first place. The other way to pursue the idea would be to describe some type of cognitive processing that I have not considered, and argue that it is necessary for perceptual consciousness. But I have examined and given reasons for rejecting the only promising suggestions of this sort that I know of.

CONSCIOUSNESS AND CONCEPTS

Robert Kirk and Peter Carruthers

II—Peter Carruthers

I

Introduction. Robert Kirk believes that the distinctive feature of conscious, as opposed to non-conscious, experiences is that they are present to the organism's main decision-making processes. I maintain, in contrast, that conscious experiences are those that are regularly made available to be thought about in acts of thinking that are, in turn, regularly made available to further thinkings—where the idea of *availability* is to be explained in terms similar to Kirk's notion of *presence to*.[1] I shall refer to my account, in what follows, as the Reflexive Thinking (or RT) Theory of consciousness. I propose to concede at the outset that Kirk's account can accommodate the examples of the absent-minded driver and of blindsight. For while in-coming information in these cases is processed to quite a high degree of sophistication, and is somehow integrated with the subject's goals in such a way as to control behaviour, it is presumably not present to the subject's *main* decision-making processes—at least on the assumption that our main decision-making processes are the ones that are generally available for reporting in speech. I shall argue that Kirk's theory of consciousness has other major weaknesses, nevertheless. I will then discuss how the RT theory can handle the problem of qualia, maintaining that its success in this regard is a further mark in its favour.

1 See my 'Brute Experience' *Journal of Philosophy* 86 (1989). For more extended development, see *The Animals Issue: moral theory in practice* (Cambridge University Press, 1992), ch. 8.

II

Hierarchies of Control. It is a presupposition of Kirk's proposal, as of my own, that cognitive systems may be organized in such a way as to contain decision-making mechanisms at various different levels in a hierarchy of control. In the case of the absent-minded car driver, for example, something very like decision-making must be taking place. Information about the distribution of obstacles on the road is integrated with the driver's goals (to reach a destination, and to avoid an accident) so as to yield detailed decisions about the appropriate direction and speed of motion of the vehicle. Yet this all takes place, remember, while the driver's conscious attention is occupied wholly elsewhere. So while decisions are made, they presumably do not result from the activity of the driver's *main* decision-making mechanism, since those decisions are not available for reporting in speech, and since the activity of driving can be returned to conscious control at any time as the situation requires.

Similarly, something very like decision-making must be taking place in the case of a blindsighted person who reaches out to grasp an object on their blind side, in response to a request from an experimenter.[2] Decisions have to be made, in the light of non-conscious perceptual information concerning the position of the object, about the direction of motion of the arm, the orientation of the hand, and the precise point at which grasping should take place. Yet these decisions are presumably non-conscious in their turn. For the subjects in such cases will declare sincerely that they saw nothing, and were just guessing when they reached out their hands.

Besides conscious decision-making and non-conscious control of routine activities, it seems highly likely that there are yet other regions of cognition where decisions, or something very like them, get made. So it seems, to reiterate, that cognitive systems may be

2 Such graspings, of objects of various sizes and orientations at various distances, are performed with 80–90% of normal accuracy. I owe this information to personal communication from Anthony Marcel, Applied Psychology Unit, Cambridge, who tells me that subjects also proved remarkably good at catching a ball thrown towards them from their blind side.

organized into hierarchies of control, containing distinct decision-making mechanisms at various different levels within them.

III

First Objection. An initial problem with Kirk's account is that it only covers the conscious status of conscious experiences. Yet it is clear that the distinction between conscious and non-conscious mental states can find application right across the mental domain, applying to beliefs, desires, intentions, acts of thinking, and decisions just as much as to experiences. Indeed, this is implicit in the very idea of a hierarchy of control in cognitive systems, discussed in II above. Yet Kirk's account has no application in these cases. For presumably neither a belief nor a decision can be *present to* an organism's main decision-making processes, in the required sense—neither beliefs nor decisions will consist of analog information held in a short-term buffer store whose function is to make such information immediately accessible to the main decision-making processes.

