FORUM 2 updated: 15 May 2002 http://www.swif.uniba.it/lei/mind/forums/forum2.htm

The Book

Peter, Carruthers. *Phenomenal Consciousness: A Naturalistic Theory.* Cambridge: Cambridge University Press, 2000. *More details from the publisher's site*

Author's précis

Peter Carruthers (homepage) is professor of Philosophy at the Department of Philosophy at the University of Maryland College Park.

Commentaries and replies

Colin Allen, Evolving Phenomenal Consciousness - Carruthers's reply.

José Luis Bermúdez, Commentary - Carruthers's reply - Reply to Carruthers: Properties, first-order representationalism and reinforcement.

Joseph Levine, Commentary - Carruthers's reply.

William Seager, Dispositions and Consciousness - Carruthers's reply.

Related Links

Alex Byrne Review of *Phenomenal Consciousness*, by Peter Carruthers, *Mind*, forthcoming. [PDF, 18K] http://web.mit.edu/abyrne/www/CarruthersReview.pdf

Précis of

Carruthers, Peter. *Phenomenal Consciousness*. Cambridge: Cambridge University Press, 2000.
ISBN 0521 78173 6

Peter Carruthers

University of Maryland, College Park
Homepage

Most contemporary philosophers of mind think that mental states are physical states of the brain, characterised in terms of their causal roles; and many hope that our common-sense conception of the mind can be incorporated smoothly into science. These are beliefs and hopes which I share. But philosophers such as Nagel (1986) and McGinn (1991) have argued that consciousness – particularly phenomenal consciousness, or the sort of consciousness which is involved when one undergoes states with a distinctive subjective phenomenology, or 'feel' – is inherently, and perhaps irredeemably, mysterious. And many would at least agree with Penrose (1994) and Chalmers (1996) in characterising consciousness as the 'hard problem', which forms one of the few remaining 'final frontiers' for science to conquer. Yet there have also been a plethora of attempts by philosophers and psychologists at explaining consciousness in natural terms. These debates have attracted a great deal of interest, both throughout the academic community and amongst the wider public.

This book reviews and contributes to these debates, with the overall objective of defending a particular kind of naturalistic (scientifically acceptable) explanation of phenomenal consciousness – namely, dispositionalist higher-order thought theory. My view is that phenomenal consciousness consists in a certain sort of intentional content ('analog', or fine-grained), held in a special-purpose short-term memory store in such a way as to be available to higher-order thoughts about the occurrence and nature of those contents; and that in virtue of such availability (given the truth of some or other form of 'consumer semantics') all of those contents are at the same time higher-order ones, acquiring a dimension of *seeming* or *subjectivity*. While the problem of phenomenal consciousness may indeed be hard, it is by no means insuperable; indeed, I claim to have provided a solution to it within the pages of this book.

Chapter 1 Assumptions, distinctions, and a map

In this opening chapter I lay out some of my background assumptions, introduce a number of important distinctions, and outline the direction which the discussions of later chapters will follow.

1. *Physicalism and naturalism*

   In this section I briefly explain and defend two default assumptions - physicalism and naturalism - which form the background to the problem of phenomenal consciousness. It is these assumptions which appear to be challenged by the very existence of phenomenal consciousness, as we shall see in chapters 2 and 3.

2. *Functionalism and theory-theory*

   The assumptions in section 1 above relate to the metaphysics of the mind. In this section I say something about how I take the mind to be conceptualised, or conceived of. I assume that some kind of functionalism - *viz*. a form of theory-theory - provides the best account of the way in which we conceptualise mental states. Again the position is not entirely mandatory, and again some of the main challenges come from considerations having to do with phenomenal consciousness. But the advantages of functionalism as an account of the mind (*viz.* its metaphysical neutrality – hence

allowing interactive dualism to be a conceptual possibility – and its solution to the problem of other minds) mean that it should not be given up lightly.

3. *Some distinctions: kinds of consciousness*

   There are a number of different notions of consciousness and/or a number of different kinds of use of the term 'conscious' which need to be distinguished carefully from one another. Failure to draw the right distinctions, and/or failure to keep the different notions apart, has vitiated much work in the area.

   The most basic distinction is between *creature*-consciousness (or *perceptual*-consciousness) and *state*-consciousness. The latter then admits of a number of conceptually-different kinds, including *phenomenal*-consciousness and various forms of *access*-consciousness (first-order and higher-order). The project of the book is to reductively explain phenomenal consciousness in terms of higher-order access-consciousness.

4. *A route map: the tree of consciousness*

This section lays out all the important theoretical positions in respect of the explanation of phenomenal consciousness, ranging from 'no-explanation' theories, through neurological explanations, first-order representationalist theories, and a variety of higher-order theories. These are related to the sequence of discussion in the book.

Chapter 2 Perspectival, subjective, and worldly facts

Many have alleged that phenomenal consciousness can neither be accommodated within a physicalist world-view, nor reductively explained in physical terms. In this chapter I confront some of these 'mysterian' arguments, concentrating on those which are more metaphysical in nature. If it is to be possible to provide a naturalistic explanation of phenomenal consciousness, as I intend, then all of these arguments must be flawed.

1. *Perspectival and 'myness' facts*

   Nagel (1974) is often credited with putting the problem of phenomenal consciousness on the map. In this section I consider his main arguments. I show that they fail, and that their failure results from conflating together notions which should be kept distinct - in particular, many of the arguments involve a conflation of *sense* and *reference*.

2. *On facts and properties*

   This is a vital section of the book. In it I argue that we can and should distinguish between two different notions of 'fact' and of 'property'. One conception is *thin*, and is arguably required for semantic theory. On this conception, facts are just true thoughts, and properties are functions from possible worlds to extensions. And on this conception, the identity and distinctness of properties can be discerned *a priori* by means of thought-experiments. But the other conception is *thick*, and is arguably required for accounts of *change*, and to serve as the *relata* in scientific laws. On this account, what properties there are in the world is ultimately a matter for science to answer, and questions of identity and distinctness of properties cannot be resolved *a priori*. Moreover, it is this 'thick' notion of *property* which is relevant to questions of reductive natural explanation.

3. *Necessary identities*

   This section discusses Kripke's (1972) famous modal arguments against physicalism. It argues that they do not succeed. Token-identities are not ruled out, since all the arguments for their contingency rely on conceivability-experiments. Type-identities are not ruled out, since there may be no unitary 'thick' mental properties common across the imagined worlds. And claims that mental properties may be physically *constituted* are left untouched, which is all that is strictly required for reductive explanation.

4. *Logical supervenience*

This section confronts one of the main arguments in Chalmers (1996), that phenomenally conscious states do *not* supervene on the physical facts in the way required for them to be physically constituted. It shows that the argument turns crucially on the 'thin' conception of properties as functions from worlds to extensions, and should thus be rejected by any would-be reductive naturalist.

Chapter 3 Explanatory gaps and qualia

In this chapter I continue my review and rebuttal of 'mysterian' arguments concerning phenomenal consciousness, focusing particularly on those which are epistemic in nature, having to do with possibilities of explanation, knowledge, or understanding.

1. *Cognitive closure*

   This section discusses and criticises McGinn's (1991) view that consciousness, while not intrinsically (metaphysically) mysterious, must always remain mysterious to us, because the answers to the questions we can frame here are cognitively closed to us.

   One general moral to emerge is that any explanation of phenomenal consciousness should be top-down and incremental in nature. (Somehow the rumour has got around, and become entrenched, that the problem of phenomenal consciousness is the problem of explaining how subjective properties can be constituted by processes in the brain; and most proposals on the market attempt to relate brain processes directly with properties of phenomenal consciousness.) In contrast, reflection on general scientific methodology suggests that we should initially seek our explanations in terms of the level immediately below our target – in this case intentional or computational psychology – which we will then, in turn, attempt to relate to the level below that, and so on until we ultimately reach processes which can be described in terms of the operations of neurons in the brain.

2. *Explanatory gaps*

   This section discusses and criticises Chalmers' (1996) argument that there is an unbridgeable 'explanatory gap' between phenomenal consciousness and all other worldly phenomena. I agree with him that reductive explanations normally work by specifying lower-level mechanisms for fulfilling some higher-level function. And I agree that we at least have *available* to us purely-recognitional concepts of phenomenally conscious states. But I disagree with the conclusions which Chalmers draws from these facts. His mistake is to assume that a given property or state can only be successfully reductively explained if the proposed mechanisms are what we might call *immediately cognitively satisfying*, in the sense that they mesh with the manner in which those states are conceptualised. While the 'explanatory gap' is of some *cognitive* significance, revealing something about the manner in which we conceptualise our experiences, it shows nothing about the nature of those experiences themselves.

3. *The knowledge argument*

   This section outlines and criticises Jackson's (1982, 1986) famous 'knowledge argument'. My diagnosis is this: the knowledge-argument only *seems* compelling because we covertly read the 'complete knowledge' component of the argument in the *thick* sense (that is: 'Concerning every worldly – thickly individuated – fact about colour vision, Mary knows the truth of a thought representing it'), but then we take the claim about Mary's *incomplete* knowledge of colour-experience in the *thin* sense (that is: 'Concerning some conceptual representation of colour-experience, Mary does not know its truth'). The argument commits a fallacy of equivocation.

4. *Inverted and absent qualia arguments*

This section evaluates a variety of inverted-spectrum arguments for the conclusion that our experiences possess intrinsic *qualia* (non-representational, non-relational, properties). Mere conceivability arguments are easily dismissed, but natural-possibility arguments due to Shoemaker (1981) and Block (1990) are more challenging. They can, however, be decisively answered if intentional content can be *narrow* as well as *wide.* This is left over for discussion in the next chapter.

Chapter 4 Naturalisation and narrow content

In this chapter I begin to survey the prospects for a naturalistic account of phenomenal consciousness, taking us through some of the initial options. Attention quite soon comes to focus on theories which employ some combination of *causal role* and *intentional content* – that is, theories which are both functional–boxological and representational in character. I then suggest that intentional content should be characterised *narrowly* for purposes of psychological explanation in general, and for deployment in proposed reductive explanations of phenomenal consciousness in particular.

1. *Neural identities and consciousness boxes*

   This section argues that neural identities, even if true, cannot provide reductive explanations of phenomenal consciousness. It also argues that the postulation of a consciousness *box* - again, even if correct - cannot provide a sufficient explanation. Rather, the account will need additionally to advert to the *contents* of the box.

   At the end of this section the main *desiderata* for a successful explanation of phenomenal consciousness are set out. Such an account should (1) explain how phenomenally conscious states have a *subjective* dimension; how they have *feel*; why there is something which it is *like* to undergo them; (2) why the properties involved in phenomenal consciousness should *seem* to their subjects to be *intrinsic* and non-relationally individuated; (3) why the properties distinctive of phenomenal consciousness can *seem* to their subjects to be *ineffable* or indescribable; (4) why those properties can *seem* in some way *private* to their possessors; and (5) how it can *seem* to subjects that we have *infallible* (as opposed to merely *privileged*) knowledge of phenomenally conscious properties.

2. *Naturalisation by content*

   This section briefly reviews various kinds of attempt to *naturalise* intentional content. It argues that intentional properties are already in good natural standing by virtue of figuring in scientific- intentional psychology. While there are issues concerning the integration of such psychology with the rest of science, we have good reason to believe that this can be done, and so good reason to believe that the property of *intentionality* is both natural and real. It is therefore fit to serve in a naturalistic explanation of phenomenal consciousness.

3. *Wide versus narrow content*

   In this section the distinction between wide and narrow content is explained, and the coherence of a notion of narrow content is defended. Then it is briefly argued that the notion of content employed for purposes of psychological explanation should be (and is) narrow (these arguments are pursued more fully in other publications - see especially Botterill and Carruthers, 1999).

4. *Phenomenal consciousness and narrow content*

This section rounds off the chapter by discussing ways in which narrow content can be deployed in reductive explanations of phenomenal consciousness, and by re-visiting the inverted spectra examples from chapter 3.

Chapter 5 First-order representationalism

This is the first of two chapters which assess the prospects for a naturalistic explanation of phenomenal consciousness in first-order representational (FOR) terms, focusing particularly upon the accounts presented by Dretske (1995) and Tye (1995). Part of the point of these discussions is to develop an account of first-order perceptual contents which can then be fed, as a component, into the higher-order theories to be discussed in chapters 7, 8 and 9.

1. *FOR theory: elucidation*

   This section outlines Tye's PANIC theory (PANIC is for Poised Abstract Non-conceptual Intentional Content). It focuses especially on showing how the theory can handle the case of bodily sensations like pain, and emotions like fear. Tye's view that pains are best understood on the model of secondary qualities like redness is set out and defended - so the pain itself is what is *represented in* a state of feeling pain, just as redness is what is represented in a state of seeing red.

2. *FOR theory: defence*

   In this section the main arguments supporting FOR theory are set out. Such theories explain the so-called 'transparency' of experience, they mesh with our intuitions regarding the consciousness of animals, they admit of plausible evolutionary explanation, and they can meet at least some of the desiderata for a successful theory, set out in 4:1.

3. *Non-conceptual content versus analog content*

   This section tackles the proper characterisation of the contents of perception. Are they non-conceptual, as Tye would have it? or merely analog, but imbued with concepts? The section argues that the latter option is preferable. It also goes on to discuss the sense in which perceptual contents can be concept-involving, given that they can be non-judgemental.

4. *More varieties of FOR theory*

It is manifest that FOR theories admit of more varieties than are actually represented in the published literature. One set of options comes from the different ways of drawing the contrast between belief and perception, reviewed in section 3 above. Another choice concerns whether the intentional contents appealed to by FOR theory should be individuated widely (externally) or narrowly (internally). Both Tye (1995) and Dretske (1995) endorse forms of externalism about content. But the considerations adduced in chapters 3 and 4 above make it seem likely that a FOR theorist both can and should appeal, rather, to narrowly individuated contents.

In this section two further sets of options are considered. One is whether a FOR theory of phenomenal consciousness should adopt a reductive, or rather a non-reductive, account of intentional content. The other is whether the intentional content of perception is best explicated in terms of informational (that is, causal co-variance) relations to the environment, or rather in terms of some or other form of what Millikan (1984) calls 'consumer semantics' – either teleosemantics, or some sort of functional or inferential role semantics. The latter alternative is defended in each case.

Chapter 6 Against first-order representationalism

This chapter sets out the case against all FOR accounts, of whatever variety. It focuses merely on the *first-orderness* of such theories, and the argument turns crucially on the real existence of non-conscious experience. The first two sections of the chapter are concerned to argue for the reality of non-conscious experience, from both common-sense and scientific perspectives. The final two sections then develop the argument against FOR theory, in the form of a trilemma. The upshot is that FOR theory fails because it cannot really explain the *feel*, or 'what-it-is-likeness', of phenomenally conscious experience.

1. *Non-conscious experience: the case from common-sense*

This section presents a common-sense argument for the reality of non-conscious experience, using examples of absent-minded activity and experiences undergone while sleeping. It also sketches an outline of the *two-layered mind* which seems supported by such arguments.

2. *Non-conscious experience: the scientific case*

   This lengthy section contains some discussion of the now-familiar phenomenon of *blindsight*. But most of it is devoted to exposition of Milner and Goodale's (1995) *dual-function* theory of vision, and some of the evidence in its support. On this view the parietal-lobe stream of visual analysis is to produce ego-centric representations for use in the on-line guidance of movement, whereas the temporal-lobe stream is to produce allocentric representations for conceptualisation, planning, and consciousness. The outputs of the parietal-lobe steam are not conscious, although they guide movement in just the sort of fine-grained way which fits one aspect of our intuitive idea of experience.

3. *A trilemma for FOR theory*

   One horn of the trilemma is to dismiss the data. This is dealt with briskly. Another is to accept that the outputs of the parietal system aren't phenomenally conscious, and to try explaining *why* they aren't in purely first-order terms. The problem here is to explain why availability to concepts and conceptual thought should transform intentional contents from ones which aren't phenomenally conscious into ones which are. I argue that no adequate explanation can be forthcoming here. The final horn of the trilemma is discussed in section 4.

4. *Non-conscious phenomenality?*

The final option is to insist that the contents of the parietal stream are phenomenally conscious, but unknowingly so to their subjects because they are not *access-conscious*. This option is hard to believe, and various additional objections are raised against it.

Chapter 7 Higher-order representationalism: a first defence

In this chapter I take the first steps towards a defence of higher-order representational (HOR) accounts of phenomenal consciousness. I argue that these accounts have considerable explanatory advantages over first-order representational (FOR) theories, and that they have the resources to rebut a number of potentially-devastating objections.

1. *Overview and preliminaries*

   This section gives a brief review of the tasks ahead, and reminds the reader of the different varieties of HOR theory (first sketched in chapter 1).

2. *HOR theory and qualia irrealism*

   This section shows how HOR theory can explain *desiderata* (2) through (5) from 4:1 - explaining why people are tempted to believe (falsely) that our experiences posses *intrinsic* properties, which are *ineffable*, *private*, and known with *certainty* by the subject. The challenge of explaining the core defining feature of phenomenal consciousness - its *subjectivity*, *feel*, or *what-it-is-likeness* - is held over until chapter 9.

3. *Of animals, infants, and the autistic*

   This section deals with a seemingly-powerful objection to HOR theory, namely that it entails that non-human animals (as well as human infants, and perhaps also autistic people) are lacking in phenomenal consciousness. It argues that the objection is entirely without force, since grounds for attributing phenomenal consciousness to animals are lacking; and since it is easy to explain why our intuitions that they *must* be phenomenally conscious are so powerful. (Roughly, the explanation runs: we correctly attribute perceptual experience to animals; but when we try to imagine

those experiences from the inside, we inevitably imagine *conscious* experiences, because our own acts of imagining are phenomenally conscious.)

4. *Moral consequences?*

This section tackles another seemingly-powerful objection to HOR theories, this time that they must withhold moral significance from animals. Here I argue that the conclusion does not follow. It may be that first-order - non-phenomenal - frustrations of desire are the most basic and appropriate objects of sympathy and moral concern.

Chapter 8 Dispositionalist higher-order thought theory (1): function

By this time I have argued that some form of higher-order representational (HOR) theory of phenomenal consciousness is to be preferred to any more modest first-order (FOR) approach. It remains to adjudicate between the different forms of HOR account. In the present chapter I deploy a variety of functional and evolutionary considerations to argue that dispositionalist higher-order thought (HOT) theory is greatly preferable to both actualist HOT theory, on the one hand, and to higher-order experience (HOE) theory on the other.

1. *Higher-order experience (HOE) theory*

   In this section I critically examine (HOE) theories, of the sort defended by Armstrong (1968, 1984) and Lycan (1987, 1996). These are 'inner sense' models of phenomenal consciousness. They postulate a set of inner scanners, directed at our first-order mental states, which construct analog representations of the occurrence and properties of those states. I argue that inner-sense theories are functionally and evolutionarily implausible by comparison with higher-order thought (HOT) accounts. The basic objection is that such inner scanners would have to be computationally complex, but that there are no plausible explanations of why they might have evolved, or of what they might be *for*.

2. *Actualist HOT theory*

   This section critically examines the form of actualist HOT theory proposed by Rosenthal (1986, 1993). On this account an experience of mine is conscious if and only if it is actually causing an activated higher-order belief that I am undergoing that experience, and causing it non-inferentially. The main difficulty for this account is the *objection from cognitive overload* - what is to explain the huge number of higher-order beliefs which would have to be caused by any given phenomenally conscious experience, given the richness and detail which is standardly present in such experience? This objection is developed at some length. But it depends upon an assumption of experiential *richness*, challenged by Dennett (1991). This is left as an issue to be returned to in chapter 11.

3. *Dispositionalist HOT theory*

   In this section I develop and defend a dispositionalist form of HOT theory, according to which the conscious status of an experience consists in its *availability* to HOT. As with actualist HOT theory, in its simplest form we have here a quite general proposal concerning the conscious status of any type of occurrent mental state, which becomes an account of phenomenal consciousness when the states in question are experiences (or images, emotions, etc.) with analog content (narrowly individuated); thus:

   Any occurrent mental state M, of mine, is conscious = M is disposed to cause an activated belief (possibly a non-conscious one) that I have M, and to cause it non-inferentially.

   In contrast with the actualist form of HOT theory, the HOTs which render M conscious are not necessarily actual, but potential. So the objection now disappears, that an unbelievable amount of cognitive space would have to be taken up with every conscious experience. There need not *actually* be *any* HOT occurring, in order for a given perceptual state to count as phenomenally conscious, on this account.

This theory is elaborated, and it is shown how phenomenal consciousness admits of a smooth and plausible evolutionary explanation on this account.

4. *Dispositional theory and categorical experience*

This section confronts an obvious challenge to dispositionalist HOT theory. When I am subject to a conscious experience, there is something actually taking place in me which constitutes my state of phenomenal consciousness. How, then, can the conscious status of my experience consist merely in the fact that I am *disposed* to have an appropriate HOT about it if circumstances should demand? For this is not something which is actually happening, but merely something which *would* happen if certain other things happened. The section argues that there are three dissociable strands in this complaint, in fact. Two are dealt with easily and briskly in the remainder of the section. The third is addressed at some length in chapter 9.

Chapter 9 Dispositionalist higher-order thought theory (2): feel

In this chapter we come to the crux. I examine how the competing higher-order accounts can explain the defining feature of phenomenal consciousness – namely its subjective feel, or 'what-it-is-likeness' – and I give a final adjudication between them on this ground. I argue that each of the two forms of higher-order thought (HOT) theory – in contrast with higher-order experience (HOE) theory – *can* advance essentially the same (fully successful) reductive explanation of phenomenal consciousness. In which case, given the strength of the arguments which we were able to deploy in chapter 8 on behalf of dispositionalist forms of HOT theory, it is the latter which should emerge as the overall winner.

1. *HOE theory and feel*

   This section argues briefly that the weaknesses of 'inner sense' theory are not compensated for by any advantage in explaining the subjectivity of phenomenally conscious experience. On the contrary, such theories face the problem of 'excess content' - if there really were the inner scanners postulated by such theories, one would expect that conscious experience should be factorable into two distinct aspects, one provided by our first-order senses, and one provided by our inner senses. But this expectation is not borne out.

2. *Actualist HOT theory and feel*

   This short section argues for the inadequacy of Rosenthal's (1998) explanation of the *feel* of experience in terms of actual higher-order thought. It is pointed out, however, that actualist HOT theory *could* embrace the very same consumer-semantic explanation to be offered in section 3, but that the result would be theoretically ill-motivated in comparison with its dispositionalist cousin.

3. *Consumer semantics and feel*

   This lengthy section constitutes the heart of the book. My claim is that the very same perceptual states which represent the world to us (or the conditions of our own bodies) can at the same time represent the fact that those aspects of the world (or of our bodies) are being perceived. It is the fact that the faculties of thinking to which experiences are made available *can make use of* those experiences in dual mode which turns them into dual-mode representations. This is because, in general, the intentional content of a state will depend upon the nature and powers of the 'consumer-systems', as Millikan (1984) would put it. The content possessed by a given state depends, in part, upon the uses which can made of that state by the systems which can consume it or draw inferences from it. And similarly, then, in the case of perceptual representations: it is the fact that perceptual contents are present to a system which is capable of discriminating between, and of making judgements about, those perceptual states *as such* which constitutes those states as second-order representations of experience, as well as first-order representations of the world (or of states of the body). And it is in virtue of this that those states acquire a subjective dimension, or *feel*.

