those cases where individuals confabulate false self-explanations is simply that in such cases the true causes of the thoughts, feelings, or behaviours in question are unknown to common-sense psychology. The best explanation of the errors, then, is that in all cases of unverbalized thought individuals are actually *employing* common-sense psychology, relying on its principles and generalizations to attribute mental states to themselves—the distinguishing feature of the cases where confabulation occurs is simply that in these instances common-sense psychology is itself inadequate.

This account is also further supported by the neuropsychological data, in particular the investigations of split-brain patients undertaken by Michael Gazzaniga and colleagues over many years (for a recent review, see Gazzaniga, 1994). For in these cases self-attributions are made in a way which we *know* cannot involve access to the thought-processes involved, but are made with exactly the same phenomenological immediacy as normal. And yet these self-attributions can involve the most ordinary and everyday of thoughts, being erroneous in a way which manifestly does *not* depend upon the inadequacies of common-sense psychology as such, nor upon any special features of the case—rather, these are just cases in which the psychology faculty lacks sufficient data to construct an accurate interpretation. So, if unwitting self-interpretation can be involved here, it can be involved anywhere. Let me briefly elaborate.

As is well known, in connection with split-brain, or commissurotomy, patients, information can be presented to (and responses elicited from) each half-brain independently. In the cases which concern us, both half-brains have some comprehension of language, but only the left-brain has access to the language-production system; the right-brain, however, is capable of initiating other forms of activity. When an instruction, such as 'Walk!' is flashed to the right-brain alone, the subject may get up and begin to leave the room. When asked what he is doing, he (that is the left-brain) may reply, 'I am going to get a Coke from the fridge'. This explanation is plainly con-

fabulated, since the action was actually initiated by the right-brain, for reasons to which, we know, the left-brain lacks access. In fact these and similar phenomena lead Gazzaniga (1994) to postulate that the left-brain houses a special-purpose cognitive sub-system, which he dubs 'the Interpreter', whose function is continually to construct rationalizing explanations for the behaviour of oneself and other people. And it then seems reasonable to suppose that it is this same sub-system which is responsible for the confabulated self-explanations in the data from normal subjects discussed by Nisbett and Wilson. Indeed, it is reasonable to suppose that this sub-system is responsible for *all* the access, or apparent access, which we have to our unverbalized thoughts.

I propose, then, that what are often described as purely propositional (non-verbal) thoughts, available to introspection (and hence conscious), are really the results of active self-interpretation. So even where the interpretations in question happen to be correct, and the thoughts are self-ascribed veridically, those thoughts are *not* conscious ones (at least, not on any of the various approaches to consciousness which entail a version of Premise 1). So, given the truth of Premises 1, 2, and 3 of the argument set out in section 2 above, it follows that we only ever engage in conscious propositional thinking at all if the cognitive conception of language is correct.

7. Conclusion

The arguments given in support of the three main premises of my initial argument (that is, the arguments given in sections 4, 5, and 6 in support of Premises 1, 3, and 4) were non-demonstrative, of course—so there are various ways in which any conclusions might be resisted. But if those premises are taken as established, then the subsidiary conclusion, Premise 5, follows: that if we engage in conscious propositional thinking at all, then natural language sentences must be constitutively involved in such thinking. In which case our choices are either: to *deny* that we ever engage in conscious propositional thinking, and to become eliminativists about such thinking; or: to *accept* that such thinking is conducted in natural language, and so to embrace at least a weak version of the cognitive conception of language. I propose that we should adopt the latter course. But for present purposes I shall be content if I have convinced my readers that there is a plausible case to be made for saying that they are faced with just this choice. (And note, too, that if we reject eliminativism, by accepting the cognitive conception of language, then this will not come cheap. For we shall then be committed to one or other of the three accounts of the causal role of inner speech which we sketched in section 3.)