There is a partially adequate response available to Kirk, that parallels the way in which the conscious status of (standing-state) propositional attitudes is handled within the RT theory. The latter claims that while conscious experiences are *present to* occurrent thought that is regularly made available to further thinking, a conscious belief or desire is one that is *apt to emerge in* an occurrent conscious thought—that is, in an act of thinking that is in turn made available to further thinking. Similarly, Kirk can claim that what is distinctive of conscious beliefs and desires is that they are apt to emerge in propositional episodes within the organism's main decision-making processes.

Such an account cannot be extended to explain the conscious status of (occurrent) conscious thinkings and decidings themselves, however. Kirk cannot say that conscious decidings are *present to* the subject's main decision-making processes, on the one hand, nor that they are *apt to emerge as* propositional episodes within those processes, on the other. The RT account, in contrast, applies equally to thinkings and decidings as to experiences and beliefs. The conscious status of an act of thinking or deciding, too, consists in

its being available to be thought about in acts of thinking that are regularly made available to further thinking.

It is hard to be confident of the force of this line of criticism of Kirk, however. This is partly because it is unclear how powerful the underlying intuition is, that the concept of consciousness is a unitary one. But it is also because Kirk himself fails to discuss the conscious status of anything but experiences. So one is left having to guess at the manner in which he might want to extend his account.

IV

Second Objection. What is genuinely puzzling about Kirk's proposal, however, is that mere height in a hierarchy of control could make the difference between an experience being conscious or non-conscious. Why should the mere fact that information is present to a cognitive system's *main* decision-making processes be sufficient to constitute it as a conscious experience? If this were so, it would mean that the experiences of any earth-worm or slug would be conscious ones, irrespective of the degree of conceptual sophistication displayed, provided that the organism does have a main decision-making mechanism to which its experiences are present. The experiences of the absent-minded car driver, on the other hand, or of someone who is blindsighted, would not count as conscious, on this proposal, despite the fact that those experiences are engaged in processes of decision-making that are considerably more complex and varied, and much more conceptually sophisticated, than those of the slug.

Another way to develop this point is to notice that on Kirk's proposal the status of an experience as conscious is a doubly relational feature of it, depending on how the decision-making processes to which it is present stand in the overall hierarchy of control. (I agree with Kirk that in order to count as an experience at all, information must be present to—and thus stand in relation to—some decision-making mechanism.) Indeed, if we suppose that the various decision-making mechanisms in a cognitive system are organized in modular fashion, then one and the same experience could change its status from conscious to non-conscious (or vice versa) merely by the decision-making mechanisms to which it is present being plugged into (or being unplugged from) a more

extensive cognitive system, in which some other decision-making mechanism is charged with overall control. This strikes me as intuitively absurd. For the conscious status of an experience is surely an intrinsic characteristic of it, in such a way that nothing could count as *that* experience which was not conscious.

This difficulty does not arise on the RT account. For an experience that is made available to be thought about in thinkings that are, in turn, regularly made available to further thinking will remain a conscious one, I claim, no matter how the relations may change between the decision-making processes in question and others. If my mind were to be (or were to become) a module in some more extensive cognitive system, for example—perhaps an institution or corporation that is a person, if this is possible—it would still remain the case that my experiences are conscious ones. For it would still be true that they are made available to thinkings that are regularly made available to further thinking. Conversely, suppose that I were completely blindsighted, and were then to have my capacity for reflexive thinking destroyed, while being left with the ability to engage in routine activities, like walking around obstacles or driving a car. This would not mean that those experiences that were formerly non-conscious would suddenly become conscious (because present to what would then be the system's main decision-making processes). For those blindsight experiences would still not be available to thinkings that are, in turn, regularly made available to thinking.