Note that on this account phenomenal consciousness is constituted by higher-order analog representations, or higher-order experiences (HOEs), just as HOE theory – or 'inner sense' theory – maintains. But there don't actually need to be two physically distinct sets of representations to carry the two sets of perceptual contents, in the way that HOE theory supposes. Rather, dual content comes for free with the availability of perceptual contents to the mind-reading faculty, or with the availability of those contents to HOT. It is in virtue of the availability of first-order perceptual contents to a mind-reading system which understands the is/seems distinction and/or contains recognitional concepts of experience, that all of those first-order contents are, at the same time, higher-order ones.

This account is elaborated at some length, and its virtues expounded.

4. *Elucidations and replies*

This is another lengthy and important section. It further elaborates on dispositionalist HOT theory's explanation of phenomenal consciousness by way of responding to a series of potential objections and counter-examples.

Chapter 10 Phenomenal consciousness and language

In this chapter I argue that the simple form of dispositionalist HOT theory of phenomenal consciousness defended in chapters 8 and 9 is preferable to three other similar but more elaborate accounts (due to Carruthers, 1996; Dennett, 1978; and Dennett, 1991 respectively). Each of these is a form of dispositionalist HOT theory, but each makes out a constitutive connection of some sort between phenomenal consciousness and language.

1. *Reflexive thinking theory and language*

In this section I contrast – favourably – the account outlined in chapter 8 with a rather more elaborate form of HOT theory of phenomenal consciousness, defended in some earlier publications of mine (e.g. 1996), which I refer to as 'reflexive thinking theory'. On this account, experiences are conscious when they are available to acts of higher-order thinking *which are reflexively available to further higher-order thinkings*. This more demanding theory is shown to be ill-motivated for explanatory purposes. But I suggest that it may well be that reflexive thinking theory does, *de facto*, describe the (language-involving) structure of human consciousness, as I claimed in my 1996.

2. *Higher-order description (HOD) theory*

In this section I begin to explore a set of views – focusing mainly on Dennett – which are like dispositionalist HOT accounts, except that they define state-consciousness in general, and phenomenal consciousness in particular, in terms of the availability of a state to higher-order *linguistic description*. In the present section I examine and criticise Dennett's early view (1978), which can be seen as a simplified, language-based, form of reflexive thinking theory.

3. *The Joycean machine*

This section examines Dennett's (1991) theory of the conscious mind as a language-based *Joycean machine*. The theory is elaborated and its connections with evolutionary issues discussed.

4. *The independence of structured HOTs from language*

In order for higher-order linguistic description (HOD) theory to be preferred to higher-order thought (HOT) theory as an account of the phenomenal consciousness of human beings, it has to be the case that all hominid thought (realistically construed, as involving discrete, structured, content-bearing states) involves natural language. More particularly, it has to be the case that there can be no structured HOTs except those which are formulated in language. The present section defends dispositionalist HOT theory by arguing, on a variety of grounds, that the human capacity for HOTs is independent of language.

Chapter 11 Fragmentary consciousness and the Cartesian theatre

I have argued for the superiority of higher-order thought (HOT) theory over higher-order description (HOD) theory: there is no reason to think that a theory of phenomenal consciousness should implicate natural language, and there is good reason to think that it should not. This leaves untouched Dennett's (1991) arguments for a 'multiple drafts' approach to phenomenal consciousness, however, and in support of the radical indeterminacy of facts concerning the latter; together with his attacks on 'Cartesian theatre' models of phenomenal consciousness (of which both his own earlier 1978 theory and my sort of HOT theory are alleged examples). These arguments form the topic of this final chapter.

1. *Multiple drafts versus integrated contents*

   This section defends the intuitive richness of phenomenally conscious experience against Dennett's attacks (this is the issue held over from 8:2). It also argues positively that practical reasoning in relation to the perceived environment requires a set of integrated perceptual contents.

2. *Fragmenting the Cartesian theatre*

   This section begins defending dispositionalist HOT theory against the charge that it is committed to a 'Cartesian theatre' conception of the conscious mind. Various distinct strands in this charge are distinguished, and a number of them are shown to be innocuous. Those which are more challenging are held over to sections 3 and 4.

3. *Time as represented versus time of representing*

   This section demonstrates that dispositionalist HOT theory - like Dennett's own theory - can maintain that time is *represented* in the brain, rather than being given by *time of representing*. However, the account is genuinely committed to the idea that there is an objective (albeit vague) time at which an experience first becomes conscious, which may be distinct from the time at which that state is *experienced as* occurring.

4. *Objective versus subjective time*

Dennett develops an argument against the very coherence of the idea that there might be an objective, determinate, time at which an experience first becomes phenomenally conscious, as distinct from the subject's *representation of* the time at which the experience first occurs (1991; Dennett and Kinsbourne, 1992). This section expounds and elaborates on that argument, and unpicks the assumptions on which it is based. The upshot, then, is that Dennett has no damaging objections to the dispositionalist HOT theory being proposed in this book.

Conclusion

The book concludes with a brief survey of the main course of the argument, and of its major premises. It closes with a slogan: *a disposition to get higher makes consciousness phenomenal*.

References

Armstrong, D. 1968. *A Materialist Theory of the Mind.* Routledge.

Armstrong, D. 1984. Consciousness and causality. In D. Armstrong and N. Malcolm, *Consciousness and Causality*, Blackwell.

Block, N. 1990. Inverted Earth. *Philosophical Perspectives*, 4.

Botterill, G. and Carruthers, P. 1999. *The Philosophy of Psychology*. Cambridge University Press.

Carruthers, P. 1996. *Language, Thought and Consciousness.* Cambridge University Press.

Chalmers, D. 1996. *The Conscious Mind.* Oxford University Press.

Dennett, D. 1978. Toward a cognitive theory of consciousness. In C. Savage ed., *Minnesota Studies in the Philosophy of Science*, 9.

Dennett, D. 1991. *Consciousness Explained*. Allen Lane.

Dennett, D. and Kinsbourne, M. 1992. Time and the observer. *Behavioral and Brain Sciences,* 15.

Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly*, 32.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy*, 83.

Kripke, S. 1972. Naming and necessity. In G. Harman and D. Davidson, eds., *Semantics of Natural Language*, Reidel.

Lycan, W. 1987. *Consciousness*. MIT Press.

Lycan, W. 1996. *Consciousness and Experience.* MIT Press.

McGinn, C. 1991. *The Problem of Consciousness.* Blackwell.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories.* MIT Press.

Milner, D. and Goodale, M. 1995. *The Visual Brain in Action*. Oxford University Press.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review*, 83.

Nagel, T. 1986. *The View from Nowhere.* Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Rosenthal, D. 1986. Two concepts of consciousness. *Philosophical Studies*, 49.

Rosenthal, D. 1993. Thinking that one thinks. In Davies and Humphreys, eds., 1993.

Rosenthal, D. 1998. State consciousness and what it's like. Paper delivered to a cognitive neuroscience seminar, Corpus Christi, Oxford. Forthcoming in D. Rosenthal, *Consciousness and Mind*, Oxford University Press.

Shoemaker, S. 1981. The inverted spectrum. *Journal of Philosophy*, 74.

Tye, M. 1995. *Ten Problems of Consciousness*. MIT Press.

Evolving Phenomenal Consciousness

Commentary on Carruthers, Peter. *Phenomenal Consciousness*. Cambridge: Cambridge University Press, 2000.

Colin Allen

Department of Philosophy
Texas A&M University(USA)
colin-allen@tamu.edu homepage

As part of the argument for the superiority of his HOT (Higher Order Thought) theory of phenomenal consciousness, Carruthers offers an answer to the question of how phenomenal consciousness evolved, and he claims that this answer is superior to alternatives based upon other theories. I wish to demur. Evolutionary considerations are the cornerstone of Carruthers' rejection of alternative HOE (Higher Order Experience) accounts of phenomenal consciousness, which he argues do not successfully explain the cognitive role of phenomenal consciousness. They are also a cornerstone of his denial that nonhuman animals, young children, and autistic individuals are phenomenally conscious. The conjectured evolutionary story is too weak for either role.

In Carruthers' story, phenomenal consciousness emerges as an evolutionary "by-product, not directly selected for" (2000:230) of a two-stage selection process: First there was selection for first-order sensory representations (unconscious experiences), then there was selection for a "mind-reading" capacity which required conceptualization of mental states. On Carruthers' view, once the organism's own first-order sensory representations become directly available for conceptualization they are *de facto* phenomenally conscious. Once experiences become phenomenally conscious, then further adaptive benefits may follow -- particularly, Carruthers thinks, for making appearance-reality distinctions (2000:232; see also Allen & Bekoff 1997).

We have a huge range of phenomenally conscious experiences, from pains and orgasms, to the taste of sour milk and the feeling of breathlessness caused by the thin air and the staggering view from atop a snow-capped mountain. I shall argue that Carruthers' account fails to explain why we are phenomenally conscious in all the ways that we are. In other words, I shall put pressure on the alleged generality of this account, which is in evidence in this passage from chapter 8:

> Now the important point for our purposes is that the mind-reading faculty would have needed to have access to a full range of perceptual representations. It would have needed to have access to auditory input in order to play a role in generating interpretations of heard speech, and it would have needed to have access to visual input in order to represent and interpret people's movements and gestures, as well as to generate representations of the form, `A sees that P' or `A sees *that* [demonstrated object/event]'. Mere conceptual events wouldn't have been good enough. For often what needs to be interpreted is a fine-grained gesture or facial expression, for which we lack any specific concept. It seems reasonable to suppose, then, that our mind-reading faculty would have been set up as one of the down-stream systems drawing on the integrated first-order perceptual representations, which were already available to first-order concepts and indexical thought. .... Once this had occurred, then nothing *more* needed to happen for people to enjoy phenomenally conscious experiences, on a dispositionalist HOT account. (2000:231)

In this passage, the phrase "access to a full range of perceptual representations" seems to be playing a dual role. First, although only two forms of perception -- hearing and sight -- are mentioned explicitly they seem to be standing service for all the forms of perception which give rise to phenomenally conscious experience, and the full range of perceptual representations should therefore encompass odor, taste, and touch, as well as nociception and other somatic sensations. Second, Carruthers' reference to those elements of

perception for which we lack specific concepts indicates another sense in which the range of perceptual representations available for second order thought is supposed to be construed broadly.

I shall have little to say about this second sense of breadth, although I do think there are questions that might be raised about the limits of this account of phenomenal consciousness in the light of such phenomena as chicken-sexing (Biederman & Shiffrar 1987) where the question of how it is accomplished cannot be answered by introspection on the part of the expert chicken-sexers (Harnad 1996). Instead, I shall focus on the first notion of breadth -- the idea that Carruthers has outlined here an account of the evolution of phenomenal consciousness that applies across a wide variety of experiences from different perceptual systems.

Carruthers draws attention to interpretive acts based on speech and gesture, as well as to a more general class of attributive acts that seem to have less to do with communication *per se*. Because intentional communication between humans takes place predominantly in the modalities of hearing and vision (and perhaps to some extent using touch), Carruthers' focus on these two modalities in the quoted passage seems designed to enhance the plausibility of his thesis that interpretation constitutes a driving force for evolution. But the thesis seems considerably less plausible with respect to other sensory modalities, particularly smell and taste. The way others look to us, sound to us, and the sensations they produce when they touch us are all possible targets of interpretation. But there seems little to interpret in the way other people smell and taste to us. I conclude that the mind-reading faculty has no need for access to smell and taste for interpretive purposes.

What about more general forms of attribution? Well, it seems trivially easy to think of scenarios in which it would be adaptively useful to know what another individual is smelling or tasting. And we also know that natural selection can operate on very small margins, so it is not out of the bounds of possibility that there could have been selection for mind reading with respect to smell and taste. But this just-so story needs fleshing out, especially in light of the fact that it is not a foregone conclusion that these perceptual systems should give rise to phenomenally conscious experience, for there is at least one perceptual system, the vomeronasal system (Monti-Bloch et al. 1998), which responds to pheromones and affects human behavior but with respect to which we utterly lack phenomenal consciousness. Indeed it seems much more straightforward to think of cases where it would be adaptively advantageous (not to mention potentially pleasurable) to know whether one's pheromones have been detected and are generating an intense desire for intercourse in a conspecific, than it is to think of adaptive scenarios for more mundane odors. It is far from clear why we have phenomenally conscious smell and taste but are oblivious to "vomerolfaction" (Cooper & Burghardt 1990).

At best, then, the evolutionary story for phenomenal consciousness with respect to taste and smell is not proven. Neither sense is important to interpretation, and a weak just-so story is all that has been suggested to explain why we should have evolved the ability to attribute gustatory and olfactory states to others. It might be thought that this case can be made stronger by pointing to facts such as that my knowledge that some food type tastes bad to you, might make me less inclined to eat it myself. But I see no reason for thinking that the ensuing adaptive advantage that accrues from lowering the chances of poisoning oneself need to be mediated by conceptual recognition of the gustatory experiences of others, when the very same advantages could be derived simply by learning some non-mental facts about their reduced tendency to ingest this type of food. Of course there could be *exaptation* for attribution of such perceptual states (although scenarios involving deceptive manipulation of appearances are harder to imagine for taste and smell than for sound and vision), but Carruthers is supposed to be providing us here with an explanation for initial selection of the capacity to attribute gustatory states to others, not its subsequent cooption.

It is not my intention, however, to trade just-so stories. Rather, the point is this: the fact that the vomeronasal system is devoid of phenomenology shows that there is no guaranteed connection between phenomenal consciousness and any given behavior-guiding perceptual system, even among mind-reading creatures such as ourselves. Thus we are entitled to demand from Carruthers a particular and specific explanation for phenomenal consciousness with respect to each of the separate perceptual systems. It is not good enough to say that "mind-reading faculty would have needed to have access to a full range

of perceptual representations" for the mind-reading faculty does not in fact even have access to the full range of perceptual systems.

Carruthers' diagrams tend to obscure the separate processing of different perceptual streams. In his figure 8.1 (fig 1a), for instance, the "Percept" box splits its output between two "short-term memory stores, C (conscious) and N (non-conscious)" (2000:228). Clearly the vomeronasal system has no direct line to C, so cannot be regarded as belonging in a general "Percept" box. As far as I know there is no neurological evidence for a single "Percept" box through which all perceptions pass. Hence the diagram should, minimally, be amended to show separate input streams from each perceptual system (even this is likely to be an oversimplification given the complex interactions between different perceptual modalities and conscious experience). There should, then, be separate links from each perceptual system to C, or perhaps even links to separate short-term memory stores, each with separate links to the "Conceptual thinking/HOT" box (fig 1b), and we are owed an explanation for the links between each perceptual system and the parts of the system responsible for consciousness.

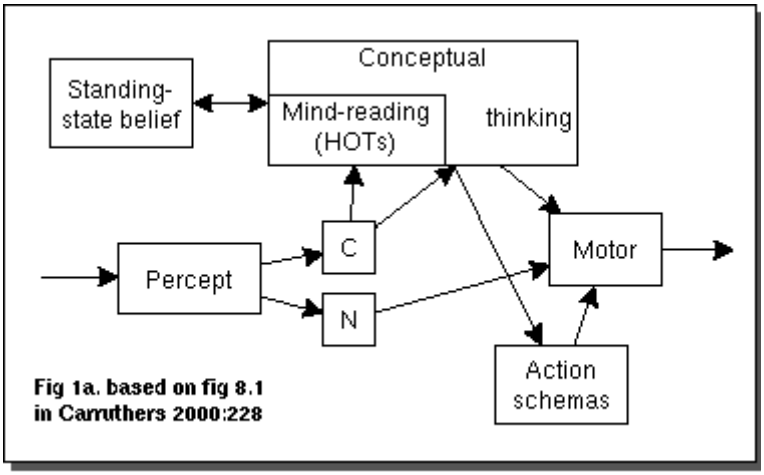

Fig 1a. based on fig 8.1 in Carruthers 2000:228

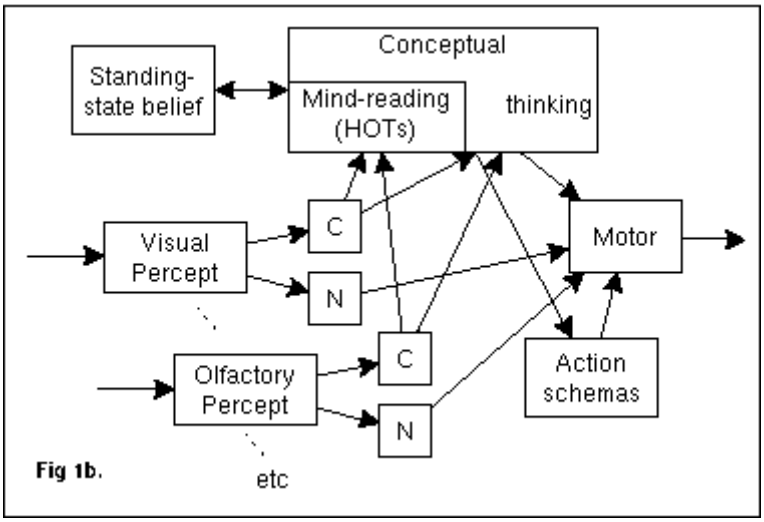

Fig 1b.

Further complexity in the relationship between perception and conscious experience consists in the fact that some modalities interact in ways that affect phenomenal consciousness. For instance conscious experiences of *flavor* (distinct from taste) are due to a synthesis of gustatory and olfactory inputs, and effects such as the McGurk illusion (McGurk & McDonald 1976) show that conscious experience of speech can be affected by the visual perception of lip movement -- the simple act of closing one's eyes can change which phoneme is heard given identical aural input. Other modalities do not interact in this way; thus, for instance, the phenomenology of vision is not, as far as we know, influenced by olfaction (although flavor perception may be affected by vision). Carruthers might argue that these kinds of phenomena are determined by perceptual systems before the perceptual contents become available to higher order thought and that therefore it is not incumbent upon his theory of consciousness to explain them. But given that olfactory and gustatory neurons project to separate brain structures, and that odor and taste can be distinguished in associative conditioning experiments, there is no clear reason (particularly with respect to mind-reading) why phenomenal consciousness should receive synthetic experiences of flavor rather than concurrent experiences of taste and odor.

I have argued that the evolutionary explanation of phenomenal consciousness delivered by Carruthers is rather weak. This conclusion is yet compatible with his claim that the HOT theory provides a better evolutionary explanation for phenomenal consciousness than other accounts of consciousness, for the truth of this claim depends on the relative strength of those alternative accounts.

In what remains of this commentary, I want to argue that Carruthers gives too short shrift to at least one of the alternatives he considers and rejects on behalf of HOE accounts, specifically that the capacity for conscious discrimination between different experiences might aid learning (2000:215, first item). Carruthers claims that, for example, learning to avoid harmful events is possible without higher order representation of pain states -- all that is required is the capacity to distinguish painful stimuli (i.e. stimuli that trigger nociception) from other stimuli such as tickles. It is "hard to see" the point, he writes, of "discriminating between *experiences* of pain and *experiences* of tickling ... in the absence of a capacity for HOT" (2000:215). He provides no empirical citations to support this claim. But it is in fact borne out to a certain extent by work on spinal nociceptive mechanisms in rats (see Grau 2001). Rats whose spinal cords have been cut in the cervical region show quite sophisticated kinds of associative learning in response to noxious stimuli applied to the hind legs, even though we can be sure that no signals are reaching the brain, and correspondingly certain that there is no phenomenal consciousness associated with these stimuli. The spinal cord distinguishes painful stimuli from other stimuli, and adaptive changes in behavior result. It is true, therefore, that learning can occur in the absence of conscious awareness of pain.

Carruthers is correct to reject simplistic adaptive hypotheses about the function of phenomenally conscious pain. Phenomenally conscious pain is not required either for withdrawal from noxious stimuli, nor for associative conditioning of pain responses. Yet learning is not a unitary phenomenon; certain forms of operant learning seem to require the brain (Grau 2001), as do some forms of classical conditioning (see next paragraph). For pain-related learning, it is by no means clear that these more advanced forms of learning can be sustained on the basis only of "first-order information-bearing states differentially caused by tissue damage in the one case, and stroking or tickling in the other" (2000:215) for while such states may indeed be sufficient to support discriminative avoidance of the noxious stimuli in some circumstances, the varieties of organismic adaptation are far more subtle than Carruthers lets on. One simplistic hypothesis has been replaced with another. Indeed, one of the chief puzzles of conscious pain is its relative independence from tissue damage; whatever role the internal state is playing is much more complicated than mere correlation with tissue damage (see Hardcastle 1997 for an overview). It is also the case that pain has an evaluative component -- we can decide, often in retrospect, whether a certain level of conscious pain was worth the reward or reasonably correlated with the danger, and such evaluations affect subsequent behavior (i.e., consciously-mediated learning results). Here, the point may not be whether there is a capacity to distinguish between experiences of pain and experiences of tickling, as Carruthers puts it, but whether there is the capacity to discriminate among experiences of different intensities. Since these different intensities are variably related to tissue damage, it might very well be adaptive to discriminate between the painful experiences as such so as to allow independent assessment of the severity of the underlying conditions that they purport to represent.

Another intriguing finding relating brain (specifically hippocampal) function to a sophisticated form of learning derives from Clark & Squire's (1998) experiments on "trace conditioning", a form of classical conditioning involving a temporal separation between the conditioned stimulus (CS) and the unconditioned stimulus (US). Clark & Squire used the eye-blink reflex as the response and showed that when presented with an audible tone (CS) that terminated one half or one second before a puff of air (the US), not all normal human subjects learned to blink in response to the tone (prior to the puff), but all those who did learn reported during debriefing that they were explicitly aware of the temporal relation between the two stimuli. In contrast, all subjects can be conditioned to blink in response to the tone when the puff is presented after the onset of the tone but before its termination -- "delay conditioning" -- regardless of their knowledge of the temporal relationship between CS and US. Trace conditioning occurs in rabbits (Clark & Squire 1998), and there are much greater individual differences in success rate for trace conditioning than for delay conditioning in these animals (Thompson et al. 1996). While Carruthers might assert that trace conditioning requires only first order representation of the temporal relation between two events, not between the experiences themselves, we know that the human capacity to represent this event depends on phenomenally conscious awareness of the stimuli; one

cannot be explicitly aware of the temporal relation between two stimuli of which one has no phenomenal consciousness. In other words, what is represented seems to be the relationship between the two experiences.

Can such considerations ultimately sustain an HOE account? Ultimately I'm not sure, and I don't have any particular stake in the outcome. But even if they cannot -- even if higher order *thoughts* are required to mediate the more sophisticated kinds of learning -- I believe that the considerations above undermine one of the more radical claims that Carruthers takes to be a consequence of his HOT account: namely his denial of phenomenal consciousness to nonhuman animals, young children, and autistic individuals, on the basis of their apparent lack of full mind-reading capabilities. Even if HOTs are required for certain kinds of learning, such learning seems within the range of at least some nonhuman animals whether or not they can "read minds".

Carruthers' denial of phenomenal consciousness to animals runs strongly counter to most people's intuitions on this matter. "It really is something of a scandal," he writes, "that people's intuitions, in this domain, are given any weight at all" (2000:199). At this stage of our understanding of behavior and neuropsychology, a far greater scandal, I believe, would be to circumscribe phenomenal consciousness on the basis of one man's inability to imagine alternative accounts for its evolution.

Bibliography

Allen, C. and Bekoff, M. 1997. *Species of Mind*. Cambridge, MA: The MIT Press.

Allen, C. 2000. "Where pigs and people fear to tread." In *The Smile of a Dolphin*, ed. M. Bekoff, pp. 90-91. New York: Discovery Books.

Biederman, I. & Shiffrar, M. M. 1987. "Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task." *Journal of Experimental Psychology: Learning, Memory, & Cognition* 13: 640 - 645.