Department of Philosophy
University of Sheffield

References

Armstrong, D. 1968: *A Materialist Theory of the Mind*. London: Routledge.
Armstrong, D. 1984: Consciousness and Causality. In D. Armstrong and N. Malcolm, *Consciousness and Causality*, 105–92. Oxford: Blackwell.
Baron-Cohen, S. 1995: *Mindblindness*. Cambridge, MA: MIT Press.
Berry, D. C. and Broadbent, D. E. 1984: On the Relationship between Task Performance and Associated Verbalizable Knowledge. *Quarterly Journal of Experimental Psychology*, 36A, 209–31.
Berry, D. C. and Broadbent, D. E. 1987: Explanation and Verbalization in a Computer Assisted Search Task. *Quarterly Journal of Experimental Psychology*, 39A, 585–609.
Block, N. 1995: A Confusion about a Function of Consciousness. *Behavioural and Brain Sciences*, 18, 227–47.
Burge, T. 1996: Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society*, 96, 91–116.
Carruthers, P. 1996: *Language, Thought and Consciousness*. Cambridge University Press.
Carruthers, P. 1998: Thinking in Language?: Evolution and a Modularist Possibility. In P. Carruthers and J. Boucher (eds), *Language and Thought*. Cambridge University Press, 94–119.
Chomsky, N. 1988: *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.
Chomsky, N. 1995: *The Minimalist Program*. Cambridge, MA: MIT Press.
Churchland, P. 1981: Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy*, 78, 67–90.
Clark, A. 1998: Magic Words: How Language Augments Human Computation. In P. Carruthers and J. Boucher (eds), *Language and Thought*. Cambridge University Press, 162–83.
Davidson, D. 1975: Thought and Talk. In S. Guttenplan (ed.), *Mind and Language*. Oxford University Press, 7–24.
Davidson, D. 1982: Rational Animals. *Dialectica*, 36, 317–27.
Davies, M. 1993: Introduction: Consciousness in Philosophy. In M. Davies and G. Humphreys (eds), *Consciousness*. Oxford: Blackwell, 9–39.
Dennett, D. 1978: How to Change your Mind. In his *Brainstorms*. Brighton: Harvester, 300–309.
Dennett, D. 1991: *Consciousness Explained*. London: Allen Lane.
Dummett, M. 1981: *The Interpretation of Frege's Philosophy*. London: Duckworth.
Dummett, M. 1991: *Frege and Other Philosophers*. Oxford University Press.
Ericsson, A. and Simon, H. 1980: Verbal Reports as Data. *Psychological Review*, 87, 215–51.
Fodor, J. 1978: Propositional Attitudes. In his *RePresentations* (1981). Brighton: Harvester, 177–203.
Fodor, J. 1983: *The Modularity of Mind*. Cambridge, MA: MIT Press.
Fodor, J. 1987: *Psychosemantics*. Cambridge, MA: MIT Press.
Frankish, K. 1998: Natural Language and Virtual Belief. In P. Carruthers and J. Boucher (eds), *Language and Thought*. Cambridge University Press, 248–69.
Gazzaniga, M. 1994: Consciousness and the Cerebral Hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press, 1391–404.

Gopnik, A. 1993: How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality. *Behavioural and Brain Sciences*, 16, 1–14.
Grice, P. 1957: Meaning. *Philosophical Review*, 66, 377–88.
Grice, P. 1969: Utterer's Meaning and Intention. *Philosophical Review*, 78, 147–77.
Heal, J. 1994: Moore's Paradox: A Wittgensteinian Approach. *Mind*, 103.
Horgan, T. and Woodward, J. 1985: Folk Psychology Is Here to Stay. *Philosophical Review*, 94, 197–225.
Hurlburt, R. 1990: *Sampling Normal and Schizophrenic Inner Experience*. New York: Plenum Press.
Hurlburt, R. 1993: *Sampling Inner Experience with Disturbed Affect*. New York: Plenum Press.
Kosslyn, S. 1994: *Image and Brain*. Cambridge, MA: MIT Press.
Levelt, W. 1989: *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Lewis, D. 1966: An Argument for the Identity Theory. *Journal of Philosophy*, 63, 17–25.
Lewis, D. 1969: *Convention*. Oxford: Blackwell.
Malcolm, N. 1984: Consciousness and Causality. In D. Armstrong and N. Malcolm, *Consciousness and Causality*. Oxford: Blackwell, 3–102.
May, R. 1985: *Logical Form*. Cambridge, MA: MIT Press.
Millikan, R. 1984: *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Nisbett, R. and Ross, L. 1980: *Human Inference*. New York: Prentice-Hall.
Nisbett, R. and Wilson, T. 1977: Telling More Than We Can Know. *Psychological Review*, 84, 231–95.
Peacocke, C. 1992: *A Study of Concepts*. Cambridge, MA: MIT Press.
Pinker, S. 1994: *The Language Instinct*. London: Penguin Press.
Russell, B. 1921: *The Analysis of Mind*. London: Allen and Unwin.
Searle, J. 1983: *Intentionality*. Cambridge University Press.
Shoemaker, S. 1988: On Knowing One's Own Mind. *Philosophical Perspectives*, 2, 183–209.
Shoemaker, S. 1990: First-Person Access. *Philosophical Perspectives*, 4, 187–214.
Stich, S. 1983: *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Vygotsky, L. 1934: *Thought and Language* (trans. Kozulin). Cambridge, MA: MIT Press, 1986.
Whorf, B. 1956: *Language, Thought, and Reality*. London: Wiley.
Wilson, T. 1985: Strangers to Ourselves. In J. Harvey and G. Weary (eds), *Attribution: Basic Issues and Applications*. New York: Academic Press, 9–38.
Wilson, T., Hull, J. and Johnson, J. 1981: Awareness and Self-Perception: Verbal Reports on Internal States. *Journal of Personality and Social Psychology*, 40, 53–71.
Wilson, T. and Stone, J. 1985: Limitations of Self-Knowledge. In P. Shaver (ed.), *Self, Situations and Social Behaviour*. New York: Sage, 167–83.
Wittgenstein, L. 1921: *Tractatus Logico-Philosophicus*. London: Routledge.
Wittgenstein, L. 1953: *Philosophical Investigations*. Oxford: Blackwell.