The only way out of the above difficulty, for Kirk, would be to drop the requirement that the decision-making processes to which an experience is present have to be the organism's main ones, in order for that experience to count as conscious. But this would then allow—indeed require—conscious experiences to proliferate wildly. It would entail, for example, that blindsighted people do indeed have conscious visual experiences on their blind sides, only not ones that are available to their main decision-making processes, which is why they deny having any such awareness. This proposal is hard to refute directly, but does seem extravagant. While introspection cannot be used to establish that there do not exist a whole multiplicity of conscious experiences at various levels in my cognition, of which *I* lack awareness, the idea had better be strongly theoretically motivated, if it is to be believable. But in fact it is just

such motivation that is lacking from Kirk's account, either in the modified form above, or, indeed, in its original version. That the RT account is genuinely theoretically motivated, in contrast, will emerge from the course of the subsequent discussion.

V

Third Objection. Perhaps the most serious difficulty with Kirk's account, however, is that it fails to make any connection between the conscious status of conscious experiences and the various features generally attributed to such experiences, particularly that they have distinctive subjective feels to them—that they possess qualia. There is something that it feels like to be the subject of a conscious experience. On this much, at least, there is widespread agreement. Where there is very considerable disagreement, concerns the ontological and epistemological status of qualia. Some have argued that the *what it is like*ness of an experience can find no place within a physicalist ontology.[3] Others, while accepting that there is something that a conscious experience is like, have denied that qualia have any of the mysterious properties, such as ineffability, that philosophers commonly attribute to them.[4] Others again have maintained (rightly, in my view) that qualia raise no threat to physicalism, arguing that to know what an experience is like is merely to have certain abilities with respect to it—to recognize, remember, and imagine it, for example.[5] These disputes need not concern us directly, though I shall return to them briefly in IX below. For what is left wholly unclear on Kirk's account, is why it should necessarily feel like *anything* to be an organism with perceptual information present to its main decision-making processes. I believe, indeed, that this is demonstrably *not* necessary. I propose to argue, in fact, that the subjective feel of experience

3 See Thomas Nagel 'What is it like to be a bat?' in his *Mortal Questions* (Cambridge University Press, 1979), and Frank Jackson 'Epiphenomenal Qualia' *Philosophical Quarterly* 32, 1982.

4 See Daniel Dennett 'Quining Qualia' in W. G. Lycan (ed.) *Mind and Cognition* (Blackwell, 1990), 519–47.

5 See David Lewis 'What Experience Teaches' in the Lycan volume referred to above, 499–519. See also my *Introducing Persons* (Routledge, 1986), ch. 5.

presupposes a capacity for reflective self-awareness that is omitted from Kirk's account.

It is plain that there could be—indeed there are—cognitive systems in which information about the environment is made available in analog form to a decision-making mechanism, but where that information is presented entirely transparently. That is to say, in such systems the decision-making mechanism can respond to the events in the system that carry information about the world, but *only* on the basis of the information that they carry. These events will have *characters for* the system, in Kirk's sense, in that their causal roles will be dependent upon their intrinsic properties. But the system itself will not represent those properties. Nor will it have any means of classifying or distinguishing between its experiences as such. Yet in order for there to be a subjective feel to experience, I shall argue, it is necessary that the system must contain, not only representations of the information carried by its perceptual states, but also representations of the states carrying that information. That is to say, the system should be capable of classifying the events within it that carry perceptual information *as* events carrying such information, where those events are recognized and distinguished from one another immediately, not known by inference or relational description.

Let me first present a general argument for this conclusion, arising out of reflection on the thesis I do not believe to be controversial—namely, that for there to be conscious experience there must be something that the experience *is like*. What is clear is that an organism could not *know* what its experiences were like unless it had reflective awareness of the states within it that carry perceptual information. But it might be wondered whether the organism has to *know* what an experience is like in order for there to *be* something that it is like. Why must qualitative feel presuppose reflective self-awareness? To see why it must, consider that to say that one thing is *like* another, in general, means only that there are certain similarities between them that are relevant to whatever purposes are in hand. Now, it is plain that there may be any number of respects in which one experience is like another that are not to the point in the present context—for example, similarities in the physical constitutions of the events in question. Rather, the similarities in question must be similarities *for the organism*. In

order for there to be something that an experience is like, it is not enough that there should be some objective resemblance between that experience and another—the organism itself must be capable of detecting that resemblance. For consider: to say that there is something that an object in the world—a chair, say—is like *for an organism* is just to say that the organism is capable of detecting resemblances and differences between that object and others within its world. In the same way, then, to say that there is something that an organism's experiences are like *for the organism* is to say that the organism must be capable of detecting resemblances and differences amongst its experiences. But then this is just to say that an organism whose experiences are like anything, in the relevant sense, must be capable of representing, and distinguishing between, its experiences as such.[6]