Clark, R. E., & Squire, L. R. 1998. "Classical conditioning and brain systems: The role of awareness." *Science* 280: 77-81.

Cooper, W. E. Jr. & Burghardt, G. M. 1990. "Vomerolfaction and vomodor." *Journal of Chemical Ecology* 16: 103-105.

Grau, J. 2001. "Learning and Memory without a Brain." In *The Cognitive Animal*, ed. M. Bekoff, C. Allen, G.M. Burghardt, in press. Cambridge, MA: The MIT Press.

Hardcastle, V.G. 1997. "When Pain is Not." *Journal of Philosophy* XCIV: 381-409.

Harnad, S. 1996. "Experimental Analysis of Naming Behavior Cannot Explain Naming Capacity." *Journal of the Experimental Analysis of Behavior* 65: 262-264.

McGurk, H. & McDonald, J. 1976. "Hearing lips and seeing voices." *Nature* 264: 746-748.

Monti-Bloch, L., Jennings-White, C., & Berliner, D. L. 1998. "The human vomeronasal organ: A Review." *Annals of the New York Academy of Sciences* 855:373-389.

Thompson, L.T., Moyer, J.R., and Disterhoft, J.F. 1996. "Trace Eyeblink Conditioning in Rabbits Demonstrates Heterogeneity of Learning Ability Both Between and Within Age Groups." *Neurobiology of Aging* 17:619-629.

Acknowledgments

I am grateful to Gordon Burghardt, Lori Gruen, and Gary Varner for comments.

Reply to Colin Allen

Peter Carruthers

Department of Philosophy
University of Sheffield, Sheffield, UK.
Homepage

Allen focuses on some of the evolutionary / cognitive-engineering arguments which I deployed in my 2000, chapters 8 and 11. These were supposed to support my sort of dispositionalist higher-order thought (HOT) theory against theories of 'inner sense', which maintain that we have a set of inner scanners charged with producing higher-order experiences (HOEs) of our first-order perceptual, imagistic, and emotional states. (Note that inner-sense theory might more accurately be called 'second-order sense theory'. For pain is physically 'inner', of course, but the outputs of pain-perception will need to be scanned too, according to inner-sense theory - producing HOEs of its contents - in order to become phenomenally conscious.)

Allen makes two allegations against me. First, that the evolutionary explanation provided for dispositionalist HOT theory is itself much weaker than I think it is. (So even if there isn't a plausible explanation of the evolution of inner sense, as I claim, still the two theories are near enough level-pegging.) And second, that there are possibilities which I have overlooked for the early evolution of a faculty of inner sense in a variety of creatures besides ourselves. (In which case our common-sense intuition that such creatures are phenomenally conscious may be salvageable.) I shall take these allegations in turn. On closer examination, neither is very powerful.

1 Evolution, HOEs and HOTs

The alleged weakness in the account provided from the perspective of dispositionalist HOT theory, of how phenomenal consciousness may have evolved, is that it cannot explain the full range of phenomenally conscious states which we enjoy. Since the evolutionary account of our capacity for HOT which lies at the heart of my approach has to do with the benefits of psychological interpretation or 'mind-reading', Allen concedes that the availability of visual and auditory contents to HOT is plausible. But he argues that there is little plausibility in the view that taste, smell, and pain-contents would play a significant role in other-interpretation. So we have no evolutionary account of the availability of such contents to HOT, and so no evolutionary account of their phenomenal consciousness, on a dispositionalist HOT approach.

Allen misconstrues the import of the passage he quotes from page 231 of my 2000, however. That passage wasn't - or wasn't primarily - about explaining why the mind-reading faculty needs access to different kinds of perceptual content. (I confess that the phrase 'full range of perceptual representations' may have been misleading.) Rather, it was about explaining why the mind-reading faculty needs access to perceptual contents, rather than running off mere beliefs as inputs. It was taken for granted (having been argued briefly two paragraphs previously, and in much more detail in chapter 11:1.3) that the perceptual outputs from different sensory modalities were already available (prior to the evolution of a capacity for HOT) to conceptualizing and practical reasoning (or 'executive function') systems. My task was just to explain why the evolving mind-reading system would have been set up with access to this set of outputs, rather than as a mere inference system operating on beliefs.

What Allen sees as questions for the evolution of a dispositionalist HOT architecture - namely, a demand to explain, in connection with each distinct sense-modality, why the mind-reading system would need to have evolved access to the outputs of that modality - I see rather as questions for the prior evolution of the first-order conceptualizing and planning systems which we share with many other species of animal. And I think that, in outline, the answer would go something like this. Information from many different sense-

modalities can be important factors in object-recognition. Obviously, visual information will often be relevant. But so will sound, touch, smell and taste - what something sounds like, feels like, smells like and tastes like can be crucial factors in identifying it. On cognitive-engineering grounds, then, we should expect that information of each of these sorts should be simultaneously present to the mechanisms charged with conceptualization of the world. Admittedly, this doesn't yet explain why temperature-information and pain-information should give rise to phenomenal consciousness in ourselves, since these factors rarely if ever play a part in object-recognition. But they do play an important part in planning. In deciding whether to continue holding a hot object, say, or whether to continue pushing one's hand into a bees' hive to reach the honey, experiences of heat and pain will be important considerations, with obvious adaptive advantages.

The two-stage account of the evolution of phenomenal consciousness presented in my 2000, then, was this. First, there was the evolution of first-order conceptualizing and planning systems, feeding off the outputs of all those sensory systems which are relevant to the execution of these functions. I envisage this as resulting in a single functionally-defined short-term memory store (the 'E' - for 'experience' - box in Figure 11.1). This would contain analog outputs from all of these perceptual systems, held in such a way that the processes of conceptualization and planning could feed off any aspect of the contents of that store, depending on relevance to current goals, context, background beliefs and so on. This is an architecture which we probably share with all other mammals, at least. Then second, somewhere in the ape and/or hominid lineage, there was selection for a number of conceptual modules or quasi-modules - for folk-physics, folk-biology, and (crucially for my purposes) folk-psychology or mind-reading - each of which was so set up that it could draw on the contents of the already-existing E-store. And it is the resulting availability of these first-order analog contents to the HOTs generated by the mind-reading system which renders them higher-order, and hence phenomenally conscious, according to my account.

Admittedly, further questions can now be raised in the spirit of Allen's first allegation. It can be asked why there should have been just one first-order E-box, drawn on for purposes of both conceptualization and planning; and why the evolving mind-reading system should have been set up to feed off the entire set of contents of the E-box, rather than just the sub-set which might be directly relevant for purposes of other-interpretation.

One answer to the first of these questions is that a single perceptual short-term memory system is simpler - and less costly to build and maintain - than two. Another is that conceptualization and planning are intimately linked. Indeed, conceptualization is largely for planning. The point of conceptualizing the world into kinds is to facilitate inductive generalization and learning of various sorts which can generate information to guide good plans. And while the planning process itself will characteristically be conducted, at least partly, in conceptual terms, it is also often highly indexical - as can be the outputs of the conceptualization process. Conceptualization of an animal in the bushes may lead to the indexical judgment, 'That is a tiger'. From previous learning it is known that tigers are dangerous, leading to the indexical intention, 'I need to get away from that'. Here it will be important that the conceptualizing and practical reasoning systems should both have access to the same set of perceptual contents, so that the referent of 'that' can be held constant through the whole process.

But now, why would the evolving mind-reading system have been set up to feed off the entire contents of the E-box, rather than just a sub-set? One answer is that evolution characteristically builds on and co-opts whatever structures are already in place. Since an integrated first-order perceptual memory system was already in existence, feeding the processes of first-order conceptualization and planning, it is only to be expected that an evolving mind-reading system, needing access to perceptual contents, would have been set up to draw on this perceptual memory system. A further answer is that mind-reading and planning (or 'executive function') are closely linked, too. Planning and reasoning in humans is routinely meta-representational in character (Carruthers, 1996; Perner, 1998). Indeed, it can be debated whether autism, for example, is fundamentally a pathology of mind-reading or of executive function (Russell, 1996). It is only to be expected, then, that the mind-reading system should draw on the same range of perceptual contents which are needed for practical reasoning - and these will include heat and pain.

From this perspective, then, Allen's further questions - concerning synthetic experiences of flavor (as opposed to distinct experiences of taste and smell), and why detection of pheromones should remain unconscious - are not (as he construes them) questions for the

second (mind-reading) stage of the above evolutionary account. The task is not that of explaining why the demands of mind-reading should lead to integrated experiences of flavor, on the one hand, and why it shouldn't lead to experiences of pheromones (as opposed to their emotional effects), on the other. For I assume that these matters had already been determined prior to and/or independently of the evolution of the mind-reading faculty. Rather, the task is to explain why the demands of object recognition and/or planning should lead to integrated flavor-contents, and why those same demands shouldn't require the availability of pheromone-information. These are good questions, to which I might begin to construct speculative answers. But they are not questions which bear especially on theories of phenomenal consciousness. For they are faced equally by first-order theories of the sort espoused by Tye (1995) and Dretske (1995); by inner-sense theories of the sort espoused by Armstrong (1968) and Lycan (1996), and defended here by Allen; by actualist HOT theories of the kind espoused by Rosenthal (1986); as well as by my own dispositionalist HOT theory. They are questions for all of us. The lack of easy answers to these questions does not show dispositionalist HOT theory to be weak (on this score at least) in relation to these other theories.

2 HOEs and learning

Allen's first allegation is therefore based largely on a misunderstanding (or at least an unsympathetic reading) of my views. His second allegation is that I don't take seriously enough the idea that higher-order experiences (HOEs) might be needed to underpin the kinds of sophisticated learning of which many other creatures besides ourselves seem capable. So there may be a plausible evolutionary story to be told by an inner-sense theorist, which would at the same time warrant our common-sense intuition that these same creatures undergo experiences which are phenomenally conscious.

Allen makes the point that learning in humans is often mediated by pain-judgments (for example, as to whether suffering a certain level of pain was worth the benefits), and/or by discriminations of different pain intensities. Then if animals, too, can benefit from such learning, we would have good reason to think that they are phenomenally conscious. By implication, Allen must think that such judgments and discriminations are higher-order ones, since he is here supposed to be defending a form of HOE theory.

What Allen misses is that my 2000 endorses a conception of pains as analogous to secondary qualities in other sense-modalities - a conception which I take over from Tye (1995), and which I regard as essential to the plausibility of dispositionalist HOT theory. On this account, pains themselves are the first-order properties which form the intentional content of pain perceptions, just as red is the first-order property which forms the intentional content of one form of color vision. And notice that different shades and intensities of red can be perceptually represented without the need for higher-order experiences of any sort, and that behaviors grounded in color-judgments can similarly occur without the need for any form of higher-order representation. In the same way then, I maintain, discriminations of pain intensity can be entirely first-order; and judgments of pain-intensity and pain-evaluation are similarly first-order. There is no reason at all why creatures incapable of any higher-order representation shouldn't be capable of discriminating pain intensities (any more than they should be incapable of discriminating color intensities), or why they shouldn't be capable of thoughts like, 'That pain was worth that reward'. On my view, there is nothing higher-order in such thoughts, any more than there is anything higher-order in thoughts like, 'That red is deeper than that one'.

Of course it is true that in the human case such discriminations and such judgments will be accompanied by phenomenal consciousness. And it is also true, as Allen points out, that the human capacity for explicit judgments of the temporal ordering of events will be accompanied by phenomenal consciousness. But it is quite another matter to claim that these judgments are made possible by phenomenal consciousness. And on the contrary, these coincidences are readily explained on the dispositionalist HOT approach. For the perceptual contents which ground sophisticated learning and which ground explicit judgments (the contents of the E-box) are also, in the human case, available to HOT, hence rendering them phenomenally conscious (and hence transforming the 'E' box into a 'C' - for 'conscious' - box).

There are views of phenomenal consciousness according to which these discriminations (whether of color or of pain) will be phenomenally conscious ones, of course. (For example,

any first-order representationalist view - in the manner of Tye and Dretske - will imply as much.) What Allen loses sight of is that his task is to be defending HOE theory at this point. For evolutionary considerations are only supposed (by me) to help discriminate amongst the competing higher-order theories (inner-sense theory, actualist HOT theory, dispositionalist HOT theory), not between dispositionalist HOT theory and accounts of other types. On the contrary, one of the obvious strengths of first-order representationalism is the plausible evolutionary story that it can tell. (Its weakness is its inability to handle the distinction between conscious and non-conscious experience. See my 2000, chapter 6.) So what Allen really needs, in particular, are some examples of learning which require genuinely higher-order experiences in order to succeed, not just examples of learning which are accompanied by phenomenal consciousness in human beings. He doesn't come close to providing them.

The challenge to inner-sense theorists laid down in my 2000 remains, then - to provide some account of what HOEs might be doing for us in cognition, which doesn't presuppose a capacity for HOT, and which is evolutionarily important enough to explain the investment needed to build and maintain the required set of second-order sense-organs. No one whom I have debated and discussed these issues with over the years has been able to come up with a half-way plausible proposal. And neither does Allen. This doesn't prove, of course, that some function for HOEs mightn't yet be thought of; nor does it show that we mightn't yet find independent evidence for the existence of a set of second-order sense-organs. But it does provide good reason (provided that one is in the market for a higher-order account of phenomenal consciousness at all) to prefer dispositionalist HOT theory to inner-sense theory. For this gets us all of the benefits of HOEs for free, without the need to appeal to any set of inner scanners. (On this, see my 2000, chapter 9.)

References

Armstrong, D. 1968. *A Materialist Theory of the Mind*. Routledge.

Carruthers, P. 1996. Autism as mind-blindness: an elaboration and partial defence. In P.Carruthers and P.K.Smith (eds.), *Theories of Theories of Mind*. Cambridge University Press.

Carruthers, P. 2000. *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.

Lycan, W. 1996. *Consciousness and Experience*. MIT Press.

Perner, J. 1998. The meta-intentional nature of executive functions and theory of mind. In P.Carruthers and J.Boucher (eds.), *Language and Thought*, Cambridge University Press.

Rosenthal, D. 1986. Two concepts of consciousness. *Philosophical Studies*, 49.

Russell, J. 1996. Agency: its role in mental development. Lawrence Erlbaum.

Tye, M. 1995. *Ten Problems of Consciousness*. MIT Press.

Commentary on

Carruthers, Peter. *Phenomenal Consciousness*. Cambridge: Cambridge University Press, 2000.

José Luis Bermúdez

Department of Philosophy
University of Stirling (UK)
homepage

Peter Carruthers's new book is a welcome addition to the extensive literature on phenomenal consciousness. Developing and expanding the principal themes of his 1996 book *Language, Thought and Consciousness*, Carruthers offers both an incisive discussion of the principal existing theories of consciousness (including the theories which deny the possibility of explaining consciousness) and a significantly new version of the higher-order thought approach to conscious experience. Unusual among contemporary contributors to the consciousness debate, Carruthers pays far more than lip-service to the interdisciplinary nature of the topic and material from neurophysiology, developmental psychology and the scientific study of consciousness plays a pivotal role in motivating some of Carruthers's principal conclusions.

The argumentative structure of the book is given by what Carruthers terms the "tree of consciousness" (p. 22). The tree of consciousness is a series of seven choice nodes which effectively delineate the principal theories of consciousness. Carruthers tackles each node and motivates his own response to it until at the bottom of the decision-tree we are left with a single branch – Carruthers's own non-linguistic dispositionalist higher-order thought theory. Of the seven choice nodes, four are particularly significant. At the first node we have to choose between "no explanation" views (Chalmers, McGinn, Jackson et al) and the possibility of a genuine naturalistic explanation – Carruthers, of course, opts for the latter. Like almost all contributors to the debate, Carruthers thinks that the only hope for a naturalistic theory of phenomenal consciousness is via the representational contents of phenomenally conscious states. Within the general framework of representationalism the key decision (choice node 4) is between first-order and higher-order theories. First-order theories (such as those maintained by Dretske, Tye and others) maintain that phenomenally conscious states are representational states which impact, or are poised to impact, on belief formation and practical reasoning. According to higher-order theories, however, there is a further condition to be met. Phenomenally conscious states must also be the targets of higher-order representational states. Carruthers opts for the higher-order approach. Within the general framework of the higher-order approach, choice node 5 asks us to decide whether the higher-order representational states are thoughts or experiences. Carruthers takes the latter option and (in choice node 7) argues that the relevant higher-order thoughts (HOTs) do not have to be formulated in a natural language.

In many ways the most considerable hurdle for naturalistic theories of phenomenal consciousness comes with the very first choice node – with defending the very possibility of such a theory against the various "mysterian" arguments that have been put forward by Kripke, Jackson, McGinn, Chalmers and others. By seeing how a naturalistic theory confronts these arguments we can appreciate the type of explanation which it takes itself to be offering. In responding to Kripke, Carruthers makes plain the modal strength he thinks appropriate for a naturalistic theory of consciousness. It would be wrong to demand a necessary identity between types of phenomenally conscious states and types of natural properties – for any given phenomenally conscious state type might in a different possible world be differently realised. Nonetheless, a naturalistic theory must entail a logical supervenience claim, so that it cannot even be logically possible for the relevant natural properties to exist in the absence of the appropriate phenomenally conscious state. This means that Carruthers has to defuse the various arguments purporting to show that no such logical supervenience claim could possibly be true.

Carruthers's strategy here is interesting. He suggests that many of these arguments rest upon a mistaken conception of properties. For Chalmers, for example, a property is simply

a function from possible worlds to extensions, such that every coherent concept determines one such mapping function. This is what allows Chalmers to argue that, since the existence of zombie worlds is logically possible, the property *having an experience of red* cannot logically supervene on any set of natural properties. If we can show that our concepts are such that, for any candidate natural property N which we can conceptualise, it is conceivable that something should be N without having an experience of red, then it seems plausible that the concept *having an experience of red* cannot determine property N. But, according to Carruthers, if we take properties to be thickly individuated natural entities, then this conclusion does not follow. It might well be the case that our concept *having an experience of red* picks out a natural property in this world such that, in every world, every individual of whom that property is true will also instantiate the concept *having an experience of red* – even though there may be worlds in which the concept *having an experience of red* picks out a different natural property. In such a situation (which, Carruthers suggests, Chalmers et al have not ruled out) there would be logical supervenience without property identity.

Carruthers's argument here rests on denying the basic intuition underlying many of the arguments in this area, namely, that the concept *having an experience of red* picks out a phenomenological feel. He suggests, in contrast, that the concept picks out a conceptually individuated natural property – whatever property it is that makes it the case that we have an experience of red. It is hard to see, however, why Chalmers et al. should be convinced by this, since what they are interested in is the phenomenological feel itself, rather than the natural property which might be associated in this world with the experience of red. It is surely open to them to define a new concept, *having an experience* of red* that is stipulated to pick out the phenomenological feel of the experience of red, in complete independence of any natural property whatsoever. The usual thought experiments will then show that the property which this concept picks out does not supervene logically on any natural property.

Turning now to Carrruthers's own account of consciousness, his central argument against first-order representationalist theories is that such theories cannot deal with instances of non-conscious experience. That there are non-conscious experiences is, he thinks, shown not simply by exotic neuropsychological phenomena such as *blindsight* and *visual form agnosia*, but also by common-or-garden phenomena such as absent-minded perception and non-conscious accommodation to one's environment as well as by some of the striking dissociations between perception and action in normal subjects which have been taken as evidence for the two visual systems hypothesis. If such non-conscious phenomena have roughly the same functional role as conscious experiences then it follows that simply being poised to impact on belief formation and practical reasoning cannot be sufficient for phenomenal consciousness.

As one would expect, of course, the evidence is far from unequivocal. In all the cases he discusses there is clearly some form of perceptual registration (to use a deliberately neutral term). The question is whether, in each case, the perceptual registration is (a) correctly described as an experience at the personal-level (b) genuinely non-conscious and (c) more or less functionally equivalent to a conscious perception. It is far from clear that all three of these criteria are met by each of the examples which Carruthers adduces. It will be objected to many of the common-or-garden phenomena that they fail to satisfy both (a) and (b). It is hard to see (*pace* Carruthers's section 6.1.1) how we might determine whether the absent-minded lorry-driver is non-consciously perceiving the road or simply has an unimpressive short-term memory – here it is tempting to deny (b). As far as subliminal learning is concerned there is a certain plausibility in denying (a). Many of the neuropsychological phenomena, on the other hand, seem to fall foul of (c). The crucial point here is that (c) requires that the non-conscious phenomena do two things – feed into action *and* feed into belief formation – and it is arguable that the phenomena to which Carruthers draws attention at best do one but not the other. It may well be the case that the perceptual registrations of blindsight and visually agnosic patients do feed into action in more or less the way that conscious perceptions do (although many researchers would deny this), but they certainly do not feed into the processes of belief formation in the right sort of way – which is why such patients describe themselves as merely guessing. Something similar holds for the visual illusions cited to support the two visual systems hypothesis. In the Titchener illusion, for example, normal subjects report that two circles which are in fact equally sized appear differently sized – even though when asked to reach towards the circles the map between finger apperture and target showed that the sizes were being correctly estimated. Again, however, (c) does not appear to be met. It is presumably the

non-conscious perceptual registration of the correct size which is supposed to qualify as a counter-example to first-order representationalism – but this perceptual registration does not feed into the processes of belief formation in the right sort of way, which of course is why the Titchener illusion is an illusion.

It looks, therefore, as if the case against first-order representationalism is far from watertight. And one would be forgiven for thinking that this is just as well, since the higher-order version of representationalism which Carruthers offers in its place has some fairly counter-intuitive consequences. If, as higher-order representationalism maintains, higher-order thoughts are essential for phenomenal consciousness, then it follows that creatures incapable of higher-order thought will not be phenomenally consciousness – and, since Carruthers thinks that higher-order thought is the unique preserve of humans and perhaps some of the great apes, he ends up denying sentience to the vast majority of the animal kingdom. In fact, he sees this as an advantage of his theory, since he can find no reason for thinking that non-primates are phenomenally conscious except crude intuitions. We only think that non-primates are conscious because, realising that they have experiences, we attempt to conceptualise those experiences, which are in fact non-conscious, on the model of our own conscious experiences.

Carruthers is no doubt correct that intuitions are not to be trusted in this area. But he neglects the most significant reason for attributing sentience to non-primates. The most plausible model we possess for explaining the vast majority of animal behaviour is that provided by conditioning theory. The basic principle of conditioning theory is that certain patterns of behaviour are reinforced by being associated with primary positive reinforcers, and inhibited by being associated with primary negative reinforcers. But learning through conditioning works because primary reinforcers have qualitative aspects. It is impossible to divorce pain's status as a negative reinforcer from its feeling the way it does. It is impossible to divorce soothing vocalizations being positive reinforcers from their sounding the way they do. The success of stimulus-reinforcement models of learning therefore provides a powerful motivation for doubting higher-order thought theories of consciousness.

In conclusion, Peter Carruthers has written a rich and rewarding book which both imposes a rigorous framework within which we can evaluate existing theories of consciousness and significantly advances the debate. The level of argumentation is consistently high and a wide range of empirical evidence is brought to bear. *Phenomenal Consciousness: A Naturalistic Theory* repays careful study and no one working in the philosophy of mind and/or psychology can afford to ignore it.

Reply to José Luis Bermúdez

Peter Carruthers

University of Maryland, College Park

My thanks to José Bermúdez for his kind comments on my book (Carruthers, 2000). I shall focus here on his three main criticisms.