Let me now present an argument by example to make the same point. Suppose that blindsighted patients could become so familiar with their condition that they no longer have to guess what is where on their blind sides. Rather, with practice they can reach a point where they automatically form beliefs about the distribution of objects on their blind sides, and they automatically respond to the presence or absence of those objects in the course of their actions. In such a case we would want to say, I think, that the blind side perceptual information was present to the subject's main decision-making processes—being regularly and reliably made available to enter into the subjects' practical reasonings, as well as into the control of their actions. Yet it seems plain that these changes are consistent with the subjects in question remaining blindsighted—that is, with their blind side experiences remaining non-conscious ones, lacking any qualitative feel. But what would be missing here *except* a capacity for reflective awareness of their blind side experiences as such? If the subjects can be aware of what is where without their experiences being *like* anything for them, then this must be because they have no access to the states on whose

6 Note that the argument of this paragraph will survive translation into Nagel's preferred mode of expression, in which one speaks, not of what *an experience* is like, but of what it is like to be *the subject of* that experience. For the likeness in question must still be one that is available *to the subject*, which means that the organism in question must be capable of distinguishing between its own states of experience.

basis they acquire their knowledge—that is, because they have no awareness of their blind side experiences, as and when they occur.

I conclude, then, that Kirk's conditions are by no means sufficient for experiences to have qualitative feel, or for there to be anything that those experiences are like. On the contrary, conscious experience requires that the subject be capable of distinguishing between its experiences as such.

VI

The Status of the RT Theory. I propose to argue, in what follows, that instantiating the RT model is both necessary and sufficient for a creature to be a subject of qualia, or for its experiences to have distinctive feels. But it may be claimed at the outset that such a thesis is highly implausible. We can surely conceive, for example, of a creature that has conscious experiences but lacks the capacity to think about its own thoughts. Conversely, we can surely conceive of a creature that can think about its own thoughts while lacking the capacity for conscious experience. But in fact I do not intend to deny either of these claims. For I am not in the business of conceptual analysis—rather, the RT model is intended as a substantive theory of what consciousness *is*.

I subscribe to a version of the current orthodoxy in the philosophy of mind, according to which our concepts of the mental are theoretical ones, getting their life from their position within a complex and highly sophisticated implicit theory of the mind, embodied in our common-sense, or 'folk', psychology.[7] Elucidating a psychological concept will then involve articulating some relevant aspect of the theory. So there is no basis, here, for an objection that the RT account fails to establish a conceptual connection between consciousness and qualia. For that is not the business it is in.

In rejoinder it may be said that, even granting this, it is highly implausible that when ordinary people think of a conscious experience they think of one that is available to reflexive thinking. So the RT account does not even seem to be descriptively true of

7 This does not mean, however, that common-sense psychology is intended as a scientific theory of the mind, or that mental kind terms are used in the manner of natural kind terms. See my *Human Knowledge and Human Nature* (Oxford University Press, 1992), ch. 8.

common-sense psychology. But there are two replies to this. The first is that the theory of common-sense psychology is largely implicit, embodying principles and generalisations that are by no means fully articulated by its practitioners. Since much of the theory will not in any case be at the front of people's minds when they employ common-sense psychological concepts, it would hardly be surprising that when they think of an experience being conscious, they do not think of it in the terms characterised by the RT account.

The second reply is that the concept of consciousness that forms our target may, in fact, be a boundary-concept of common-sense psychology, setting the limits of its theoretical domain. For until very recently in the history of our species most people would have regarded—indeed, many did regard—consciousness as definitive of the mental. It is only with the development of scientific psychology—first Freudian, now Cognitive—that the idea that there are non-conscious mental states has come to enjoy general currency. Elucidating the concept of consciousness may then involve extending our common-sense theories beyond the original bounds of folk-psychology.

It is, therefore, too much to demand of the RT account that it should establish that it is either conceptually necessary or intuitively obvious that information present to a faculty of reflexive thinking will at the same time have a qualitative feel to it. Nevertheless, the RT account does need to provide us with enough to make it seem plausible that all and only those states that fit the theory will, of *natural* necessity, also have distinctive feels to them. It would also be a considerable advantage for the account if it could explain some of the claims that philosophers have perennially been tempted to make about qualia. These are, particularly, that qualia are non-relationally defined, ineffable, and knowable with complete certainty by the subject (and only the subject) whose states they are. For even if one regards these claims as false, a theory of consciousness should be able to explain why people who reflect on the nature of conscious experience so easily come to make them. Kirk's theory provides us with none of these things, nor is it at all easy to see how it might be extended or developed in such a way as to do so. It remains to be shown that the RT account can fare better.

VII

Necessary Conditions for Feel. In this section I shall argue that the availability of experiences to a reflexive faculty of thought, in which thinkings are regularly made available to further thinking, is a naturally necessary condition for those experiences to enjoy a subjective, qualitative, feel. I propose to argue, in fact, that the RT theory of consciousness articulates the sort of cognitive structure that must, of natural necessity, be in place if a creature is to be a subject of qualia. I shall approach this conclusion in a number of stages, starting at some apparent distance from my target.

I have already argued in V above that in order for there to be a feel to experience, the subject must be capable of reflective awareness of its states of experience as such. I now want to get from here, to the thesis that such a system would also have to be capable of thinking, reflexively, about its own acts of thinking. I shall do this by arguing that such capacities are components in a basic package of cognitive abilities (in something like Kirk's sense) that necessarily occur together. This would be a cognitive system that is capable of handling information across the mental domain, not only in terms of *content*, but also in terms of *mode of expression*. That is, it can respond to an occurrent thought, not just *qua* its truth-condition, but also *qua* act of thinking expressive of that truth-condition. And it can respond to an experience, not just *qua* state of the world represented, but also *qua* state representative of that state of the world. This would be a cognitive system in which thinkings are regularly fed back to be possible objects of further thinking (the crucial element in the RT definition of consciousness), *and* in which experiences are regularly made available to be thought about *qua* representing state, as well as *qua* state of the world represented.

I claim that each of the elements in such a system can only occur, of natural necessity, in the presence of the others.[8] But if this is so, it is far from obvious. For what is the connection between the two aspects? Why could there not be a cognitive system in which there

8 Strictly speaking, the claim only extends to cognitive systems that are instantiated in living organisms, which are a product of evolutionary selection. I shall say nothing about the possible cognitive powers of computers.

was thought about thought, but no capacity for thought about those features of experiences in virtue of which they carry information about reality? Or, conversely, why could there not be a system capable of thinking about those features, but not about its own acts of thinking? Discussion of the former possibility I shall defer to the next section, since it is an aspect of the question whether availability to thinking that is regularly made available to thinking is a *sufficient* condition for experiences to possess qualitative feels. Here I shall concentrate on the possibility that there might be qualia without a faculty for reflexive thinking.

My first claim is that in order for a cognitive system to have reflective self-awareness of the events within it that carry perceptual information, it must be capable of deploying a distinction between appearance and reality—between how things seem to the system and how they really are. For what, otherwise, would be the point of the system enjoying such self-awareness? It is surely naturally necessary—because a condition for it to have evolved in the first place—that a cognitive system capable of responding to its own perceptual information-bearing states as being intrinsically distinct from one another, in addition to being able to respond to the conditions in the world (or in the organism) that they represent, should also be able to distinguish between the way things seem to the system and the way they really are. It is very hard indeed to see what possible value in survival would accrue to a cognitive system that had self-awareness of its own states of experience, unless the system could make use of that awareness to begin to learn that in some circumstances certain of its information-bearing states are *not* reliable bearers of information. But to do that, it has to be able to make a distinction between appearance and reality.[9]

Someone may object that there can be no natural necessity here, since evolution operates by random gene mutation. But this is not to the point. My claim is not that there could never have been an *individual* creature having self-awareness of its own perceptual states without the capacity to distinguish between appearance and reality. It is rather, that such an individual would have had no

9 Colin McGinn makes essentially this point in *Mental Content* (Blackwell, 1989), 90–4.

advantages over other members of its species, in which case its capacity for self-awareness could not have become general to the species as a whole.