1 Mysterianism and the nature of properties

Bermúdez notes that one of my main points against those in general who are 'mysterian' about phenomenal consciousness, and against Chalmers (1996) in particular, turns on a distinction between two ways of thinking of properties. I claim that Chalmers relies on a 'thin' notion of property (roughly: a function from worlds to extensions), where properties are individuated in terms of the concepts which we use to express them. Would-be naturalizers, in contrast, can legitimately (and characteristically do) work with a 'thick' conception of properties-as-worldly-entities, such that one and the same property might be picked out by a number of distinct concepts. Given such a conception, we can claim that the properties which we pick out using phenomenal concepts can also be characterized in terms of physical or functional concepts. And we can, moreover, claim that the teeth can be drawn from the familiar zombie conceivability experiments. For from the fact that we can *conceive of* phenomenal concepts failing to apply to a creature despite all physical and functional facts remaining the same, it doesn't follow that those concepts don't in fact pick out (some subset of) those very facts.

Now Bermúdez claims that this move involves denying that the concept *having an experience of red* picks out a phenomenal feel. Rather, I am said to hold that the concept in question is really just the concept of *whatever property it is which makes it the case that we have an experience of red*. But this is just wrong; and what Bermúdez says here misses the point.

I repeatedly allow – indeed, insist – that we have available to us purely recognitional concepts for our phenomenally conscious experiences. Such concepts pick out a phenomenal feel, and (so far as their content goes), they do *nothing but* pick out a phenomenal feel. My point is just that, given a thick construal of 'property', it is perfectly consistent to claim that such concepts pick out a property which can *also* be characterized in physical or functional terms. And while the zombie thought-experiments can show that the *thin* properties *such-and-such a feel* and *such-and-such a physical state* are distinct (this is obvious: in general where concepts are distinct, the thinly-individuated properties picked out by those concepts are distinct also), those thought-experiments do nothing to show that these are distinct properties thickly understood.

2 The case against first-order representationalism

Bermúdez gives a sketch of the initial case I set out against first-order representationalist accounts of phenomenal consciousness (of the sort defended by Dretske, 1995, and Tye, 1995, 2000). Roughly, the case is that there exist forms of first-order perceptual content which *aren't* phenomenally conscious; and so it must be something else (something higher-order, I claim) which explains why *some* perceptual states *are* phenomenally conscious. Bermúdez rightly points out that in many of the examples of non-conscious perception which I discuss, it isn't obvious that we have states which are fully first-order-equivalent to conscious perception. More specifically, in most such cases it is doubtful whether we have perceptual contents which are available *to conceptual thought and reasoning* (as opposed to being available to guide movement). In which case it is open to the first-order theorist to

insist that it is the former sort of availability which is distinctive of phenomenal consciousness.

This is a perfectly reasonable point; but it is one which I myself make and elaborate on at some length, in chapter 6:3.3 of my 2000. The challenge which I present to the first-order theorist in reply, is to say what it is about *presence to conceptual thought* which makes the difference. I claim that there is nothing illuminating which can be said, here; in which case we don't yet have a reductive explanation of phenomenal consciousness. In contrast, I claim that my own higher-order account *can* explain why contents which are available to higher-order concepts should become different: it is because they acquire, at the same time, a higher-order perceptual content (hence giving them a dimension of *seeming*, or of *subjectivity*), resulting from the truth of some form of consumer semantics.

Of course my overall case against first-order representationalism 'is far from watertight', as Bermúdez points out. Nothing is watertight in this domain, in my view, since our task is to find the best overall explanation of the phenomena. There will always be other choices which *can* be made, with their attendant costs and benefits. My point is just that first-order theories don't achieve all that one might hope for in the way of an explanation, whereas higher-order theories fare significantly better.

3 Do reinforcers need to be phenomenal?

Finally, Bermúdez thinks that he has a knock-down argument that non-human animals have to be regarded as phenomenally conscious, at least in so far as conditioning theory applies to them. This is because (he says) positive and negative reinforcers (such as pain) have to be phenomenal in order to be effective. I quote, 'It is impossible to divorce pain's status as a negative reinforcer from its feeling the way it does.'

There are two distinct mistakes here. The first is that it might well be possible, on the contrary, for the motivational effects of pain to be achieved in the absence of any felt quality. As is now well known, pain perception is subserved by two distinct neural pathways: the old path, which projects primarily to the limbic system, and which is responsible for the awfulness of pain; and the new path, which has multiple projection points throughout the cortex, and which is responsible for fine discrimination and feel. It is also well known that certain types of morphine, and certain kinds of neurological damage, can suppress the old path while leaving the new path intact. In such circumstances people say that their pains *feel* just the same as they did, but that they no longer care about them. Although unlikely in practice, in principle it might then be possible to secure the reverse effect in humans: to suppress the feel of pain while leaving the motivational side intact.

Much more importantly, however, Bermúdez doesn't seem to realize the extent of the resources available to representationalist theories of phenomenal consciousness. I (like Tye, 1995) am a first-order representationalist about pain perception. I think that *feeling a pain* is a matter of being in a state which represents a certain secondary quality (*pain*) as being distributed in a region or surface of one's body, just as *seeing red* is a matter of being in a state which represents a secondary quality (*redness*) as distributed in a certain region or surface of the external world. Neither of these kinds of state is intrinsically phenomenally conscious, for me. Just as I think that there can be percepts of red which aren't phenomenally conscious (in blindsight, for example), so I think that there can be percepts of pain which aren't phenomenally conscious (this would be true of most animals, on my account). (See my 2000, chapter 7:3.5.) So conditioning theory can perfectly well apply to non-human animals: we can say that the reason why the rat learns not to touch its food-dish when a light is on (or whatever), is because it *feels pain* if it does. But it is entirely consistent to claim that the pains which the rat feels are nevertheless not phenomenally conscious ones. (Of course this isn't *intuitive*; but then who ever thought that a scientific account of phenomenal consciousness – or of anything else, for that matter – should be intuitive?)

References

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press.

Dretske, F. (1995). *Naturalizing the Mind* . MIT Press.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

Tye, M. (2000). *Consciousness, Color and Content*. MIT Press.

Properties, first-order representationalism and reinforcement: Reply to Carruthers

José Luis Bermúdez

Department of Philosophy
University of Stirling (UK)
homepage

Peter Carruthers's comments on my review of his *Phenomenal Consciousness* raise interesting points that it is worth pursuing in more detail. I have followed the headings under which Carruthers organized his own comments.

1. The nature of properties

Carruthers is absolutely right that I was too quick with his distinction between thin and thick ways of construing natural properties. I remain unsure, however, that there really is room in logical space for the position he sketches out on the basis of that distinction. Carruthers, in his general line of response to Kripke's argument against type-identities is trying to offer a sense in which there can be identities between phenomenological feels and (thickly individuated) natural properties even though there are worlds in which the relevant phenomenological feels are instantiated in different properties. He needs, therefore, to explain how these identities can be contingent rather than necessary. It is not clear to me that he succeeds.

It will helpful to put my concern in terms of the analogy he draws with the term "manifest water". "Manifest-water" is put forward as an example of precisely the type of contingent identity claim that Carruthers thinks holds between phenomenological feels and thickly individuated natural properties.

> Suppose that I am interested, not in the underlying constitution of water, but in its manifest properties (*clear, colourless, potable when pure* and so on). And suppose that I introduce a special term 'manifest-water' whose use is to be tied to just those properties (thinly individuated). 'Manifest-water' will track whatever has the requisite properties across worlds, just as 'this type of feel' and 'pain' (used purely recognitionally) track whatever has the requisite phenomenology across worlds.
>
> This need not prevent it from being true, in the actual world, that manifest-water = H_2O, however. It is just that this truth has no bearing on the application of 'manifest-water' in other possible worlds. For although the term 'manifest-water'does not use the properties in question as a mere contingent way of referring to an underlying nature, that need not prevent them from *having* an underlying nature. And *because* the term 'manifest-water' does not use those properties as a mere contingent way of picking out an underlying nature, the modal status of the identity 'Manifest- water = H_2O' will be not necessary but contingent. (Carruthers 2000, 47).

The crucial question here is, To what does the term 'manifest-water' refer? There are two obvious candidates. The first is that it refers simply to the cluster of thinly-individuated manifest properties, in such a way that once one has identified the presence of colourness, potability etc then one has said all that there is to say about the presence of manifest-water. The second is that it is used in a reference-fixing way to pick out whatever property *in this world* is the underlying nature of those manifest properties (which would of course be the natural property of having a molecular structure of two hydrogen atoms and an oxygen atom). Carruthers rejects both of these options. He rejects the second because it would make the identity of water and H_2O a necessary identity, while he rejects the first because it does not seem to allow one to raise the question to which the identity claim 'Manifest-water = H_2O' might be the answer. His position is that 'manifest-water' really

does refer to a natural property, but not in the sort of reference-fixing way that would make the identity into a necessary identity.

The position is attractive, but seems difficult to sustain. How exactly does 'manifest-water' pick out a natural property? We are told that 'manifest-water' will track whatever has the requisite properties across worlds, but it is very unclear what this amounts to. It is certainly true that whenever there is something that counts as manifest-water because it has the appropriate thinly-individuated properties, that thing will also have certain thickly-individuated natural properties. So, we might say that whatever is colourless, potable etc must have some molecular structure (H_2O in this world, XYZ in another world). This would be one way of reading what Carruthers means by saying that "'manifest-water' will track whatever has the requisite properties across worlds". But it hardly follows from this that the term 'manifest-water' *refers to* the thickly-individuated natural property that it tracks in this weak sense – and even less so that manifest-water *just is* that thickly individuated natural property. So, my question for Carruthers is, How exactly should the concept of tracking be understood in order to make the idea of an identity claim plausible?

Carruthers's second strategy against Kripke offers one way of responding to this difficulty – viz. by retreating from the claim that there is an identity between manifest-water and H_2O to the claim that manifest-water is *constituted* by H_2O. This gives one clear sense to the idea that manifest-water tracks whatever has the requisite properties across worlds – on the plausible assumption that wherever there are thinly individuated properties those properties will be constituted by some thickly-individuated natural properties. But will it do the work that Carruthers requires? The notion of constitution is rather slippery. Carruthers interprets it as essentially involving (and perhaps being exhausted by) a logical supervenience claim to the effect that there are no worlds in which the thickly individuated properties are as they are and the thinly individuated properties different or absent. There are various quibbles one might have about this. The first is that the logical supervenience claim has to be significantly more robust than Carruthers envisages – as it stands it allows for a zombie world that is one molecule different from this one.

More importantly, though, is the overall dialectical situation. Authors such as Chalmers have challenged the logical supervenience claim. Carruthers has his own arguments against these challenges. It turns out, however, that his objections to Chalmers hinge crucially upon the conception of property identities that we have been discussing. Carruthers offers his "tracking" account as a way of defusing Chalmers's modal arguments. But there is a danger that he may be moving in a rather tight circle here. If the tracking account relies ultimately on the notion of consitution and the notion of constitution rests upon a logical supervenience claim, then it is hard not to be suspicious of an attempt to use the tracking account of property identity to defend the logical supervience claim. Of course, it may well be that Carruthers does not need the notion of constitution to explain what is going on in the tracking account. But then my earlier question stands. We need to know more about the grounds of the identity claim that he makes.

2. First-order representationalism

I am somewhat puzzled by Carruthers's response to my points about first-order representationalism. I expressed some doubts about his argument from cases of non-conscious experience to the falsity of first-order representationalism. Carruthers's target argument draws on a range of different types of non-conscious experience (ranging from blindsight to absent-minded perception) and suggests that these non-conscious phenomena have roughly the same functional role as conscious experiences – from which he takes it to follow that simply being poised to impact on belief formation and practical reasoning cannot be sufficient for phenomenal consciousness.

As I pointed out in my original review, this line of argument will only work if the examples he offers of non-conscious experiences can plausibly be taken to satisfy the following three criteria:

a) that they be describable at the personal-level

b) that they be genuinely non-conscious

c) that they be more or less functionally equivalent to conscious perceptions

The basic point I made was that each of the examples of non-conscious perception that Carruthers mentions can plausibly be found wanting on at least one of these criteria. The ones that really do look as if they are functionally equivalent to conscious perceptions might well not be genuinely non-conscious, and so on. But at no point did I suggest that the real issue here is whether non-conscious experiences are "available to conceptual thought and reasoning (as opposed to merely guiding movement)".

The area to which "availability to conceptual thought and reasoning" might seem to be relevant is the broadly neuropsychological examples of non-conscious perception. I suggested that there are grounds for thinking that blindsight, prosopagnosia and related disorders may not be functionally equivalent to non-pathological conscious perceptions. But the issue here should not be understood in terms of the simple opposition which Carruthers employs between availability to conceptual thought and reasoning, on the one hand, and availability to guide action on the other. The crucial issue with respect to blindsight patients is the fact that they are generally unable to initiate action in the blindfield without prompting. Blindsight patients are capable of controlling action in the blindfield, but in a purely reactive way. This seems to mark a significant difference between the functional role played by "experiences" in the blindfield and experiences elsewhere in the field of vision. But it is not a difference that needs obviously to be characterised in terms of availability to conceptual thought and reasoning.

More broadly, one might wonder whether Carruthers is not being too crude in thinking that the relation between vision and action can be understood solely in terms of the operation of two systems – an on-line action-guiding system, on the one hand, and "a concept-wielding or concept-involving system whose job it is to build a detailed integrated representation of the environment to guide belief formation and medium-term and long-term planning" (p.166). Surely what is missing in blindsight patients and present in normal subjects falls somewhere between the two – and presents an important clue as to the potential functional role of phenomenal consciousness. One possible suggestion would be that phenomenal consciousness makes possible the on-line *initiation* of action. This is distinct both from the on-line control and guidance of action and from medium-term and long-term planning. Nor need it necessarily go together with the formation of beliefs about the environment.

3. Phenomenal consciousness and reinforcement

Turning now to the final point at issue between us, Carruthers is unhappy with my suggestion that "it is impossible to divorce pain's being a (negative) reinforcer from its feeling the way it does". He makes two points. The first starts from the existence of documented dissociations between the motivational side of pain and its phenomenal side, while the second is that we can understand what it is to feel a pain in a manner that is independent of pain's phenomenal dimension. Let me take these in order.

Here is what Carruthers has to say about the possibility of pain having a negative reinforcement effect without feeling the way it does.

> As is now well known, pain perception is subserved by two distinct neural pathways: the old path, which projects primarily to the limbic system, and which is responsible for the awfulness of pain; and the new path, which has multiple projection points throughout the cortex, and which is responsible for fine discrimination and feel. It is also well known that certain types of morphine, and certain kinds of neurological damage, can suppress the old path while leaving the new path intact. In such circumstances people say that their pains *feel* just the same as they did, but that they no longer care about them. Although unlikely in practice, in principle it might then be possible to secure the reverse effect in humans: to suppress the feel of pain while leaving the motivational side intact.

It is hard to know what to make of this. One of the key methodological tenets in neuropsychology is that we cannot demonstrate the independence of two cognitive phenomena or abilities without demonstrating a *double dissociation* between them. That is, in order to show that A and B are distinct we need to identify examples, not just of A existing without B, but also of B existing without A. The empirical data to which Carruthers refers show merely a single dissociation and it is widely accepted that the existence of a single dissociation does not provide evidence for the existence of a double dissociation (otherwise neuropsychology would be a significantly easier subject). I think, on balance,

that this line of argument is unlikely to succeed until we have some positive evidence that the dissociation holds in the opposite direction.

But what about the second point? Carruthers thinks that feeling pain is not an intrinsically conscious state. He adopts a first-order representationalism about pain according to which "*feeling a pain* is a matter of being in a state which represents a certain secondary quality (*pain*) as being distributed in a region or surface of one's body, just as *seeing red* is a matter of being in a state which represents a secondary quality (*redness*) as distributed in a certain region or surface of the external world". And, as he point out, if one is a first-order representationalist about pain then it makes perfectly good sense to say that a negatively conditioned rat acts the way it does because of the pain it feels, even though that pain is not conscious.

This is a perfectly reasonable point, but it fails to address the significant issue, which is one of explanation. Has Carruthers really explained what makes negative reinforcers negatively reinforcing by giving us an account of what it is to feel a pain that does not involve that pain being a conscious pain? Why, one might wonder, should "representing a certain secondary quality as being distributed in a region or surface of one's body" reinforce behaviour that leads to that secondary quality no longer being distributed in that region or surface of one's body – unless there is some phenomenological reason for wanting to get rid of the secondary quality in question? Such evidence as there is suggests that simply representing a harmful stimulus, without the appropriate affective/phenomenological response, will not be sufficiently motivating. I quote from the entry on pain in the *MIT Encyclopedia of the Cognitive Sciences*:

> These extreme cases of pain and the need to control them and other lesser or more acute nociceptive events often raise questions of the advantage conferred by painful affect. Rare clinical cases of patients who perceive a painful event as differing from an innocuous stimulus but who experience no affect accompanying that event are test cases for such a question. Most of these patients die at an early age, victims of numerous destructive wounds and crippling conditions of joints. Apparently the failure of these patients to avoid or discontinue actions that are painful significantly shortens their lives, despite intensive training in detecting and responding to painful stimuli. (p. 624)

If this is right then it looks very much as if the processes of conditioning do *not* work as they should in the absence of phenomenal/affective consciousness. This does not bode well for Carruthers's account of negative conditioning in terms of (non-consciously) feeling pains.

Commentary on

Carruthers, Peter. *Phenomenal Consciousness*. Cambridge: Cambridge University Press, 2000.

Joseph Levine

Department of Philosophy
Ohio State University
Homepage

In Phenomenal Consciousness, Peter Carruthers defends a dispositionalist higher-order theory of phenomenal consciousness. On this view, to enjoy a phenomenally conscious experience is to occupy an analog representational state that is accessible to a faculty of higher-order thought. In the course of defending his view, Carruthers takes on an impressive array of arguments and opposing positions. He addresses "mysterians", who believe that no materialist theory is adequate to an explanation of phenomenal consciousness, as well as first-order representationalists and actualist higher-order theorists. Since his theory is itself a representational theory, he also engages the debate over narrow content, arguing that narrow content, in the form of a consumer semantics, is the appropriate theory of content for psychological explanation.

As a mysterian myself, let me begin my critical comments by addressing his response to that position. As he notes, there are really two positions here, one metaphysical and the other epistemological, though they both share a reliance on conceivability considerations. The metaphysical argument goes like this. Since zombies (or zombie worlds) are conceivable, they are possible. Though it is not a straightforward matter to infer what is possible from what is conceivable, in this case there is reason to think the inference goes through; so goes the argument, at any rate. (Both Kripke 1980 and Chalmers 1996 defend the inference in this case, using different, but related, arguments.) But if zombies are possible, then the physical facts do not metaphysically determine the mental, or phenomenal facts, and thus some form of dualism is true.

I am generally sympathetic to Carruthers's response to the metaphysical argument (see Levine 1998 and 2001, chapter 2). He insists on distinguishing concepts from properties, at least if by "property" one means "natural property" - the sort of feature by virtue of which objects change and events happen. Distinct concepts can pick out the same property, and so what seems conceivable may not be possible. For instance, one might be thinking that some object could be F without being G, but just not realize that the property of being F just is the property of being G. The situation envisaged, though conceivable in the sense that there is no incoherence or contradiction in the description of the situation, is still impossible. Though there are complications here, I believe this general line is basically right.

Still, I want to quibble with some of what Carruthers has to say on this point. With respect to Kripke's argument against the type-identity of pain and cfiber-firing, Carruthers claims that we have two options by way of reply: First, maybe "pain" isn't a rigid designator, so "pain = cfiber-firing" isn't necessary after all. Second, maybe physicalism about phenomenal consciousness doesn't require an identity claim, so "pain is constituted by/realized in cfiber-firing" is good enough. I think neither of these replies works.

With regard to the first option, it's supposed to go like this. "Pain" is akin to "manifest-water", which picks out whatever substance in a world instantiates the manifest properties of water. But "manifest-water" also expresses a property that is shared across worlds, the property by virtue of which something in a world falls under that concept. What is the corresponding property for "pain", if not a phenomenal property? This is of course Kripke's (and Chalmers's) point.

Carruthers seems to think this point is already addressed by his adoption of what he calls his "thick" theory of properties. This comes up prominently in his reply to Chalmers as well,

when he denies significance to the fact that the primary intension of "pain = cfiber-firing" is contingent. He argues that primary intensions are mere functions from possible worlds to extensions, and therefore irrelevant to the identity of honest-to-God properties; that is "natural" properties, the mind-independent features of objects that figure in what happens to them. He says that "grue" and "bleen" also determine such functions, yet no one thinks of them as genuine properties (unless of course they are just collapsing the concept/property distinction altogether).

But I don't think this quite gets him off the hook. First of all, I think there is an intermediate level of property in between the full-fledged natural property (the sort that figures in causal generalizations, etc.) and the one that just collapses into a concept. Being grue, after all, is an objective condition that objects either satisfy or not, and this is so whether or not human beings had ever thought of the property. I agree that it is not a causally significant property, but it seems to me pretty clear that there is a determinate fact of the matter whether or not something is grue. It's also the case that we can form distinct concepts of grue: for instance, the concept expressed by the canonical description "green before 2001 and blue after" and the concept expressed by "has reflectance G [put in the technical details here] before 2001 and reflectance B after".

The point is that it would be a real problem for physicalism if there weren't a supervenience claim of the form "Grue metaphysically supervenes on P", where "P" designated a physicalistically respectable property. Of course in the case of grue there is such a property (or group of properties). If you specify all the reflectance properties of all the objects at all times, you have thereby metaphysically determined the distribution of grue-facts. So even if "pain-feeling" is on a par with "grue", it seems to me we still have a problem if Chalmers is right that the pain-feeling facts do not logically (or metaphysically) supervene on the physical facts.

So why do I think, nevertheless, that by distinguishing properties from concepts one can avoid the anti-physicalistic consequence of the conceivability argument? My long-winded argument is cited above, but the short answer is this. I think one must reject the assumption that we have a priori access to primary intensions. That is, I endorse what Chalmers calls "strong metaphysical necessity" - the idea that a primary intension can be both necessary and a posteriori. Though Chalmers argues that this involves a bizarre metaphysics of modality, I argue that it's quite benign in the end. For the details, see the works cited above.

Let's turn now to the epistemological side of the mysterian position. As I argue (see Levine 1983, 1993, and 2001, chapter 3), there is an explanatory gap between the physical (broadly construed, so as to include functional or computational states as well) and the phenomenal. Support for the existence of the explanatory gap comes partly from the fact that zombies - physical and functional (including representational) duplicates of me but without phenomenal consciousness - are conceivable. Again, I don't infer that they are therefore possible as well. But merely from the fact that they are conceivable, I argue, we see that we can't explain phenomenal consciousness by reference to those properties we and the zombie are stipulated to share. The point is, we have a conception of phenomenal consciousness such that we can't really understand how it could be constituted by (broadly) physical processes.

Carruthers employs two strategies to overcome the explanatory gap. On the one hand, he lists what he takes to be the principal desiderata on an adequate explanation of phenomenal consciousness, and then argues that his theory meets them all. On the other hand, whatever qualms remain over the conceivability of zombies he deals with by appeal to the notion of a recognitional concept. In other words, whatever legitimate explanatory demands there are can be met by his theory, and what can't be met is explained away.

The legitimate explanatory demands are these: What endows phenomenal states with subjectivity, there being something it's like to occupy them? Why do they seem ineffable, private, to involve intrinsic properties of experience, and to be accessible in a privileged manner? I have my doubts that Carruthers's theory does successfully meet these explanatory demands (my doubts chiefly concern the first one), but let me put this aside for now and turn to the question of what seems to remain.