More challengingly, someone may object that a property can become general to a species without possessing survival value of its own, by virtue of being an epiphenomenon of some genetically determined property that *does* have survival value. So all members of a species might come to have self-awareness of the appropriate sort while lacking an appearance/reality distinction, yet without such self-awareness having any distinct survival value. The difficulty with this suggestion, however, is to give it flesh. For there are simply no plausible candidates in the offing, for the property of which self-awareness might be an epiphenomenon. It is more reasonable to believe that a capacity for self-awareness of one's own experiences can only appear as part of a basic package of capacities that carries value in survival—especially since this capacity is, plainly, of such complexity that it would have to have evolved incrementally, rather than through a single mutation.

My second claim is that in order to deploy a distinction between appearance and reality, in turn, it is naturally necessary—again because it is a condition for such cognitive structures to have evolved—that a system should be capable of thinking about its own thinkings. For of what use would be the distinction between appearance and reality, unless the system could also employ the concepts of truth and falsity? What would be the point of being able to distinguish between the way things appear and the way they are, unless the subject were able to put this distinction to work in practical reasoning, judging, for example, that beliefs formed under certain perceptual conditions are less likely to be *true* than those formed in other circumstances? But the concepts of truth and falsity, in turn, are only possible for a system that is capable of recognizing certain objects as *bearers* of truth and falsity. That is to say, only a system that can think, of a belief or occurrent thought, that it is true or false, can properly be said to possess the concepts of truth and falsity. But in order to think such things, the system's beliefs and occurrent thoughts must be *available* to thought—which is just the reflexive structure I claim to be a necessary condition (at least) of conscious experience.

My hypothesis is thus that it was the evolution of a system capable of thinking about its own thinkings on a regular basis that provided a necessary condition for the remainder. What a capacity to think about your own thinkings gets you, is an indefinitely improvable problem solving capability. This plainly has survival value in its own right. Such a capacity must involve, at a minimum, the distinction between true and false thought. With these structures in place (and only if they are in place) there would then be further survival value for the organism if it were to evolve a capacity for reflective awareness of its own perceptual states, becoming capable of distinguishing between appearance and reality.

VIII

Sufficient Conditions for Feel. I shall now argue that in any (evolved) cognitive system in which acts of thinking are regularly made available to further thinkings there will, of natural necessity, be a further capacity to recognize and distinguish between perceptual states. It might initially seem difficult to see how such a thesis could be established. For why could not a system have evolved with the capacity to think about its own thinkings, all of whose experiences are somewhat similar to those involved in blindsight? This would be a system in which perceptual information is somehow made available to be integrated into the subject's actions, yet where it is only, at most, the information carried that is available to thought, rather than the intrinsic features of the states that carry it. For a system could presumably have the notions of true and false thought without being able to employ an appearance/reality distinction, just by being capable of framing, at will, thoughts that conflict with the best available perceptual information. Such a system would still have considerable survival value. For, by being able to think about its own thinkings, it would be capable of exploring alternative courses of action in thought prior to executing any of them in reality, and of reflecting on and improving the pattern of its practical reasonings.

There are only two conceivable ways in which such a system might evolve, however. The first would be by development out of a less sophisticated cognitive system in which perceptual information is available to simple decision-making processes,

leading to a system capable of thinking about its own thinkings, but at the same time losing altogether the availability of perceptual information to the resulting faculty of reflexive thinking. It is plain that such a change must involve an overall loss to the organism, however, and thus would not become general to the species. For, whatever may be the advantages of being able to think about one's own thinking, the crucial thing for any organism is plainly the ability to relate its own thinkings to features of its immediate surrounding environment, formulating plans of action that are indexicalized ('I shall put this there', 'I must get out of the way of that', and so on). But this, by hypothesis, the envisaged creature could not do.

The second suggestion would again be that the system might evolve from a simpler entity in which perceptual information is available to decision-making processes, leading to a system capable of thinking about its own thinkings, this time retaining the availability of perceptual information to the resulting faculty of thought, but somehow *without* the system thereby becoming capable of thinking about its own perceptual states as such. (This would be a system analogous to the case of the confident blindsighted patients, envisaged in V above.) But what such a system could not do, would be to employ visual (or other forms of) imagination in the course of its conscious reflections. This is because imagination presupposes awareness of the intrinsic characteristics of perceptual states. Imagination, by its very nature, employs mental events that represent how something would appear in some sense modality. To form a visual image is to represent to yourself how something would look, to form an auditory image is to represent to yourself how something would sound, and so on. But this is only possible if you are capable of distinguishing between the look of the thing and the thing itself—that is to say, if you are already master of the distinction between appearance and reality. But to have this capacity is, surely, to be able to distinguish between your perceptual states in terms of their intrinsic qualities.

There is a good deal that a creature lacking in imagination could not do, plainly. But, even more than this, it is plausible that the capacity for reflexive thinking must itself presuppose imagination. This is because, in order to be capable of thinking about its own thinkings, in the required sense, a system must keep at least a brief

record of its thinkings *qua* symbols, as well as *qua* states of affairs signified. It must then have access to the productive medium in which its occurrent thoughts are expressed. But it is hard to see what events *could* serve as vehicles for our acts of thinking, except structures constructed in imagination. Occurrent thinkings will have to be expressed in imaged spoken or heard sentences, or imaged combinations of objects, if we are to have access to their forms as well as to their contents. So a cognitive system containing a faculty of reflexive thinking could not evolve, which lacks the ability to distinguish between its experiences as such, because it would also be without the power of imagination.

IX

The Qualia Problem. It may be objected that all of the above is, in any case, to miss the point. For even if I have been successful in showing that cognitive systems capable of recognizing and distinguishing between their perceptual states, and of thinking about their own acts of thinking, constitute a basic package (inevitably evolving together), this is not yet to find a place for qualia within such systems. What needs to be shown is that any such system capable of immediate recognition of its perceptual states will at the same time be a subject of qualia. What needs to be established is that such a capacity is a sufficient condition for a system to enjoy perceptual states that possess distinctive subjective feels.

I now propose to argue that qualia will emerge, of natural necessity, in any system where perceptual information is made available to thought in analog form, where the system is capable of recognizing its own perceptual states, as well as the states of the world perceived. For by postulating that this is so, we can explain why qualia should be so widely thought to be non-relationally defined, ineffable, and knowable with complete certainty. I shall argue, in fact, that any subject who instantiates such a cognitive system will very naturally come to form just such beliefs about the intrinsic characteristics of its perceptual states. I know of no better way of arguing that a capacity for thinking about thinking is a naturally sufficient condition for the enjoyment of experiences possessing a subjective feel.

Let us first consider the thesis of non-relational definition for qualia terms. A system instantiating the RT theory would have the capacity to classify informational states according to the manner in which they carry their information, not by inference or description, but immediately. The system would be able to recognize the fact that it has an experience of red, say, in just the same direct, non-inferential way that it can recognize red. (This is just what it means to say that perceptual states are available to conscious thought, in the sense I intend.) Then absent and inverted qualia will immediately be a conceptual possibility for someone applying these recognitional concepts. If I instantiate such a system, I shall immediately be able to think '*This* type of experience might have had some quite other cause', for example.