One way to put the matter is this. When I wonder how a (broadly) physical process could be like this (mentally ostending my visual experience of red, say), I'm not wondering how

such a process could be such that I'm tempted to judge it to have this feature or that. I can certainly imagine a computational architecture that would lead the subject of that architecture to judge that certain of its states were ineffable, say, or couldn't be doubted. But the explanatory gap isn't between a description of the computational architecture and a description of certain propensities to make judgments. (Notice that zombies would definitely have the same propensities for judgments that we do, so the conceivability argument doesn't even apply here.) Rather, it's between the description of the computational architecture and a description of the experience itself. It's the character of my visual experience of red that I want explained, not my tendency to judge that it's ineffable, etc..

Fair enough, Carruthers might say, but this is where the appeal to the notion of a recognitional concept does its work. If what you want out of an explanation, goes the argument, is that zombies should no longer be conceivable - in other words, that one should be able to derive a description in first-person terms of someone's having a visual experience of red from a description in third-person terms of their physical or computational structure - then you won't get it. But the reason is perfectly benign, from a naturalistic perspective. The problem is that one's first-person concepts of phenomenal properties are recognitional concepts, and much as one can't derive indexical or demonstrative descriptions from purely non-indexical or non-demonstrative descriptions (see Perry 1979), one can't derive descriptions involving recognitional concepts from descriptions not containing such concepts. So in this sense the explanatory gap is not exactly bridged, but still explained away, and therefore rendered harmless.

Perhaps the best way to make the case, and also, to my mind, to see what's wrong with it, is by way of Carruthers's own example of "bare color". He asks us to imagine that there exists a creature who makes "bare-color" judgments, purely recognitional judgments unconnected to any concepts about normal lighting conditions and normal observers, or to any phenomenal properties. Like chicken-sexers, they just judge "this is red" without having any idea how they do it, without there being any basis they are aware of for their judgment. They just respond. Such creatures would experience an explanatory gap, since they could conceive of the physical story for red holding without something's being red, i.e. without their applying their recognitional concept. Yet it is clear that these creatures pose no problem for the explanatory reach of a physical (or computational) theory of color recognition.

But what strikes me as the real import of the example is how different bare-color responses are from genuine phenomenal experiences, and this bears directly on the adequacy of the appeal to recognitional concepts as a way of responding to the explanatory gap. After all, we don't feel tempted in the chicken-sexing case, nor in the hypothetical bare-color example, to see a problematic explanatory gap present. What, then, explains the difference between these cases and genuine phenomenal experience?

When I wonder how this experience (again, mentally ostending my current visual experience of something red) could be constituted by such-and-such computational/physical processes, I have a substantive and determinate conception of - I am, as it were, directly acquainted with - a feature of my experience. The reddishness doesn't have the character for me of a something-I-know-not-what that happens to be prompting this response, as it does for the bare-color responder. The point is it's the reddishness I am wondering about, not about my tendency to regard it as intrinsic, nor about my tendency to respond to it with a certain judgment, or anything else of that sort. In the bare-color case, as in the chicken-sexer case, there isn't any substantive and determinate content to my judgment over and above the response itself. Put another way, "that again" carries a much richer content in the phenomenal case than it does in the bare-color case. What's more, it is precisely what corresponds to that richer content that is the focus of my puzzlement when judging that there is an explanatory gap.

This last point is crucial, for it is natural at this point to respond that our phenomenal concepts are not in fact purely recognitional, and therefore of course they possess a richer content than is possessed by the corresponding concepts of the bare-color responder. In addition to the recognitional component, on Carruthers's view, is a descriptive component that derives from our stock of folk-psychological concepts. So we have concepts of perception, experience, and the like, with which we can fill out our concept of "that again", to make it more like "that perceptual experience of something red, again". But this doesn't address the sense of richer content at issue here, for there is no explanatory gap when it

comes to folk psychological concepts, so long as they are understood functionally, as Carruthers understands them. (Of course, if our concepts of experience and perception are not understood in this third-person way, but rather in the first-person manner of phenomenal concepts generally, then the argument begins all over again with respect to them.)

Above I used the term "direct acquaintance", which of course carries a lot of philosophical baggage. Let me hasten to say that I have no theory of our epistemic access to our own phenomenally conscious experiences. What does seem clear to me is that not only is the explanatory gap a problem about phenomenally conscious experiences themselves, but also about how we can have the sort of cognitive access to their content that we seem patently to have. The two aspects of the problem of phenomenal consciousness are just two sides of the same coin. How can we be subjects of experience, where this involves our occupying states that are somehow themselves cognitive apprehensions of themselves? In fact, this question leads naturally to a discussion of Carruthers's positive view, since his view can be seen as precisely an attempt to answer it.

However, there is one more point about conceivability arguments I want to address first, and this has to do with Carruthers's discussion of the possibility of inverted qualia. On his view, two factors make a state phenomenally conscious: its being an analog representational state and its being available to a faculty of higher-order thought. The zombie, or absent qualia problem is a challenge to that aspect of the theory. But a theory of phenomenal consciousness has to tell us not only what makes a state phenomenally conscious to begin with, but also what determines its precise phenomenal content; what distinguishes reddish visual experiences from greenish ones. With respect to this question, Carruthers maintains that phenomenal content is determined by representational content, and it is at this point that the problem of inverted qualia becomes relevant.

If inverted qualia are possible, then two individuals might be occupying the same functional/intentional state, and yet be having distinct experiences - one that is reddish, say, and the other greenish. But then phenomenal content couldn't be determined by intentional content. So Carruthers must deny the possibility. He employs two strategies. First, he allows that inverted qualia are indeed conceptually possible, but argues, as with absent qualia, that their conceptual possibility does not entail their metaphysical, or "natural" possibility, and it is only the latter that really matters to a naturalistic theory of phenomenal consciousness. Second, he appeals to his version of narrow content theory - an inferential role consumer semantics - to counter arguments to the effect that inverted qualia are indeed naturally possible.

With respect to the first strategy, I basically agree that if inverted qualia are only conceptually possible that this doesn't constitute an objection to the intentional theory of phenomenal content. But I'm puzzled by Carruthers's insistence on the distinction between metaphysical and natural possibility, and his claim that it is only the latter that matters here. I would have thought that it is metaphysical possibility that mattered. Now, given the arguments Carruthers in fact uses, the distinction may not be crucial. However, I can think of a situation in which the distinction might matter, and it's unclear to me what Carruthers would say about it.

So, suppose that as a matter of natural law it's impossible for a perfectly symmetrical information processing system to be physically realized. Let's also suppose that from the point of view of the theory of computation, however, there is no obstacle to such systems (and suppose they could include analog representations as well). It's just that for some reason (which, by the way, it seems hard to imagine what it could be) you can't actually build (or evolve) such a device. So we know that our perceptual system couldn't be symmetrical, and therefore invertible, and we know no naturally possible (I assume that means physically possible) perceptual system could be, but there's no reason to suppose that no metaphysically possible system could be.

Well, consider such a system in world W1, call it S1. By Carruthers's theory, according to which phenomenal consciousness supervenes on functional/intentional properties, S1 (assuming it's attached to an HOT faculty) contains phenomenally conscious states whose phenomenal contents are determined by their intentional contents. Let system S2, in W2, have intentional contents inverted around the axis of symmetry. We can of course insist that the phenomenal contents go with the intentional contents, but we have the Shoemaker and Block scenarios, discussed by Carruthers, to contend with. My only point is that if one

found those scenarios convincing in the first place, they ought to bother one in this case, and, by stipulation one can't appeal to constraints on the realization mechanisms to get around the problem. So it looks as if cases of inverted qualia can be metaphysically possible even if not naturally possible.

As I say, it's not clear to me how much the distinction between natural and metaphysical possibility matters here because I take Carruthers in the end to be arguing that inverted qualia are not even metaphysically possible, though they are conceptually possible. But his argument for this is obscure, at least to me. At one point he says: "if [the notion of narrow content] is legitimate, then the teeth can drawn from all forms of experience inversion...". My problem is that I can see how the appeal to narrow content insulates him from the sort of inversion argument that troubles externalist representationalists - in particular Block's Inverted Earth argument (see Block 1990) - but how does it help with the sort of inversion argument, like Shoemaker's, which is aimed at traditional functionalist theories? After all, these theories have usually been understood to be internalist. (That's certainly how Shoemaker conceived of functionalism when worried about inverted qualia in Shoemaker 1984.)

In fact there's a curious transition in Carruthers's discussion. He begins by discussing Shoemaker's argument, which involves memory switching, etc., and concludes that it's inconclusive on the question of genuine possibility (I'll use this term to cover both metaphysical and natural possibility), as opposed to conceptual possibility. He then brings up Block's Inverted Earth argument as something that pushes us in the pro-qualia direction, but then replies that if we have the requisite notion of narrow content we don't have to worry about Block's argument (that's where the quote above comes in). But what happened to the original Shoemaker argument? Why does the legitimacy of the requisite notion of narrow content "draw the teeth from all forms of experience inversion", rather than just from those that address externalism, as Block's case does?

Finally, let me say why in the end I think it's very plausible to think inverted qualia are genuinely possible. First, I assume that symmetrical internal systems of functional roles are genuinely possible. Again, it's hard to see how one could rule them out, especially their metaphysical possibility. Second, I assume that symmetrical external quality spaces - of the sort that would be the external contents of some perceptual system - are also genuinely possible. Once we have these two symmetrical systems - internal roles and external quality spaces - how could one determine the identity of a single phenomenal content by its location in either space?

Suppose one wanted to do it purely internally. True, qualitative difference can be accounted for internally, so red and green can be distinguished (or the analogues for red and green in the symmetrical system under consideration). But appeal to location in the relevant system of relations won't account for what makes red red and green green, and not the other way around. To do that, it looks like one would need to appeal to the identities of the external properties to which each is responsive. But if the relevant quality space is itself symmetric, then this won't work either, for the sorts of reasons anti-externalists like Block point out. This problem - that you can get a functional account of qualitative difference but not a functional theory of particular qualitative characters - was precisely Shoemaker's point in the papers cited above, and why he opted in the end for a physiological type-reduction theory for individual qualia (not that I think that move works either; see Levine 1989 for a detailed discussion of Shoemaker's position).

Let's turn now to consider Carruthers's positive theory. As mentioned above, Carruthers defends the view that phenomenally conscious states are analog representational states that are available to a faculty of higher-order thought. He distinguishes his view from a number of alternatives, but two in particular take up the bulk of his discussion: first-order representationalism (FOR) and actualist higher-order representationalism (HOR). FOR is the view that phenomenally conscious states are analog (or non-conceptual, the difference doesn't matter for our purposes) representational states that are available ("poised", as Tye 1995 puts it) for cognitive processes such as control of behavior, rational planning, etc.. Actualist HOR theory is the view that to be phenomenally conscious is to occurrently entertain a higher-order representation that one is in the state in question. Carruthers's position differs from the latter in not requiring that there actually be an HOR directed on the phenomenally conscious state, just that it be accessible to the faculty so that one could direct an HOR on it. His position differs from FOR in that mere availability to planning,

behavior control, and the like is not sufficient; it must be a capacity for forming HOR's to which the state is available.

The main problem with FOR, according to Carruthers, is that it can't provide a principled distinction between "experiences" or perceptual states that are phenomenally conscious and those that aren't. There is abundant evidence both from ordinary experience and from scientific psychology - from absent-minded drivers to cases of blindsight - to the effect that there are perceptual states that are not phenomenally conscious. While advocates of FOR might want to appeal to the "poised" condition to account for the non-conscious status of these perceptual states, Carruthers argues that there is solid evidence that they do contribute to behavior control, which includes interacting with unconscious beliefs and desires as well. So why don't they count as phenomenally conscious?

There is one point Carruthers makes repeatedly in response to various attempts on the part of FOR advocates to provide a principled distinction between conscious and non-conscious states: it is not sufficient to merely find some difference that is extensionally accurate, but the difference in question must actually explain why the phenomenally conscious states have this feature of subjective feel and the others don't. So, for instance, in response to the suggestion that the crucial property is being accessible to the highest-level decision-making faculty, Carruthers complains: "How can the mere fact that an analog content is now in a position to have an impact upon the highest-level decision-making processes confer on it the subjective properties of feel and "what-it-is-likeness" distinctive of phenomenal consciousness?" (page 170) I am quite sympathetic to this complaint, and it is crucial to the development of his own view.

Now some have pressed a similar complaint against (actualist) HOR theory, but advocates reply, with some justice I think, that this objection betrays a misunderstanding of their position. The entire point of (actualist) HOR theory is that being phenomenally conscious is not a monadic property of the first-order state, but a relation holding between the higher-order state and its target state. This idea has problems of its own (see Neander 1998, Byrne 1997, and Rosenthal 2000), but still it has a principled answer to what makes some states conscious and others not: namely, states are conscious when we are conscious of them, which means that they are the intentional objects of higher-order thoughts.

The point I'm getting to is this. Carruthers levels a serious objection against FOR, one to which, at least on the surface, actualist HOR is not vulnerable. However, once Carruthers abandons the actualist version of HOR for the dispositionalist version, the objection comes back to rear its ugly head; why, after all, should mere availability to a higher-order faculty render a first-order analog perceptual state phenomenally conscious? The actualist response that for a state to be phenomenally conscious is for one to be phenomenally conscious of it only works if there is an occurrent higher-order state. A mere disposition for there to be such an would seem to endow the first-order state with at most a disposition to be phenomenally conscious.

Carruthers is of course aware of this objection, and it is precisely to meet it that he employs his consumer semantics. The idea is this. First-order analog perceptual states, when not functionally attached to a higher-order faculty, have intentional contents that are about the features of distal objects (though the contents are narrowly individuated). But when serving within a mental system that contains a higher-order faculty, one that, in particular, is capable of noting an appearance-reality distinction, the analog perceptual states, by virtue of their connection to this faculty "down-stream" (hence a consumer semantics), acquire a new content, one that not only represents the features of distal objects, but also themselves as representing these distal features. Instead of "red at 4 o'clock", the content is "seems to be red at 4 o'clock", or, maybe, even more explicitly, "I'm detecting red at 4 o'clock". We now have back the analysis of "is conscious" in terms of "is conscious of", but without the need to appeal to an occurrent accompanying higher-order state.

I have three problems with this view. The first is actually not aimed at the dispositionalism of the view, but applies to any version of HOR. It's crucial to Carruthers's case, especially against FOR, that it be possible for there to be unconscious experiences. Again, this is a feature of any version of HOR. Now Carruthers makes a strong case that there are unconscious analog perceptual states - and of course if you want to stipulate that these be called "experiences", there's no stopping you. But I think it's clear that in conscious experience we encounter a property that is essentially phenomenal, in the sense that its

instantiation is "for" a subject. Put another way, pains, tickles, and sensations of red all seem to be essentially modes, or bits of conscious experience. Of course you can have the representation of bodily damage or distal color properties without any phenomenal feel, no one doubts that. But whether such registrations of properties would be the same sort of event as a painful feeling, or a reddish experience, seems highly doubtful.

I present this objection to register my disagreement with Carruthers on this crucial point. But I mention it only to put it aside, since Carruthers, with some justice, can accuse me of begging the question against him. After all, he denies there are qualia in the sense of intrinsic properties of experience, and my objection appeals to such qualia. Unfortunately I can think of no clearly non-question-begging way to put the objection. If I do, I'll let you know. (For a discussion of how the burden of argument plays out here, see Levine 2001, chapter 5).

To state my second problem I refer back to my remarks above about the conceivability argument. Appeal to a disposition to form HOR's doesn't really explain phenomenal consciousness, and the conceivability of having the former without the latter is evidence of this. Of course Carruthers has his general reply to the conceivability argument, and I have already commented on that. The reason I bring this up again here is that it might seem as if his argument above concerning the role of recognitional concepts in explaining the appearance of an explanatory gap is strengthened by his account of the dual content expressed by analog perceptual states when attached to an HOR faculty. So it is only this aspect of his response to conceivability considerations that I want now to address.

Carruthers argues that just as one can form a recognitional concept of distal features like red on the basis of one's first-order analog states, so too one can form a recognitional concept of experience-of-red on the basis of one's higher-order representations of one's first-order analog states. These, then, are the recognitional concepts appeal to which allegedly accounts for the appearance of an explanatory gap.

Now part of my objection to this move above had to do with what I claimed was a clear difference between the cognitive relation that obtains between the recognitional concept of "bare-color" and our standard cognitive relation to experiences of red and green. I claimed that the former lacked the substantive and determinate character of the latter, which pointed to a kind of cognitive intimacy not provided for in the relation between recognitional concept and what it is a concept of.

However, it might seem now that Carruthers has a way of addressing this concern. It's not that the recognitional concept which is applied when I am aware of my own experience of red is a distinct representation from its first-order target, but rather it is one and the same representation, though its secondary content. When my first-order analog state simultaneously says "red at 4 o'clock" and "seems red at 4 o'clock", it is simultaneously applying the concept "red" to the distal object (or spatial region) and the concept "is a red appearance" (or a "red-seeming") to itself. This reflexivity in the secondary content might plausibly account for the sort of cognitive intimacy that seems apparent in our apprehension of our own sensory experiences.

It isn't clear to me if Carruthers in fact takes this line, but, whether or not he does, I don't think it holds up. The problem is that the thoughts we have when entertaining conceivability thought experiments, or puzzling about the explanatory gap, are genuine thoughts, with full-fledged concepts as constituents. That we might form recognitional concepts of first-order perceptual states within the representational system of the HOR faculty makes perfect sense. But then these representations are distinct from those in the first-order perceptual faculty. When I think, "Oh, another one of those sorts of experiences again", it isn't the experience itself but a thought about it that's in play.

This leads naturally to my third problem, which relates directly to the idea that because of their effects down-stream, first-order analog perceptual states take on a secondary higher-order content. The proposal just seems ad hoc. I don't see why, merely by virtue of its availability to a higher-order faculty, a state's content should thereby take on some of that higher-order character itself. What basis is there for assigning this secondary higher-order content when interpreting the system as a whole, instead of merely leaving the analog perceptual states with their first-order contents and assigning the higher-order cognitive faculty the content that there's been a representation of a certain distal feature reported by

the perceptual faculty? I see nothing in the consumer semantics that forces the former interpretive scheme, and therefore it seems merely ad hoc to insist on it.

The idea, as I understand it, is that once the down-stream cognitive faculties are capable of discerning the difference between something's seeming red (i.e. causing the red-registers to go off) and really being red, then the consumer, these cognitive faculties, come to interpret (as it were) the red register going off as meaning "seems red". Consumer semantics, after all, is a matter of the significance of the representation for the consumer - in this case, higher-level cognitive faculties. But given the lack of any syntactic markers of reflexivity in the analog perceptual states, it seems a much better overall interpretation of the system to assign the usual distal contents to the first-order analog states and locate the appreciation of their constituting a "seeming" in the representational states of the higher-order faculty itself, where the appearance-reality distinction is explicitly appreciated. What you get then is a down-stream state saying "seems red at 4 o'clock", which, in effect, can be glossed as, "a report of red at 4 o'clock coming from the visual faculty". Pushing the "seeming" all the way down into the perceptual contents themselves seems gratuitous.

Consider the following analogy. I wear my watch all the time and rely on it constantly to give me the time of day. Of course, I know that watches are notoriously prone to run fast or slow, so I take its reading with a grain of salt. When asked what time it is, I can imagine being in a careful enough mood to say, "well, by my watch it's 2:15", or "it seems to be 2:15". Does that mean that my watch is reporting on its own state when the big hand is on the three and the little hand on the two? I think my watch is indicating the time, and then I qualify the content of my own representation when I don't completely trust it. Of course there is a somewhat metaphorical sense in which the watch is also telling me that its hands are in the requisite positions. But that isn't a matter of its content, but rather of its being so - its having its hands where they are - and my ability to detect that fact. My detecting that fact is where to locate the representational content that says the hands are where they are. Similarly, that it's seeming to me as if there's red at 4 o'clock is a content that is appropriately located in the representation that detects the first-order perceptual state; not in the first-order state itself.

Though I'm no fan of consumer semantics, let me emphasize here that my objection is not an objection to this sort of semantics per se. I take the point of positing narrow contents, including consumer-based ones, to be to overcome the Frege problem. By appeal to inter-representational relations one can more finely individuate contents and thereby account for differences in cognitive significance that don't track differences in reference. Let that be so. Indeed, let it be the case that we can use these down-stream relations to distinguish the perceptual contents of two creatures both of whose perceptions are tracking the same distal feature. In particular, let it be that "red at 4 o'clock" in a creature with an HOR faculty has a different narrow content from its corresponding state in a creature without such a faculty, because narrow content is a function of, among other things, intra-mental relations. Still, it's a big leap from acknowledging this difference to then saying that the one attached to the HOR faculty is itself partly higher-order. Why can't the mere difference in its causal relations be the difference in content? What makes it necessary to put a reflexive gloss on that difference? I don't see the motivation, except that it helps with the theory of phenomenal consciousness. But that's precisely why it's ad hoc.

In this commentary I've criticized a number of aspects of Carruthers's view. As one with mysterian leanings, it's not unexpected that I would find material with which to disagree. But this shouldn't obscure my appreciation for the impressive achievement this book represents. By pressing the anti-mysterian case as forcefully and honestly as he does, with such a wide array of philosophical and psychological argumentation to back him up, Carruthers has done the "consciousness business" a great service.

References

Block, N. (1990). "Inverted Earth", Philosophical Perspectives, 4: Action Theory and Philosophy of Mind, J. Tomberlin, ed., Atascadero, California: Ridgeview Publishing Co., 53-80.

Byrne, A. (1997). "Some Like It HOT", Philosophical Studies, 86, 103-129.

Chalmers, D. (1996). The Conscious Mind. Oxford: Oxford University Press.

Kripke, S. (1980). Naming and Necessity. Cambridge, MA: Harvard University Press.

Levine, J. (1983). "Materialism and Qualia: The Explanatory Gap," Pacific Philosophical Quarterly 64, 354-361.

(1989). "Absent and Inverted Qualia Revisited", Mind & Language, vol. 3, no. 4, 271-287.

(1993). "On Leaving Out What It's Like", in Davies, M. and Humphreys, G., eds., Consciousness: Psychological and Philosophical Essays, Oxford: Blackwell, 121 - 136.

(1998). "Conceivability and the Metaphysics of Mind", Nous 32, 449-480.

(2001). Purple Haze: The Puzzle of Consciousness, Oxford Univ. Press.

Neander, K. (1998). "The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness", in J. Tomberlin, ed., Philosophical Perspectives 12: Language, Mind, and Ontology, 411-434.

Perry, J. (1979). "The Problem of the Essential Indexical". Nous 13, 3 - 21.

Rosenthal, D. (2000). "Metacognition and Higher-Order Thoughts", Consciousness and Cognition, 9, 231-242.

Shoemaker, S. (1984). Identity, Cause, and Mind. Cambridge: Cambridge University Press.

Tye, M. (1995). Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. Cambridge, MA: Bradford Books/M.I.T. Press.

Reply to Joe Levine

Peter Carruthers

University of Maryland, College Park

My thanks to Joe Levine for his extensive and insightful comments on my book (Carruthers, 2000). In this reply I shall focus on the main points of disagreement between us - namely, the existence of an alleged 'explanatory gap'; the supposed possibility of experiential inversion; and Levine's criticisms of my proposed dispositionalist HOT theory.

1 The 'explanatory gap'

Levine concedes some force to my attempts to mark a divide between those aspects of the phenomenon of phenomenal consciousness which can be successfully reductively explained by a naturalistic theory such as dispositionalist HOT theory, and those aspects which are best explained away. A large part of the work of the latter sort is borne by the notion of a purely-recognitional concept; and Levine uses my analogy based on the example of 'bare color' as a stalking horse on which to ground his critique. Unfortunately, he gets the example wrong.

Unlike the example of the chicken-sexer (which does some, but only some, of the same work) the bare-color example isn't supposed to involve a concept which we find ourselves applying for no apparent reason, with applications of it having no 'substantive and determinate content … over and above my response itself'. On the contrary, it is supposed to be an example of a concept whose applications are grounded in first-order (but non-phenomenal) analog intentional states. We are to imagine a creature possessing analog (or non-conceptual) intentional states representing reflective properties of surfaces, and with the capacity to apply purely-recognitional concepts grounded in the intentional character of those states, but who is altogether lacking in phenomenal consciousness of color.

(Of course, whether you think the example is even so much as possible will depend upon your prior rejection of first-order representationalist theories of phenomenal consciousness. But this assumption creates no special problem as part of an argument against mysterianism. And anyone who believes in the possibility of zombies will have no difficulty with the bare-color example.)

Now, the point is that such a creature would feel the same kind of explanatory gap as is alleged to exist in the case of phenomenal consciousness, but in this case between physical / functional / intentional facts and the facts of bare-color. For lacking any beliefs about the processes which ground its color-recognition judgments, the creature would always be capable of thinking, 'Yet all of that could be true, and still THIS could be lacking', where 'THIS' expresses a recognitional concept grounded in a first-order analog color-content. Yet this creature lacks any phenomenally conscious color-states. Moreover, most mysterians will allow that non-phenomenal intentional states and concepts admit of reductive explanation in principle. So this is a case where a reductive explanation could exist, although the subject in question would be vulnerable to the same persistent worries which dog attempts to explain phenomenal consciousness.

(Of course it is true, as Levine notes, that WE don't feel any temptation to think that there is any problematic explanatory gap present in the case of the bare-color subject. But that only goes to help my case. The point is that THE SUBJECT in this example WOULD feel a problematic explanatory gap to exist.)

At its weakest, the example shows that the so-called 'explanatory gap' for phenomenal consciousness doesn't really have anything to do with phenomenal consciousness per se. It is, rather, a gap which can arise more generally wherever there are recognitional concepts grounded in analog intentional contents. At its strongest, the example shows that the

'explanatory gap' is not really a gap at all. That certain persistent questions remain, doesn't show that anything is going unexplained.

On the stronger of the above readings, the analogy works because it gives us a case of a recognitional concept grounded in (non-phenomenal) analog intentional states. And in the case of phenomenal consciousness (according to dispositionalist HOT theory) what we have is a set of recognitional concepts whose application is grounded in higher-order analog intentional states (intentional states whose higher-order content derives from the possibility of deploying higher-order concepts in response to them - see section 3 below). So I can grant to Levine that (phenomenal) reddishness 'doesn't have the character for me of something-I-know-not-what that happens to be prompting this response'. But then the same is true for bare-color too. For the bare-color subject doesn't find himself deploying concepts in response to something-he-knows-not-what. Rather, those concepts are grounded in perceptual (but non-phenomenal) awareness of analog properties of surfaces.

2 Inverted spectra

Levine concedes that an appeal to narrow content can undermine externalist forms of argument for the possibility of inverted experiences, such as Block's (1990) example of Inverted Earth. But he worries that I have lost sight of - and have failed to respond to - Shoemaker's original intra-personal inversion case (Shoemaker, 1981). Before I get to that, however, I want to say a word about burden of proof and plausibility.

Mysterians like Levine bring up the possibility of inverted experience as part of an argument for their mysterian position. They are therefore required to argue their case - the burden of proof falls on them to show that inverted experience is possible. And, given any robust form of concept / property distinction, they also need to show that inverted experiences are genuinely - metaphysically - possible, and not just conceptually so.

As I understand it, this is why Shoemaker saw the need to concentrate, in the first instance, on developing realistic intra-personal cases of experience inversion. This is the crucial battle-ground on which the possibility of experience inversion needs to be fought. Merely pointing to the possibility of symmetrical functional roles and symmetrical external quality spaces - as Levine does in his comments - is not going to convince anyone of anything more than a conceptual possibility. For we are given no reason to believe that experiences either would or could be reversed in such a case. It is open to reductive naturalists to claim that if the symmetrical functional roles are genuinely functionally equivalent, then experiential contents will be identical also. No reason would yet have been given to shift us from such a position.

Returning now to the charge that I had neglected Shoemaker's original argument - in fact I did return briefly to such cases following my discussion of Inverted Earth, on page 86 of my book. I claim that there are two stages necessary to liberate us from feeling any force in the intra-personal inversion examples. The first is to see the possibility of appealing to some notion of narrowly-individuated intentional content. For the sake of concreteness, let us suppose that this takes the form of a functional-role semantics. Then the second step is to see that functional roles are individuated, not just by actual causes and effects, but also by counter-factual causes and effects. (Even philosophers of psychology as distinguished as Fodor are apt to forget this.)

So, consider a post-amnesial subject in a supposed intra-personal inversion case: he uses 'seems green' to describe his experiences of green grass, and all his other beliefs and behaviors are as normal; although not long ago, before his amnesia, he could still recall that seeming-green was the experience he used to get when looking at fresh blood (before the pathways in his optic nerve were switched). Is he genuinely the functional equivalent of a normal person? (And so is this a genuine case of experiential inversion with functional / intentional equivalence?) I claim not.

Take the recognitional capacity which the subject now deploys when using the term 'seems green' and ask how that capacity WOULD HAVE responded had it been present pre-optic-nerve-switching, and had it been confronted by the state normally caused by looking at green grass - would it have been activated? Surely not. On the contrary, that very recognitional capacity would have been activated in response to the state normally caused by seeing fresh blood. So the narrow intentional state which now causes him to say 'seems

green' continues to have the content 'seeming-red' - and what we have here is no longer as case of experiential inversion with functional / intentional symmetry.

So although the intra-personal inversion subject is BEHAVIORALLY the same as a normal person, we can say that his internal states are functionally distinct, because different counter-factuals are true of them. And so we can explain, in functional-role terms, how it is that his experiences are different from normal too. And such cases therefore provide us with no reason to embrace mysterianism concerning phenomenal consciousness.

3 Dispositionalist HOT theory

Levine raises three difficulties for my dispositionalist HOT theory. I shall take them each in turn.

The first is that I have failed to make out the case for the existence of non-conscious experiences. Levine (like me) is inclined to deny that the cases I adduce involve phenomenal consciousness; but he thinks that they therefore don't deserve the title 'experience'. I am not going to quarrel about words. He can insist on calling them 'analog perceptual states' if he likes. But Levine has missed the dialectical position, here. The conscious / non-conscious distinction is NOT supposed to be part of any argument against mysterianism - so it is not supposed to be something which should worry Levine. By the time we get to chapter 6 in my book (where these matters are discussed), mysterianism has already been set to one side. Rather, some sort of representationalist approach to phenomenal consciousness is presupposed. And the conscious / non-conscious distinction forms the main premise in my arguments against first-order representational (FOR) theories, of the sorts defended by Dretske (1995) and Tye (1995, 2000).

Levine's second objection I have difficulty in getting straight. It has something to do with an alleged inadequacy in my appeal to recognitional concepts in defusing the conceivability arguments for mysterianism. The worry seems to be about whether dispositionalist HOT theory has provided for the 'substantive and determinate character' of the states to which our recognitional concepts are applied, and for the 'cognitive intimacy' which exists between our phenomenally conscious states and our recognitional concepts for them. But this is followed by a confused and inaccurate presentation of my view.

So here, briefly, is how I think the story should go. When I enjoy a phenomenally conscious experience as of red, I am in a perceptual state with the analog (or non-conceptual) content 'red', which also possesses the analog content 'seems red' or 'experience of red'. (The state in question is thus BOTH a first-order experience of color AND a higher-order experience of an experience of color.) I am then capable of enjoying purely-recognitional concepts for my phenomenally conscious experience, grounded in my higher-order experience of it. The perceptual state with the analog content 'seems red' provides the grounds for me to apply my recognitional concept, in much the same way that a perceptual state with the analog content 'red' provides the grounding for a recognitional application of the concept 'red'. These higher-order recognitional concepts are genuine concepts, capable of figuring in conceivability thought-experiments or in puzzling about the explanatory gap. And in particular, since they lack any conceptual connections with the various elements which go to make up dispositionalist HOT theory, it will always be possible to think, 'Dispositional HOT theory might be true of a creature, and still that creature could lack THESE kinds of states'.

It seems to me that this story provides fully for the substantive and determinate character of the states which ground our thoughts about phenomenal consciousness; and that the cognitive intimacy between such states and our judgments about them is also explained.

Levine's third objection to dispositionalist HOT theory is that he doesn't see why availability to higher-order concepts should transform the contents of our perceptual states, conferring them, at the same time, with higher-order analog contents. He uses the example of his watch-face to make the point. Surely the mere fact that Levine knows that his watch doesn't always keep time isn't enough to make the position of the hands represent 'seems 4 o'clock' as well as 'is 4 o'clock'.

The example is a misleading one, however. For watch-faces are not representations in their own right - they depend upon the minds of people to confer content upon them. Or if they don't - if they are considered representational - then the appropriate notion to apply must be an informational, or some sort of causal-covariance, conception of content. But this is

not the notion presupposed by consumer-semantics. According to consumer-semantics, states which are representational-in-their-own-right acquire their content from the inferential powers of the systems which consume, or make use of, those states.

I don't believe that any sort of special case has to be made out for saying that perceptual states will acquire higher-order content on becoming available to higher-order consumer systems. For this is just what consumer semantics would predict - in general, the content of a state depends upon (and is a reflection of) the inferential powers of the systems which are the immediate and prime consumers for that state; and states which become available to new systems will thereby acquire new contents.

Consider an example from developmental psychology. Perner et al. (1994) maintain that there is a stage in development when children operate with an undifferentiated concept of 'prelief', which is a sort of amalgam of 'belief' and 'pretence'. They understand that preliefs don't always correspond to the way the world is, but don't yet have a conception of a type of state which purports to represent the world as being in a particular way, but does so incorrectly - they don't yet have a conception of 'belief' as such. But as the child's theoretical and inferential powers develop, it reaches the stage at which it can treat pretence and belief as distinct states.

Here we may have a state - e.g. ascribing a prelief to someone - which acquires a new content - it becomes an ascription of belief - as a result of the new inferential powers of the theory-of-mind system which can operate upon that state. Whether or not this particular account of this stage in child development is true, this is all standard consumer-semantic stuff. But would Levine object (echoing the remarks in the penultimate paragraph of his comments), 'It is ad hoc to say that the state in question has acquired a new content'? Would he say, 'Granted there may be some difference in content, but why say that the new content is the content "belief"? Why not say that the mere difference in causal relations is the difference in content?' Such objections would be ill-motivated in the case of belief, unless they are intended as objections to consumer semantics per se. They are equally ill-motivated as objections to dispositionalist HOT theory.

4 Finally ...

One final point of clarification: Levine asks why I sometimes address questions of natural possibility: isn't metaphysical possibility always the relevant modal operator to consider? The answer is: yes, it is. But on many views of natural properties, their identity is tied to worlds in which the laws of nature remain the same. So when questions of property-identity or property-supervenience are at issue, conceivable circumstances which are naturally impossible are also very likely to be metaphysically impossible. Put differently: the relevant notion of supervenience to consider is global, where laws of nature as well as physical / functional facts are held fixed.

References

Block, N. (1990). "Inverted Earth.", *Philosophical Perspectives*, *4: Action theory and philosophy of mind*, 53-80. Ridgeview Publishing.

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Perner, J., Baker, S., & Hutton, D. (1994). "Prelief: The conceptual origins of belief and pretence." In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: origins and development* (261-286). Lawrence Erlbaum Associates.

Shoemaker, S. (1981). "The inverted spectrum." *Journal of Philosophy*, 74.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

Tye, M. (2000). *Consciousness, Color and Content*. MIT Press.

Reply to Joe Levine

Peter Carruthers

University of Maryland, College Park

My thanks to Joe Levine for his extensive and insightful comments on my book (Carruthers, 2000). In this reply I shall focus on the main points of disagreement between us - namely, the existence of an alleged 'explanatory gap'; the supposed possibility of experiential inversion; and Levine's criticisms of my proposed dispositionalist HOT theory.

1 The 'explanatory gap'

Levine concedes some force to my attempts to mark a divide between those aspects of the phenomenon of phenomenal consciousness which can be successfully reductively explained by a naturalistic theory such as dispositionalist HOT theory, and those aspects which are best explained away. A large part of the work of the latter sort is borne by the notion of a purely-recognitional concept; and Levine uses my analogy based on the example of 'bare color' as a stalking horse on which to ground his critique. Unfortunately, he gets the example wrong.

Unlike the example of the chicken-sexer (which does some, but only some, of the same work) the bare-color example isn't supposed to involve a concept which we find ourselves applying for no apparent reason, with applications of it having no 'substantive and determinate content … over and above my response itself'. On the contrary, it is supposed to be an example of a concept whose applications are grounded in first-order (but non-phenomenal) analog intentional states. We are to imagine a creature possessing analog (or non-conceptual) intentional states representing reflective properties of surfaces, and with the capacity to apply purely-recognitional concepts grounded in the intentional character of those states, but who is altogether lacking in phenomenal consciousness of color.

(Of course, whether you think the example is even so much as possible will depend upon your prior rejection of first-order representationalist theories of phenomenal consciousness. But this assumption creates no special problem as part of an argument against mysterianism. And anyone who believes in the possibility of zombies will have no difficulty with the bare-color example.)

Now, the point is that such a creature would feel the same kind of explanatory gap as is alleged to exist in the case of phenomenal consciousness, but in this case between physical / functional / intentional facts and the facts of bare-color. For lacking any beliefs about the processes which ground its color-recognition judgments, the creature would always be capable of thinking, 'Yet all of that could be true, and still THIS could be lacking', where 'THIS' expresses a recognitional concept grounded in a first-order analog color-content. Yet this creature lacks any phenomenally conscious color-states. Moreover, most mysterians will allow that non-phenomenal intentional states and concepts admit of reductive explanation in principle. So this is a case where a reductive explanation could exist, although the subject in question would be vulnerable to the same persistent worries which dog attempts to explain phenomenal consciousness.

(Of course it is true, as Levine notes, that WE don't feel any temptation to think that there is any problematic explanatory gap present in the case of the bare-color subject. But that only goes to help my case. The point is that THE SUBJECT in this example WOULD feel a problematic explanatory gap to exist.)

At its weakest, the example shows that the so-called 'explanatory gap' for phenomenal consciousness doesn't really have anything to do with phenomenal consciousness per se. It is, rather, a gap which can arise more generally wherever there are recognitional concepts grounded in analog intentional contents. At its strongest, the example shows that the

'explanatory gap' is not really a gap at all. That certain persistent questions remain, doesn't show that anything is going unexplained.

On the stronger of the above readings, the analogy works because it gives us a case of a recognitional concept grounded in (non-phenomenal) analog intentional states. And in the case of phenomenal consciousness (according to dispositionalist HOT theory) what we have is a set of recognitional concepts whose application is grounded in higher-order analog intentional states (intentional states whose higher-order content derives from the possibility of deploying higher-order concepts in response to them - see section 3 below). So I can grant to Levine that (phenomenal) reddishness 'doesn't have the character for me of something-I-know-not-what that happens to be prompting this response'. But then the same is true for bare-color too. For the bare-color subject doesn't find himself deploying concepts in response to something-he-knows-not-what. Rather, those concepts are grounded in perceptual (but non-phenomenal) awareness of analog properties of surfaces.

2 Inverted spectra

Levine concedes that an appeal to narrow content can undermine externalist forms of argument for the possibility of inverted experiences, such as Block's (1990) example of Inverted Earth. But he worries that I have lost sight of - and have failed to respond to - Shoemaker's original intra-personal inversion case (Shoemaker, 1981). Before I get to that, however, I want to say a word about burden of proof and plausibility.

Mysterians like Levine bring up the possibility of inverted experience as part of an argument for their mysterian position. They are therefore required to argue their case - the burden of proof falls on them to show that inverted experience is possible. And, given any robust form of concept / property distinction, they also need to show that inverted experiences are genuinely - metaphysically - possible, and not just conceptually so.

As I understand it, this is why Shoemaker saw the need to concentrate, in the first instance, on developing realistic intra-personal cases of experience inversion. This is the crucial battle-ground on which the possibility of experience inversion needs to be fought. Merely pointing to the possibility of symmetrical functional roles and symmetrical external quality spaces - as Levine does in his comments - is not going to convince anyone of anything more than a conceptual possibility. For we are given no reason to believe that experiences either would or could be reversed in such a case. It is open to reductive naturalists to claim that if the symmetrical functional roles are genuinely functionally equivalent, then experiential contents will be identical also. No reason would yet have been given to shift us from such a position.

Returning now to the charge that I had neglected Shoemaker's original argument - in fact I did return briefly to such cases following my discussion of Inverted Earth, on page 86 of my book. I claim that there are two stages necessary to liberate us from feeling any force in the intra-personal inversion examples. The first is to see the possibility of appealing to some notion of narrowly-individuated intentional content. For the sake of concreteness, let us suppose that this takes the form of a functional-role semantics. Then the second step is to see that functional roles are individuated, not just by actual causes and effects, but also by counter-factual causes and effects. (Even philosophers of psychology as distinguished as Fodor are apt to forget this.)

So, consider a post-amnesial subject in a supposed intra-personal inversion case: he uses 'seems green' to describe his experiences of green grass, and all his other beliefs and behaviors are as normal; although not long ago, before his amnesia, he could still recall that seeming-green was the experience he used to get when looking at fresh blood (before the pathways in his optic nerve were switched). Is he genuinely the functional equivalent of a normal person? (And so is this a genuine case of experiential inversion with functional / intentional equivalence?) I claim not.

Take the recognitional capacity which the subject now deploys when using the term 'seems green' and ask how that capacity WOULD HAVE responded had it been present pre-optic-nerve-switching, and had it been confronted by the state normally caused by looking at green grass - would it have been activated? Surely not. On the contrary, that very recognitional capacity would have been activated in response to the state normally caused by seeing fresh blood. So the narrow intentional state which now causes him to say 'seems

green' continues to have the content 'seeming-red' - and what we have here is no longer as case of experiential inversion with functional / intentional symmetry.

So although the intra-personal inversion subject is BEHAVIORALLY the same as a normal person, we can say that his internal states are functionally distinct, because different counter-factuals are true of them. And so we can explain, in functional-role terms, how it is that his experiences are different from normal too. And such cases therefore provide us with no reason to embrace mysterianism concerning phenomenal consciousness.

3 Dispositionalist HOT theory

Levine raises three difficulties for my dispositionalist HOT theory. I shall take them each in turn.

The first is that I have failed to make out the case for the existence of non-conscious experiences. Levine (like me) is inclined to deny that the cases I adduce involve phenomenal consciousness; but he thinks that they therefore don't deserve the title 'experience'. I am not going to quarrel about words. He can insist on calling them 'analog perceptual states' if he likes. But Levine has missed the dialectical position, here. The conscious / non-conscious distinction is NOT supposed to be part of any argument against mysterianism - so it is not supposed to be something which should worry Levine. By the time we get to chapter 6 in my book (where these matters are discussed), mysterianism has already been set to one side. Rather, some sort of representationalist approach to phenomenal consciousness is presupposed. And the conscious / non-conscious distinction forms the main premise in my arguments against first-order representational (FOR) theories, of the sorts defended by Dretske (1995) and Tye (1995, 2000).

Levine's second objection I have difficulty in getting straight. It has something to do with an alleged inadequacy in my appeal to recognitional concepts in defusing the conceivability arguments for mysterianism. The worry seems to be about whether dispositionalist HOT theory has provided for the 'substantive and determinate character' of the states to which our recognitional concepts are applied, and for the 'cognitive intimacy' which exists between our phenomenally conscious states and our recognitional concepts for them. But this is followed by a confused and inaccurate presentation of my view.

So here, briefly, is how I think the story should go. When I enjoy a phenomenally conscious experience as of red, I am in a perceptual state with the analog (or non-conceptual) content 'red', which also possesses the analog content 'seems red' or 'experience of red'. (The state in question is thus BOTH a first-order experience of color AND a higher-order experience of an experience of color.) I am then capable of enjoying purely-recognitional concepts for my phenomenally conscious experience, grounded in my higher-order experience of it. The perceptual state with the analog content 'seems red' provides the grounds for me to apply my recognitional concept, in much the same way that a perceptual state with the analog content 'red' provides the grounding for a recognitional application of the concept 'red'. These higher-order recognitional concepts are genuine concepts, capable of figuring in conceivability thought-experiments or in puzzling about the explanatory gap. And in particular, since they lack any conceptual connections with the various elements which go to make up dispositionalist HOT theory, it will always be possible to think, 'Dispositional HOT theory might be true of a creature, and still that creature could lack THESE kinds of states'.

It seems to me that this story provides fully for the substantive and determinate character of the states which ground our thoughts about phenomenal consciousness; and that the cognitive intimacy between such states and our judgments about them is also explained.

Levine's third objection to dispositionalist HOT theory is that he doesn't see why availability to higher-order concepts should transform the contents of our perceptual states, conferring them, at the same time, with higher-order analog contents. He uses the example of his watch-face to make the point. Surely the mere fact that Levine knows that his watch doesn't always keep time isn't enough to make the position of the hands represent 'seems 4 o'clock' as well as 'is 4 o'clock'.

The example is a misleading one, however. For watch-faces are not representations in their own right - they depend upon the minds of people to confer content upon them. Or if they don't - if they are considered representational - then the appropriate notion to apply must be an informational, or some sort of causal-covariance, conception of content. But this is

not the notion presupposed by consumer-semantics. According to consumer-semantics, states which are representational-in-their-own-right acquire their content from the inferential powers of the systems which consume, or make use of, those states.

I don't believe that any sort of special case has to be made out for saying that perceptual states will acquire higher-order content on becoming available to higher-order consumer systems. For this is just what consumer semantics would predict - in general, the content of a state depends upon (and is a reflection of) the inferential powers of the systems which are the immediate and prime consumers for that state; and states which become available to new systems will thereby acquire new contents.

Consider an example from developmental psychology. Perner et al. (1994) maintain that there is a stage in development when children operate with an undifferentiated concept of 'prelief', which is a sort of amalgam of 'belief' and 'pretence'. They understand that preliefs don't always correspond to the way the world is, but don't yet have a conception of a type of state which purports to represent the world as being in a particular way, but does so incorrectly - they don't yet have a conception of 'belief' as such. But as the child's theoretical and inferential powers develop, it reaches the stage at which it can treat pretence and belief as distinct states.

Here we may have a state - e.g. ascribing a prelief to someone - which acquires a new content - it becomes an ascription of belief - as a result of the new inferential powers of the theory-of-mind system which can operate upon that state. Whether or not this particular account of this stage in child development is true, this is all standard consumer-semantic stuff. But would Levine object (echoing the remarks in the penultimate paragraph of his comments), 'It is ad hoc to say that the state in question has acquired a new content'? Would he say, 'Granted there may be some difference in content, but why say that the new content is the content "belief"? Why not say that the mere difference in causal relations is the difference in content?' Such objections would be ill-motivated in the case of belief, unless they are intended as objections to consumer semantics per se. They are equally ill-motivated as objections to dispositionalist HOT theory.