Does this then count against the acceptability of the functionalist conceptual scheme that forms the background to my own account of consciousness? If it is conceptually possible that a red quale should regularly be caused by perception of blue sky, then does this mean that the crucial facts of consciousness must escape the functionalist net, as many have alleged? I think not. For, as was pointed out earlier, the RT account is not in the business of conceptual analysis, but of theory development. So it is no objection to that account, that there are some concepts of the mental that cannot be analysed in terms of functional role, but are purely recognitional, provided that those concepts, and the states they recognize, can be adequately characterised within the theory. All that is needed at this point, in fact, is that it can plausibly be held to be metaphysically necessary, for us,[10] that the quale of an experience of red should be caused, normally, by perception of red. If, as the RT account suggests, the quale of an experience of red just *is* analog information about red, presented to a cognitive apparatus with the power to classify states as information carriers, as well as to classify the information carried, then there can be no world in which the one exists but not the other. For there will, in fact, be no 'one' and 'other', but only one state differently thought of.

10 I say 'for us' because it might be held that qualia would vary with varying types of physical instantiation of functionally equivalent perceptual systems—for example, between carbon-based and silicon-based. On this, see William Lycan *Consciousness* (MIT Press, 1987), ch. 5.

Now, given the thesis of non-relational definition, the supposed ineffability of qualia is easily accounted for on the RT approach. For the recognition-instances of qualia concepts cannot, of course, be exhibited to another person. Yet any attempt to describe in relational terms the character of a quale will seem to miss what is essential to the latter. Moreover, it is a general feature of cognitive systems to which perceptual information is presented in analog form, that the system will be capable of discriminations that slip through its conceptual net. For example, imagine yourself watching the leaves of a tree shimmering in the breeze. You will be able to discriminate subtle changes in the pattern of movement, and will be aware of the distinctive quality of each pattern, that you are incapable of describing further. In the case of colour perception, similarly, you will be able to discriminate shades from one another where you are incapable of describing the precise difference between them, having to resort to such generalities as 'a slightly darker shade of red'. You will also be aware of the distinctive quality of each shade without being able to describe it other than as 'the shade of *that* object over there'.

Equivalently, then, in the case of awareness of the qualities of the experience itself—you will be able to recognize and respond to subtle distinctions, where you lack the concepts to express the precise difference in question. All you will be able to say is something like 'it is what it is normally like to perceive *that* shade of red as opposed to *that* shade of red'. And you will be aware of the distinctive quality of each perceptual state without being able to describe it other than as 'the way it feels to see *that* shade'. Note that this description is relational—it describes the feeling in terms of its normal cause. So anyone who thinks that qualia are not relationally defined will believe that the crucial characteristic of a subjective feel must remain wholly inexpressible. All we can do is indicate what that feeling is indirectly, by its relationships with other things.

Finally, in order to explain the temptation to think that qualia are knowable with certainty, we can make use of the idea that perceptual information, and the marks by which it is carried, are *present to* thinking that is available to further thinking. There might then easily seem to be no space for error in the classification of those states, other than conceptual error. If all that is involved, when one recognizes a quale, is an act of classifying a state that is directly present to the classifier mechanisms, then provided that the classifier is in order, it can seem that there is no further room for mistake.

Although this picture is tempting, it is erroneous. For there may be ways in which a classifier mechanism can cease to operate normally that are not dramatic enough for us to say the system has thereby lost its grasp of the concepts it employs in its judgements. For example, it might prove to be the case that mood can have an effect upon colour judgements. Perhaps anger makes us literally see red, in so far as it slightly skews our colour judgements towards the red end of the spectrum. In which case, it will equally have an effect upon judgements of colour qualia. Then, knowing that this is so, you may have grounds to doubt your qualia judgements.

What emerges from the discussion of the last three sections is that the RT theory of consciousness is not only superior to Kirk's in the various respects indicated earlier, but that it can also accommodate the distinctive facts of consciousness—namely, that conscious experiences have characteristic feels to them, that many philosophers have thought must be non-relationally defined, ineffable, and knowable with certainty. This lends yet further support to the proposal that conscious experiences attain their status as such by forming part of a basic cognitive package in which perceptual states are made available to thinking that is, in turn, regularly made available to further thinking.[11]

11 I am grateful to David Archard, George Botterill, Peter Smith, and Tim Williamson for written comments on an earlier draft. Some of the material in this paper was presented to seminars at the Universities of Sheffield, London (University College and King's College), and Nottingham. I am grateful to all those who participated in the subsequent discussions.