4 Finally ...

One final point of clarification: Levine asks why I sometimes address questions of natural possibility: isn't metaphysical possibility always the relevant modal operator to consider? The answer is: yes, it is. But on many views of natural properties, their identity is tied to worlds in which the laws of nature remain the same. So when questions of property-identity or property-supervenience are at issue, conceivable circumstances which are naturally impossible are also very likely to be metaphysically impossible. Put differently: the relevant notion of supervenience to consider is global, where laws of nature as well as physical / functional facts are held fixed.

References

Block, N. (1990). "Inverted Earth.", *Philosophical Perspectives*, *4: Action theory and philosophy of mind*, 53-80. Ridgeview Publishing.

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Perner, J., Baker, S., & Hutton, D. (1994). "Prelief: The conceptual origins of belief and pretence." In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: origins and development* (261-286). Lawrence Erlbaum Associates.

Shoemaker, S. (1981). "The inverted spectrum." *Journal of Philosophy*, 74.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

Tye, M. (2000). *Consciousness, Color and Content*. MIT Press.

Reply to Joe Levine

Peter Carruthers

University of Maryland, College Park

My thanks to Joe Levine for his extensive and insightful comments on my book (Carruthers, 2000). In this reply I shall focus on the main points of disagreement between us - namely, the existence of an alleged 'explanatory gap'; the supposed possibility of experiential inversion; and Levine's criticisms of my proposed dispositionalist HOT theory.

1 The 'explanatory gap'

Levine concedes some force to my attempts to mark a divide between those aspects of the phenomenon of phenomenal consciousness which can be successfully reductively explained by a naturalistic theory such as dispositionalist HOT theory, and those aspects which are best explained away. A large part of the work of the latter sort is borne by the notion of a purely-recognitional concept; and Levine uses my analogy based on the example of 'bare color' as a stalking horse on which to ground his critique. Unfortunately, he gets the example wrong.

Unlike the example of the chicken-sexer (which does some, but only some, of the same work) the bare-color example isn't supposed to involve a concept which we find ourselves applying for no apparent reason, with applications of it having no 'substantive and determinate content … over and above my response itself'. On the contrary, it is supposed to be an example of a concept whose applications are grounded in first-order (but non-phenomenal) analog intentional states. We are to imagine a creature possessing analog (or non-conceptual) intentional states representing reflective properties of surfaces, and with the capacity to apply purely-recognitional concepts grounded in the intentional character of those states, but who is altogether lacking in phenomenal consciousness of color.

(Of course, whether you think the example is even so much as possible will depend upon your prior rejection of first-order representationalist theories of phenomenal consciousness. But this assumption creates no special problem as part of an argument against mysterianism. And anyone who believes in the possibility of zombies will have no difficulty with the bare-color example.)

Now, the point is that such a creature would feel the same kind of explanatory gap as is alleged to exist in the case of phenomenal consciousness, but in this case between physical / functional / intentional facts and the facts of bare-color. For lacking any beliefs about the processes which ground its color-recognition judgments, the creature would always be capable of thinking, 'Yet all of that could be true, and still THIS could be lacking', where 'THIS' expresses a recognitional concept grounded in a first-order analog color-content. Yet this creature lacks any phenomenally conscious color-states. Moreover, most mysterians will allow that non-phenomenal intentional states and concepts admit of reductive explanation in principle. So this is a case where a reductive explanation could exist, although the subject in question would be vulnerable to the same persistent worries which dog attempts to explain phenomenal consciousness.

(Of course it is true, as Levine notes, that WE don't feel any temptation to think that there is any problematic explanatory gap present in the case of the bare-color subject. But that only goes to help my case. The point is that THE SUBJECT in this example WOULD feel a problematic explanatory gap to exist.)

At its weakest, the example shows that the so-called 'explanatory gap' for phenomenal consciousness doesn't really have anything to do with phenomenal consciousness per se. It is, rather, a gap which can arise more generally wherever there are recognitional concepts grounded in analog intentional contents. At its strongest, the example shows that the

'explanatory gap' is not really a gap at all. That certain persistent questions remain, doesn't show that anything is going unexplained.

On the stronger of the above readings, the analogy works because it gives us a case of a recognitional concept grounded in (non-phenomenal) analog intentional states. And in the case of phenomenal consciousness (according to dispositionalist HOT theory) what we have is a set of recognitional concepts whose application is grounded in higher-order analog intentional states (intentional states whose higher-order content derives from the possibility of deploying higher-order concepts in response to them - see section 3 below). So I can grant to Levine that (phenomenal) reddishness 'doesn't have the character for me of something-I-know-not-what that happens to be prompting this response'. But then the same is true for bare-color too. For the bare-color subject doesn't find himself deploying concepts in response to something-he-knows-not-what. Rather, those concepts are grounded in perceptual (but non-phenomenal) awareness of analog properties of surfaces.

2 Inverted spectra

Levine concedes that an appeal to narrow content can undermine externalist forms of argument for the possibility of inverted experiences, such as Block's (1990) example of Inverted Earth. But he worries that I have lost sight of - and have failed to respond to - Shoemaker's original intra-personal inversion case (Shoemaker, 1981). Before I get to that, however, I want to say a word about burden of proof and plausibility.

Mysterians like Levine bring up the possibility of inverted experience as part of an argument for their mysterian position. They are therefore required to argue their case - the burden of proof falls on them to show that inverted experience is possible. And, given any robust form of concept / property distinction, they also need to show that inverted experiences are genuinely - metaphysically - possible, and not just conceptually so.

As I understand it, this is why Shoemaker saw the need to concentrate, in the first instance, on developing realistic intra-personal cases of experience inversion. This is the crucial battle-ground on which the possibility of experience inversion needs to be fought. Merely pointing to the possibility of symmetrical functional roles and symmetrical external quality spaces - as Levine does in his comments - is not going to convince anyone of anything more than a conceptual possibility. For we are given no reason to believe that experiences either would or could be reversed in such a case. It is open to reductive naturalists to claim that if the symmetrical functional roles are genuinely functionally equivalent, then experiential contents will be identical also. No reason would yet have been given to shift us from such a position.

Returning now to the charge that I had neglected Shoemaker's original argument - in fact I did return briefly to such cases following my discussion of Inverted Earth, on page 86 of my book. I claim that there are two stages necessary to liberate us from feeling any force in the intra-personal inversion examples. The first is to see the possibility of appealing to some notion of narrowly-individuated intentional content. For the sake of concreteness, let us suppose that this takes the form of a functional-role semantics. Then the second step is to see that functional roles are individuated, not just by actual causes and effects, but also by counter-factual causes and effects. (Even philosophers of psychology as distinguished as Fodor are apt to forget this.)

So, consider a post-amnesial subject in a supposed intra-personal inversion case: he uses 'seems green' to describe his experiences of green grass, and all his other beliefs and behaviors are as normal; although not long ago, before his amnesia, he could still recall that seeming-green was the experience he used to get when looking at fresh blood (before the pathways in his optic nerve were switched). Is he genuinely the functional equivalent of a normal person? (And so is this a genuine case of experiential inversion with functional / intentional equivalence?) I claim not.

Take the recognitional capacity which the subject now deploys when using the term 'seems green' and ask how that capacity WOULD HAVE responded had it been present pre-optic-nerve-switching, and had it been confronted by the state normally caused by looking at green grass - would it have been activated? Surely not. On the contrary, that very recognitional capacity would have been activated in response to the state normally caused by seeing fresh blood. So the narrow intentional state which now causes him to say 'seems

green' continues to have the content 'seeming-red' - and what we have here is no longer as case of experiential inversion with functional / intentional symmetry.

So although the intra-personal inversion subject is BEHAVIORALLY the same as a normal person, we can say that his internal states are functionally distinct, because different counter-factuals are true of them. And so we can explain, in functional-role terms, how it is that his experiences are different from normal too. And such cases therefore provide us with no reason to embrace mysterianism concerning phenomenal consciousness.

3 Dispositionalist HOT theory

Levine raises three difficulties for my dispositionalist HOT theory. I shall take them each in turn.

The first is that I have failed to make out the case for the existence of non-conscious experiences. Levine (like me) is inclined to deny that the cases I adduce involve phenomenal consciousness; but he thinks that they therefore don't deserve the title 'experience'. I am not going to quarrel about words. He can insist on calling them 'analog perceptual states' if he likes. But Levine has missed the dialectical position, here. The conscious / non-conscious distinction is NOT supposed to be part of any argument against mysterianism - so it is not supposed to be something which should worry Levine. By the time we get to chapter 6 in my book (where these matters are discussed), mysterianism has already been set to one side. Rather, some sort of representationalist approach to phenomenal consciousness is presupposed. And the conscious / non-conscious distinction forms the main premise in my arguments against first-order representational (FOR) theories, of the sorts defended by Dretske (1995) and Tye (1995, 2000).

Levine's second objection I have difficulty in getting straight. It has something to do with an alleged inadequacy in my appeal to recognitional concepts in defusing the conceivability arguments for mysterianism. The worry seems to be about whether dispositionalist HOT theory has provided for the 'substantive and determinate character' of the states to which our recognitional concepts are applied, and for the 'cognitive intimacy' which exists between our phenomenally conscious states and our recognitional concepts for them. But this is followed by a confused and inaccurate presentation of my view.

So here, briefly, is how I think the story should go. When I enjoy a phenomenally conscious experience as of red, I am in a perceptual state with the analog (or non-conceptual) content 'red', which also possesses the analog content 'seems red' or 'experience of red'. (The state in question is thus BOTH a first-order experience of color AND a higher-order experience of an experience of color.) I am then capable of enjoying purely-recognitional concepts for my phenomenally conscious experience, grounded in my higher-order experience of it. The perceptual state with the analog content 'seems red' provides the grounds for me to apply my recognitional concept, in much the same way that a perceptual state with the analog content 'red' provides the grounding for a recognitional application of the concept 'red'. These higher-order recognitional concepts are genuine concepts, capable of figuring in conceivability thought-experiments or in puzzling about the explanatory gap. And in particular, since they lack any conceptual connections with the various elements which go to make up dispositionalist HOT theory, it will always be possible to think, 'Dispositional HOT theory might be true of a creature, and still that creature could lack THESE kinds of states'.

It seems to me that this story provides fully for the substantive and determinate character of the states which ground our thoughts about phenomenal consciousness; and that the cognitive intimacy between such states and our judgments about them is also explained.

Levine's third objection to dispositionalist HOT theory is that he doesn't see why availability to higher-order concepts should transform the contents of our perceptual states, conferring them, at the same time, with higher-order analog contents. He uses the example of his watch-face to make the point. Surely the mere fact that Levine knows that his watch doesn't always keep time isn't enough to make the position of the hands represent 'seems 4 o'clock' as well as 'is 4 o'clock'.

The example is a misleading one, however. For watch-faces are not representations in their own right - they depend upon the minds of people to confer content upon them. Or if they don't - if they are considered representational - then the appropriate notion to apply must be an informational, or some sort of causal-covariance, conception of content. But this is

not the notion presupposed by consumer-semantics. According to consumer-semantics, states which are representational-in-their-own-right acquire their content from the inferential powers of the systems which consume, or make use of, those states.

I don't believe that any sort of special case has to be made out for saying that perceptual states will acquire higher-order content on becoming available to higher-order consumer systems. For this is just what consumer semantics would predict - in general, the content of a state depends upon (and is a reflection of) the inferential powers of the systems which are the immediate and prime consumers for that state; and states which become available to new systems will thereby acquire new contents.

Consider an example from developmental psychology. Perner et al. (1994) maintain that there is a stage in development when children operate with an undifferentiated concept of 'prelief', which is a sort of amalgam of 'belief' and 'pretence'. They understand that preliefs don't always correspond to the way the world is, but don't yet have a conception of a type of state which purports to represent the world as being in a particular way, but does so incorrectly - they don't yet have a conception of 'belief' as such. But as the child's theoretical and inferential powers develop, it reaches the stage at which it can treat pretence and belief as distinct states.

Here we may have a state - e.g. ascribing a prelief to someone - which acquires a new content - it becomes an ascription of belief - as a result of the new inferential powers of the theory-of-mind system which can operate upon that state. Whether or not this particular account of this stage in child development is true, this is all standard consumer-semantic stuff. But would Levine object (echoing the remarks in the penultimate paragraph of his comments), 'It is ad hoc to say that the state in question has acquired a new content'? Would he say, 'Granted there may be some difference in content, but why say that the new content is the content "belief"? Why not say that the mere difference in causal relations is the difference in content?' Such objections would be ill-motivated in the case of belief, unless they are intended as objections to consumer semantics per se. They are equally ill-motivated as objections to dispositionalist HOT theory.

4 Finally ...

One final point of clarification: Levine asks why I sometimes address questions of natural possibility: isn't metaphysical possibility always the relevant modal operator to consider? The answer is: yes, it is. But on many views of natural properties, their identity is tied to worlds in which the laws of nature remain the same. So when questions of property-identity or property-supervenience are at issue, conceivable circumstances which are naturally impossible are also very likely to be metaphysically impossible. Put differently: the relevant notion of supervenience to consider is global, where laws of nature as well as physical / functional facts are held fixed.

References

Block, N. (1990). "Inverted Earth.", *Philosophical Perspectives*, *4: Action theory and philosophy of mind*, 53-80. Ridgeview Publishing.

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Perner, J., Baker, S., & Hutton, D. (1994). "Prelief: The conceptual origins of belief and pretence." In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: origins and development* (261-286). Lawrence Erlbaum Associates.

Shoemaker, S. (1981). "The inverted spectrum." *Journal of Philosophy*, 74.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

Tye, M. (2000). *Consciousness, Color and Content*. MIT Press.

Dispositions and Consciousness

Commentary on Carruthers, Peter. *Phenomenal Consciousness*. Cambridge: Cambridge University Press, 2000.

William Seager

Department of Philosophy
University of Toronto at Scarborough, Ontario (Canada).
Homepage

Though my job and my aim is to criticize certain aspects of Peter Carruthers's theory of consciousness as presented in his *Phenomenal Consciousness* (2000), I want to begin by enthusing about Carruthers's book and the philosophical advance of which it is a part. One of the most important developments in the philosophy of mind over the last fifteen years or so is the rise of the representational theories of consciousness. Of course, some of the core ideas in this tradition can be traced back further than fifteen years! Notions of the essential "reflexivity" of conscious thought can be found in Kant, who perhaps took the idea from Leibniz, who clearly prefigures the representational account of consciousness where he says: "... it is well to make the distinction between perception which is the internal state of the monad representing external things, and apperception, which is consciousness, or the reflective knowledge of this internal state" (1714/1989, p. 208). A somewhat less historical figure is David Armstrong, whose account of consciousness as a kind of "self-monitoring" inner perception in *A Materialist Theory of the Mind* (1968) at least anticipates the modern representational theories of consciousness. Armstrong, almost echoing Leibniz, wrote: "consciousness is simply a further mental state, a state "directed" towards the original inner states" (p. 94). Significant developments in the representational account of consciousness that followed Armstrong include David Rosenthal's (1986) "Two Concepts of Consciousness" (Rosenthal has continued to develop and extend his theory in a number of influential papers), William Lycan's (1987) *Consciousness* (Lycan's theory is further developed in his 1996), and Gilbert Harman's "The Intrinsic Quality of Experience" (1990).

In 1995 a most remarkable event occurred: the publication of two books, written independently of each other, which espoused remarkably similar views on consciousness. There were Fred Dretske's *Naturalizing the Mind* and Michael Tye's *Ten Problems of Consciousness*. Both books argued that consciousness could be "reduced" to representation (in a certain sense) and that in particular the qualitative character of experience (or what philosophers call qualia) was itself a matter of representation. Their theories maintain that to experience red consciously is to represent a certain part of the environment (or imagined environment) as having a certain property; to experience pain is to represent a part of one's body (perhaps - in phantom limb pain for example - an "imaginary" part) as having a rather disagreeable kind of property. The fact that Tye and Dretske came to so many of the same conclusions once they had set the basics of their views seems to indicate that the representational theory of consciousness has a lot of content which is constraining its final form. Which is good, and rather unusual in philosophical constructions.

The work of Dretske and Tye, as my examples illustrate, made it clear there was a fundamental distinction within representational theories of consciousness: that between what Carruthers calls first order representation (FOR) based and higher order representation (HOR) based theories. Whereas Dretske and Tye opted for FOR theories, Rosenthal and Armstrong (not to mention Leibniz) opted for HOR theories. The debate between proponents of FOR and HOR theories is one of the most interesting and important issues in current philosophy of mind.

Both approaches have many virtues. Abstractly speaking, they promise to open a route towards the naturalization of mind and consciousness (thus perhaps solving the so-called hard problem), although this promise is dependent on the distinctly non-trivial task of providing an adequate naturalistic theory of representation and/or intentionality. Less abstractly, it seems these theories will also be able to integrate smoothly into the mainstream of cognitive science with its long standing emphasis on representation and the

processing of representations. In addition to providing a strikingly novel approach to consciousness itself, the representational theories are a wellspring of potential insights into the nature of the mind with interesting, if sometimes contentious, things to say, for example, about introspection and its relation to consciousness, emotional consciousness, animal consciousness, and the role of the "theory-theory" account of our commonsense understanding of mind (here we see another interesting and more concrete integration with science, this time psychology).

Carruthers's book is a marvelous and wide-ranging critical introduction to the problem of consciousness from a representational theory point of view. One of the main reasons it provides such an incisive review of current views about consciousness is that Carruthers also develops and defends a highly original version of the representational theory of consciousness which boasts a remarkable degree of depth and precision. The book is so rich that I will be able to comment on only a few points within it and the fact that my discussion will be critical indicates only that I want to focus on a central theme about which I have some worries.

I think we can outline Carruthers's position, briefly and crudely, as follows. First, Carruthers prefers a HOR to a FOR theory of consciousness. The basic reason for this preference is that the FOR theory lacks the resources to distinguish, in a principled way that springs naturally from the underlying theory, conscious from non-conscious representations. This is a serious problem since given the kind of cognitive science inspired background theories of the mind which most FOR theorists would endorse there are inevitably going to be huge numbers of representations loose and active within any cognitive system, only a tiny fraction of which, presumably, are conscious. HOR theories avoid this problem by explicitly providing a mechanism by which only some representations become conscious, namely by being "taken up" into a higher order representation. This mechanism is well able to distinguish conscious from non-conscious representations without impairing the cognitive function of the non-conscious representations.

The HOR theories can also be divided up in various ways. One distinction is between higher order experience (HOE) theories and higher order thought (HOT) theories. Carruthers prefers the latter. Roughly speaking the reason is that our consciousness of the world is of the world rather than of the mind, or, as it is sometimes put, consciousness is transparent; we see right through our experience to the world beyond, so to speak. This is odd if our account of consciousness is based on our having quasi-perceptual experiences of inner states, for we might expect then that these experiences then have their own set of perceptible (to inner sense now) phenomenological qualities and, in all honesty, there is just no trace of any such qualities (or else in perception we are completely cut off from the world, never "observing" anything but our own sense data). Higher order thought theories avoid this problem simply because thoughts do not come with any intrinsic phenomenology but they can be about any kind of phenomenlogical state, or indeed any kind of state, whatsoever.

We need one more distinction to situate Carruthers's view, which is the distinction between actualist and dispositionalist theories. The actualist HOT theories define consciousness in terms of the actual occurrence of an appropriate higher order thought about a lower order state so that only those states are conscious which in fact bring about an occurrent higher order thought about them; the dispositionalist HOT theories require less for a lower order state to be a conscious state, namely that the lower order states be disposed to produce such higher order thoughts. But there is no requirement that they exercise their ability. I believe that the dispositional higher order thought theory of consciousness is original to Carruthers. It might seem odd to analyse the vivid, totally occurrent phenomenon of consciousness in terms of mere dispositions, but there are advantages. For example, the dispositionalist account offers - literally in this instance - a kind of economy in that it reduces the number of actual mental states which the brain has to produce and Carruthers uses that in an argument to the effect that evolution would be unlikely to build a brain that is given to the promiscuous proliferation of higher order thoughts for every aspect of our apparently exceedingly rich and varied conscious experience.

In any case, I don't want to consider the arguments which Carruthers advances against actualist HOT theory but instead raise an objection against two versions of the dispositionalist account, both of which have been advanced by Carruthers. The first is the reflexive account which first appeared in Carruthers's Language, Thought and Consciousness (1996) and from which Carruthers steps back in Phenomenal Consciousness. However, the

present retreat is merely tactical so far as I can see and in fact Carruthers goes so far as to say that the reflexive account, which is described in some detail, may be an accurate description of the mind as it actually is (see pp. 271 ff.) It is just that the full featured reflexive account is not needed for the "restricted purpose" of the present volume - I have to put that in scare quotes seeing that this restricted purpose is nothing less than that of explaining consciousness. The difference between the old reflexive account and the non-reflexive account actually endorsed by Carruthers now depends upon yet another distinction. Whether or not one is a dispositionalist HOT theorist, one can wonder about the status, with respect to their consciousness, of the higher order thoughts. These thoughts can themselves either possess or not possess the disposition to produce yet higher order thoughts. The reflexive account asserts that the higher order thoughts do possess these dispositions; that is, it makes the higher order thoughts themselves conscious thoughts.

Now, we might pause here to ask whether it is so much as even possible for actualist HOT theories of consciousness to require that the higher order thought, call it T[x], which makes the lower order state x conscious, be itself a conscious mental state. The answer is apparently an obvious "no" since such a requirement would generate a vicious infinite regress of nested conscious states. Not only is it the case that it is phenomenologically plain that when I am conscious of some mental state, x, I am not also conscious of each of an infinite hierarchy of states T[x], T[T[x]], ..., T[...T[x]...], etc. but there must also be neurologically founded limitations on the number and complexity of thoughts that any of us can actually entertain at one time. But on the other hand, HOT theory cannot rule out the possibility of any particular higher order thought's being conscious since, generally speaking, it is certainly possible to become consciously aware of any particular lower order thought. Actualist HOT theory has no difficulty here. One can become conscious of the higher order thought, T[x], that makes x a conscious state if T[x] should happen to bring about the still higher order thought T[T[x]], but there is no requirement that this thought should occur to one in order for x to be a conscious state. Since it seems evident that we are often conscious without being conscious of being conscious but that sometimes we do enjoy such higher order consciousness, this would appear to be a required feature of any theory of consciousness.

Although the actualist account cannot allow that the consciousness conferring higher order states are themselves conscious states, it seems clear that the dispositionalist HOT theory can avoid the regress without denying consciousness to the higher order thoughts. So if such reflexivity was indeed a feature of consciousness this would be an significant advantage for a dispositionalist account. I doubt that such reflexivity could be a feature of consciousness however, just because only dispositional HOT theory can accommodate it. For, I suspect, the dispositional reflexive account suffers from a fatal flaw itself which is closely related to the regress problem which dooms an actualist reflexive theory. The objection is very simple.

Let us consider whether there is a limit, imposed by the finiteness and particularity of our cognitive architecture, on the complexity of nested thoughts we can entertain. It seems as certain that there are thoughts of the form "I am aware that I am aware that I am aware .... that P" which are sufficiently deeply nested as to be in fact entirely incomprehensible, given normal human cognitive capacities, as that there are numbers which are too big for my calculator to multiply. I suspect that this limit is in fact quite low - at least for me - as I have a good deal of difficulty being aware of just a few levels of awareness. This psychological difficulty or weakness, which I think everyone will readily agree is real, in fact shows that higher order thoughts are indeed more complex - require more mental machinery to entertain - than the lower order thoughts they are about.

Suppose, then and without loss of generality, that the level of nesting at which nested thoughts become unentertainable is n. Then it is easy to show that no thought of level n-1 could be conscious. For if it is impossible to entertain, because of inherent cognitive limitations, a nested thought of level n, it cannot be the case that any thought is apt to cause a nested thought of level n. (Any more than a certain kind of ill-made match could be disposed to light if struck if, because of inherent chemical conditions, no striking could raise the temperature sufficient to ignite it.) Obviously, if a level n-1 thought cannot cause a level n thought it cannot cause a level n conscious thought, and so, according to the dispositional conscious, or reflexive, HOT theory, the level n-1 thought cannot be conscious. But if no level n-1 thought could be conscious then no thought of level n-2 could be disposed to produce a conscious level n-1 thought (even if it might be disposed to produce a level n-1 thought). Hence, no level n-2 thought could be a conscious thought. This

argument by "vicious descent" can clearly be generalized as far as necessary, with the disastrous result that no mental state can be conscious according to the reflexive HOT theory.

Biting, instead of dodging, the bullet, it might perhaps be replied that there is no level of nested thought which is impossible to entertain. It is true that the concept of thought, and that of the entertaining of thoughts, admits of no intrinsic, purely abstract, limitation in complexity. But the point here is that there is a natural limitation, imposed by the cognitive architecture implemented by the finite brain, to the complexity of entertainable thoughts. This limitation is based upon natural laws which reveal to us the range of possible dispositions which our neurological machinery can instantiate. The dispositional conscious HOT theory cannot avail itself of the mere abstract structural possibilities of thought, since it depends upon the actual dispositions inherent in the brains or minds we actually possess. The distinction between the abstract structure of thought and our actual dispositions to entertain thoughts appealed to here is of course reminiscent of Chomsky's between performance and competence. While competence - the abstract structural possibilities of language - is not limited by natural constraints, it would be ludicrous to claim that there was an actual human disposition either to produce or to understand sentences with, say, 1037 nested relative clauses.

I think the only conclusion that can be drawn is that the dispositional conscious HOT theory, or the reflexive theory, cannot be correct (at least for finite, real-world cognitive systems).

But even if this objection is granted, it does not touch the non-reflexive account which Carruthers officially advances in Phenomenal Consciousness. This theory is immune to the vicious descent argument which yields only the "theorem" that there is a level of nested thought of which we cannot be conscious (which is one level below the level at which we can no longer entertain thoughts of that complexity at all). Such cognitive limitations on consciousness seem entirely acceptable, not unexpected and in fact phenomenologically highly plausible.

Still, there is a somewhat more subtle difficulty which can be raised against the non-reflexive account and which I think is at least an interesting challenge to the dispositionalist approach in general and perhaps threatens to undercut it entirely.

In actualist HOT theory, if we can prevent the higher order thoughts from occurring we can prevent the lower order states from being conscious. Thus, for example, if we had some kind of machine, call it a "neural meddler", that interfered with the causal mechanisms which normally enable the lower order state to produce the higher order thought which makes the former conscious then we would have a "consciousness inhibitor". Of course, such mechanisms exist on either the actualist or dispositionalist versions of HOT theory. For example, Carruthers imagines (see pp. 228 ff.) that contents are sent from perceptual systems to a special short-term memory buffer which has access under certain conditions to the systems which realize conceptual thinking. This access is limited to the direct, non-inferential production of higher order thoughts about the current contents of this buffer. It is important that the production of the higher order thoughts be non-inferential to avoid strong objections based upon coming to know that one is in a lower order state in ways that intuitively do not lead to a state of consciousness (for example, if a reliable source simply informs one that one is in a certain mental state one will come to think a higher order thought about the original state but not in a way that seems to entail that the original state is a conscious state). But this issue does not matter for what follows here.

In the case of the dispositional HOT theories, the operation of a neural meddler is slightly less straightforward than in actualist theories. Part of the reason for this has to do with the nature of dispositions. Consider an ordinary match. It has a disposition to light if struck. Does it have this disposition in a vacuum, where it cannot light no matter whether or not it is struck? I'm not sure if there is a definite answer to this question, but I feel sure that if we meddled with the match itself - as opposed to altering the external circumstances - we could destroy the disposition. If, for example, we actually struck and lit the match and then blew it out we would thereby have a match that definitely lacked the disposition. More subtle methods of "disabling" the disposition are easily imagined - such as coating the match with some kind of irremovable wax. Similarly, if we meddle with someone's brain so that the lower order states are made incapable of causing the appropriate higher order states we have, as the phrase "incapable of causing" suggests, eliminated the disposition to cause higher order states. In the kind of "boxology" envisioned by Carruthers there are

doubtless very complex neural processes mediating the relation between lower order states and the higher order thoughts which make the latter conscious states, and they could presumably be interfered with in a multitude of ways (some of which might even be invoked to account for certain of the bizarre deficits of consciousness, such as Capgras syndrome, blindsight, etc.). Thus a neural meddler would eliminate consciousness under the dictates of the dispositional HOT theory, in much the same way that consciousness would be eliminated by the prevention of the occurrence of higher order states under the dictates of the actualist HOT theories. Another way to put this point, in terms that Carruthers favours, is that the neural meddler would prevent the lower order states from being available to consciousness, and without this availability there can be no consciousness of those states (on either the dispositional or actualist theories).

But if this is so, dispositional HOT theories face a serious objection. To advance it I have to enter the realm of pure philosophical thought experiment but I think in a legitimate fashion. Consider, now, a modified neural meddler which blocks the disposition to cause higher order states only for those states and for those time periods when the lower order states would not, in fact, cause a higher order thought. (To continue the analogy with the match, we can imagine a device that somehow only coats with wax matches that are not actually going to be struck.) Such a meddler would be extremely difficult to produce in practice (not that the original meddler is exactly "off the shelf" machinery just yet!) since it requires an ability to know under what conditions a lower order thought will occur but will not actually cause a higher order thought. If we suppose that it is in principle possible to predict the operation of a brain at the neural level, then the information from such predictions could be fed into the meddler so that it would be active only for those lower order states which possess the disposition to bring about higher order thoughts about themselves but that are such that during the relevant time period are in fact not going to exercise this disposition and cause the higher order thought. Of course, the practical difficulty of developing such a meddler is irrelevant to the point of principle at issue here.

Now, a curious consequence follows. Let us take two people, the first with one of these modified neural meddlers implanted into to his or her brain and the second without, but who otherwise begin in identical neurological states (and in identical environments). Both of these people will have an identical history of higher order thoughts, since the meddler will never prevent a lower order state that actually was going to produce a higher order thought from causing that thought. They will also have identical histories of lower order mental states, for the meddler has no effect on these. Yet they will be markedly different in their states of consciousness, for the unfortunate person with the meddler will lack an entire set of conscious states enjoyed by the other - namely those that as a matter of fact do not produce any higher order thoughts (but - in the unmeddled brain - could have). This is a necessary consequence of the dispositional HOT theory, since it is explicitly designed to allow that states are conscious simply if they are able to produce higher order thoughts, not if they actually do produce those thoughts.
This consequence of dispositional HOT theory is not only curious, it is disturbing and highly implausible. There is absolutely no difference in the behaviour of our two individuals, no difference in their history of mental states and no difference in the neural events and processes which are occurring within them. There is nothing to mark the difference between the two of them except an entirely inert meddler. The meddler never has to actually function to produce this drastic alteration in consciousness. That is, two brains identical in their neural states and their dynamics will differ in consciousness solely because one has an inert piece of machinery within it! No such implausibility follows from occurrent HOT theory.

In a way, this objection is appealing to the feature of consciousness which I mentioned above. Consciousness is the example of something which is happening, which is, so to speak, totally occurrent. It seems impossible that in the face of the identical nature of the neural processes which are occurring in our two systems that the existence of consciousness could depend upon the purely counterfactual possibilities of differences in the system's behavioural capacities.

Perhaps the implausibility can be underlined if we imagine that the modified meddler is oscillating between being "off" - incapable of functioning - and "on" - capable of functioning, even though it never will. There will be a corresponding oscillation in consciousness (more conscious states when the meddler is disabled, fewer when it is enabled) which would presumably be very striking but in fact would be seemingly be completely unreportable by the subject despite being a huge difference in phenomenological experience.

There is a kind of "inverted" version of this objection. Consider a device that increases or boosts the aptitude of a lower order mental state to produce higher order thoughts (call it a boosting meddler - in terms of our match analogy, a boosting meddler might be something that modifies the match so that it will ignite at a lower temperature, thus increasing its aptitude to light when struck). Under actualist HOT theory, such a device would increase the number of conscious states inasmuch as it would increase the number of higher order thoughts that are actually brought about by lower order states. This effect would be apparent under dispositional HOT theory as well. But, as before, dispositional HOT theory permits a more subtle tampering with consciousness. There must be a distinction between those lower order states which do and those which do not possess the disposition to bring about higher order thoughts and so there must be some neurological basis for this difference. Thus we can imagine implanting a modified boosting meddler, which creates the aptitude to cause higher order states in lower order states which (1) would not otherwise have the disposition to bring about higher order states and which (2) even with the boosting meddler in place will never in fact exercise this disposition and actually cause a higher order thought. Again, this would require remarkable knowledge (and fore-knowledge) of how a particular neural system is going to work, but we can grant this knowledge in principle . (The analogous device in the match example would only modify those matches that are in fact not going to be struck. So the modified boosting meddler never makes any difference to which matches actually light or do not light.) So, even though there is absolutely no increase in the number of higher order thoughts, there is a striking increase in consciousness. Perhaps this result is less implausible than the previous one, insofar as in this case the meddler actually does some work, but it remains very implausible.

One reply that might be made on behalf of the dispositional HOT theory is to invent a distinction between dispositions so that only some of them are such as to yield consciousness. This would be a dangerous move however, since it would open this HOT theory to the same objection raised against FOR theories (such as Dretske's or Tye's), namely, to provide a principled distinction between those states which are and those which are not conscious. If the dispositional HOT theorist can say: conscious states are those which are properly disposed to produce higher order thoughts about them, then it seems the FOR theorists can with equal justice say that those representations are conscious which are of the proper form and function.

Perhaps there is a simple reply to these lines of objections which reveals a misunderstanding of the theory on my part, but they seem to me a serious foundational worry about the basic structure of the view. Carruthers's theory even so remains an intriguing and original addition to the growing range of representational theories of consciousness, which is the most exciting area in consciousness studies at the moment. Carruthers's book is now an essential part of the literature of this area.

References

Armstrong, David (1968). *A Materialist Theory of the Mind*, London: Routledge and Kegan Paul.

Carruthers, Peter (1996). *Language, Thought and Consciousness*, Cambridge: Cambridge University Press.

Carruthers, Peter (2000). *Phenomenal Consciousness*, Cambridge: Cambridge University Press.

Dretske, Fred (1995). *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Harman, Gilbert (1990). "The Intrinsic Quality of Experience" in J. Tomberlin (ed.) *Philosophical Perspectives*, vol. 4, pp. 31-52.

Leibniz, G. W. (1714/1989). "Principles of Nature and Grace", in R. Ariew and D. Garber (trans. ed.) *G. W. Leibniz: Philosophical Essays*, Indianapolis: Hackett.

Lycan, William (1987). *Consciousness,* Cambridge, MA: MIT Press.

Lycan, William (1996). *Consciousness and Experience*, Cambridge, MA: MIT Press.

Rosenthal, David (1986). "Two Concepts of Consciousness," *Philosophical Studies*, 49, pp. 329-59.

Tye, Michael (1995). *Ten Problems of Consciousness*, Cambridge, MA: MIT Press.

Reply to Seager

Peter Carruthers

Department of Philosophy
University of Maryland, College Park.

William Seager gives a generous assessment of my recent book (2000), for which I am grateful. But he also develops an alleged counter-example to my dispositionalist higher-order thought (HOT) theory of phenomenal consciousness. I shall concentrate on this in my response, commenting briefly on his criticisms of my earlier 'reflexive thinking' account at the end.

Seager develops his counter-example in stages. First, he envisages what he calls a 'neural meddler', which would interfere in someone's brain-processes in such a way as to block the availability of first-order perceptual contents to higher-order thought. Second, he imagines that our understanding of neural processing in the brain has advanced to such an extent that it is possible to predict in advance which first-order perceptual contents will actually become targeted by higher-order thought, and which will not. Then third, he supposes that the neural meddler might be so arranged that it only blocks the availability to higher-order thought of those perceptual contents which are not actually going to give rise to such thought anyway.

The upshot is that we can envisage two people -- Bill and Peter, say -- one of whom has such a modified neural meddler in his brain and the other of whom does not. They can be neurological 'twins' and enjoy identical neural histories, as well as undergoing identical sequences of first-order perceptual contents and actual higher-order thoughts. But because many of Bill's first-order percepts are unavailable to higher-order thought (blocked by the neural meddler, which actually remains inoperative), whereas the corresponding percepts in Peter's case remain so available, there will then be large differences between them in respect of phenomenal consciousness. Or so, at least, dispositional HOT theory is supposed to entail. And this is held to be highly counter-intuitive.

I have a number of responses to this example, which will be presented in order of increasing concessiveness.

1 Imaginary examples

The first response is to reject the example as being merely imaginary, and hence as irrelevant to the assessment of a naturalistic reductive explanation. Dispositionalist HOT theory is not, of course, intended as a conceptual analysis of our common-sense notion of phenomenal consciousness, nor as any sort of explication of our folk-theory of such consciousness. It is, rather, intended as a reductive explanation of phenomenal consciousness, parallel in kind to the reductive explanation of (one form of) heat in terms of mean molecular momentum. Such explanations are not intended to hold good in all conceivable circumstances. Rather, they are intended to provide us with an account of the actual constitution of the property being explained, which will be valid only across those possible worlds where the laws of nature remain constant (i.e. remaining the same as in the actual world).

Now, who knows whether it is naturally possible to meddle with the availability of first-order perceptual contents to higher-order thought without also having an impact on the nature of those contents, or on the higher-order thoughts entertained, or both? And who knows whether neural processing is well enough behaved to be predictable, even in principle? (It may turn out to be indeterministic; or it may turn out to be chaotic, with large differences resulting from undetectably-small differences in initial conditions.) In so far as we don't know these things, we don't know whether Seager's example is naturally possible,

either. And if it isn't naturally possible, then it isn't the right kind of case to provide a counter-example to dispositionalist HOT theory.

Stronger still, we have at least some reason to suspect that Seager's example is not naturally possible. In order for Bill and Peter to remain neural duplicates while we predict in complete detail the higher-order thoughts which Bill will entertain, and while the readiness-potential of the neural meddler is adjusted accordingly, it wouldn't be enough just to control and match their respective environments. We would also have to predict how their relations with those environments will change in response to their own actions (e.g. by turning their eyes, or by moving across the room). And we will also have to predict the sequences of thought which will occur to them, presumably influenced by all kinds of detail in their past experiences and in their background beliefs, as well as by current perceptual input. So there may actually be the same sort of deep impossibility here which Dennett (1991) argues to exist for familiar 'brain-in-a-vat' scenarios.

Moreover, all these predictions will have to be made while controlling for, and predicting, any effects which the changing state of the neural meddler might have on the rest of the brain -- changes which will take place in response to the very predictions we are presently trying to make. So there may be the same kind of principled difficulty in making accurate predictions here as occurs also in the social sciences, where the social effects of any prediction we make (effects which are partly dependent on our own future decisions) are amongst the things which we have to try to take into account when making a prediction (MacIntyre, 1981).

2 Unreliable intuitions

Suppose we allow that Seager's example is naturally possible, however. Still, on any account, the example is a fantastic one, requiring us to consider a situation lying wholly outside the orbit of our ordinary thoughts and beliefs. And there is good reason to think that once we leave that orbit our common-sense intuitions cease to have any reliability whatever. (Remembering, of course, that these intuitions are merely beliefs about the circumstances in which phenomenal consciousness would or would not be present, arrived at in the absence of any well-established theory of its underlying nature. We are doing science here, not semantics.)

Travel back in time just twenty years, and you will find that most people have a powerful intuition that visual perception is an intrinsically-conscious process, for example. Yet now, after the discovery of such phenomena as blindsight (Weiskrantz, 1986), and following the development of a two-systems theory of vision (Milner and Goodale, 1995), there are few educated people who would be prepared to make such a claim. Similarly, there are many people who continue to have the intuition that the unwelcomeness of a sensation of pain is intrinsic to its very nature. But in fact there is good evidence that the felt quality distinctive of pain sensations, on the one hand, and the awfulness of pain, on the other, can dissociate from one another in rare cases (Dennett, 1978; Ramachandran and Blakeslee, 1998), for physiological reasons which are now quite well understood (Young, 1986).

Our intuitions are, of course, grounded in the familiar and every-day. They reflect the beliefs which we have formed from experience and/or testimony concerning usual cases. Now we are asked to consider a highly unusual case, whose structure falls well beyond the range of our ordinary experience and belief. Our background beliefs, formed in the context of the ordinary, may well combine in such a way as to deliver an implication in respect of the unusual case we are imagining. This implication will have the force of an intuition. When we imagine the unusual case, and ask ourselves what we are inclined to think about it, our minds may well throw up an answer for us. But there is no reason at all to think that this answer will have been reliably generated, nor that it is likely to be true. The scientific truth about highly unusual cases needs to be settled by careful theoretical enquiry, not by appeals to intuition.

3 Kinds of disposition

Seager thinks that his neural meddler, by blocking access between first-order percepts and higher-order thought, will make it the case that the former are no longer disposed to give rise to the latter -- and so will no longer be phenomenally conscious according to a dispositionalist HOT account. He may be right. But as Seager is aware, there a number of different kinds of disposition. I think it is an open question which kind is involved here.

Suppose we take a green emerald and lock it away in an indestructible light-excluding box. Now (in one sense) the emerald is no longer disposed to reflect green light. The surrounding box has blocked the exercise of that capacity, in something like the way in which the neural meddler blocks the exercise of the capacity to give rise to higher-order thought. But is the emerald still green, as it sits there in the box? Surely so. The fact of being locked permanently in the box hasn't altered or done away with its color. Yet color, of course, is a dispositional property. So here is a case where the blocking of the exercise of a dispositional property leaves that property present and intact.

Perhaps the same is true in connection with the neural meddler. Although it blocks the access-relation between perception and higher-order thought, it may be that the disposition of the former to give rise to the latter remains present and intact. In which case Bill and Peter can be entirely alike in respect of phenomenal consciousness, according to a dispositionalist HOT account. Whether or not this is so will turn, I think, on as-yet-unresolved issues in semantic theory, as I shall discuss next.

4 The incompleteness of consumer semantics

What Seager doesn't mention in his discussion of dispositionalist HOT theory, is that much of the crucial work is done by appeal to some or other version of consumer semantics. The thought is this: if intentional content depends, in part, on the powers of the 'consumer systems' which use or draw inferences from it, then first-order analog perceptual contents can be rendered AT THE SAME TIME as higher-order, by virtue of their availability to a HOT-wielding consumer system. Each experience with the analog content 'red', for example, will at the same time have the analog content 'seems red' or 'experience of red'. And it is this which confers on those experiences -- categorically -- a dimension of 'seeming' or 'subjectivity', and constitutes them as phenomenally conscious.

Now, consumer semantics embraces a number of different varieties of theory, of which there are two main sorts -- teleosemantics, on the one hand, and various forms of inferential role semantics, on the other. (For the former, see Millikan, 1984, 1986, 1989; and Papineau, 1987, 1993. For the latter, see Loar, 1981, 1982; McGinn, 1982, 1989; Block, 1986; and Peacocke, 1986, 1992.) In the present state of scientific knowledge, it is quite unclear which of these will ultimately prove to be the victor; and their respective implications for Seager's example may not be clear either.

Actually, it does appear that on a teleosemantic account the higher-order analog content of Bill's experiences will be left unaffected, in Seager's example. Teleosemantics claims that intentional content is to be explicated in terms of biological and/or derived proper function. And if we ask what the proper function is, of those of Bill's experiences which are blocked from giving rise to higher-order thoughts by the presence of the neural meddler, the answer will be: unchanged. These states still have the proper function of feeding into thoughts about the perceived environment, and also of feeding into higher-order thoughts about themselves, despite the presence of the neural meddler. And so they will continue to have the very same first-order and higher-order analog contents which they would possess if the neural meddler weren't there. Which means, on my dispositionalist HOT account, that they remain phenomenally conscious, despite their de facto inaccessibility to higher-order thought.

In my 2000 I operated within the framework of an inferential-role version of consumer semantics, as opposed to a teleosemantic one -- for convenience only, I thought. But there are a number of issues on which the two approaches may actually generate different answers. As recent work by Sarah Clegg (2001) has brought home to me, for example, a teleosemantic version of dispositionalist HOT theory may imply that the experiences of even new-born infants are phenomenally conscious, despite the infants' lack of capacity to entertain higher-order thoughts. For their experiences may still have the FUNCTION of feeding into higher-order thought, and will become available to such thought in the course of normal development. In which case those percepts may already possess higher-order analog content, on a teleosemantic account, even though the infants are not presently disposed to draw higher-order inferences from those experiences.)

What should an inferential-role semanticist say about Seager's example? Does a device which blocks the normal inferential role of a state thereby deprive that state of its intentional content? The answer to this question is not clear. Consider a propositional variant of Seager's example: we invent a 'conjunctive-inference neural meddler'. This

device can disable a subject's capacity to draw basic inferences from a conjunctive belief. Subjects in whom this device is operative will no longer be disposed to deduce either 'P' or 'Q' from beliefs of the form 'P & Q' -- those inferences will be blocked. (Even more elaborately, in line with Seager's example, we could imagine that the device only becomes operative in those cases where we can predict that neither the inference 'P' nor the inference 'Q' will actually be drawn anyway.)

Supposing, then, that some version of the Mentalese hypothesis is true, are the belief-like states which have the syntactic form 'P & Q' in these cases conjunctive ones? Do they still possess conjunctive intentional content? I don't know the answer to this question. And I suspect that inferential role semantics is not yet well enough developed to fix a determinate answer. What Seager provides, from this perspective, is an intriguing question whose answer will have to wait on future developments in semantic theory. But it is not a question which raises any direct problem for dispositionalist HOT theory, as such.

As is familiar from science, it is possible for one property to be successfully reductively explained in terms of others, even though those others are, as yet, imperfectly understood. Heat in gasses was successfully reductively explained by statistical mechanics, even though in the early stages of the development of the latter, molecules of gas were thought to be like little bouncing billiard balls, with no relevant differences in shape or internal structure. That, I claim, is the kind of position we are now in with respect to phenomenal consciousness. We have a successful reductive explanation of the latter in terms of intentional content (provided by dispositionalist HOT theory), while much scientific work remains to be done to elucidate and explain the nature of intentional content in turn.

5 Seager on reflexive thinking theory

Although I don't think Seager has discovered anything which need trouble dispositionalist HOT theory, he does have a nice argument 'by vicious descent' against my earlier reflexive thinking account (1996). Indeed, I'm quite glad that I don't now have to defend myself against that argument! However, Seager also says that this argument undermines one of the claims made in my 2000, because of my continued commitment to the view that the reflexive thinking account 'may be an accurate description of the [human] mind as it actually is'. But this is a mistake (or a misunderstanding).

I do think that our phenomenally conscious states are not only available to HOT, but are also normally available to conscious HOT, where these HOTs will be phenomenally conscious in turn in the same sort of way. For we can formulate our HOTs linguistically, in imaged natural language sentences (in 'inner speech'), which will be phenomenally conscious in the same way, and for the same reason, that any other sensory images are. For these imaged sentences, too, are available to HOT.

No regress gets started by such claims, because there is no REQUIREMENT that the consciousness-determining HOTs should be conscious ones. Rather, at each level what constitutes a percept or an imaged sentence as phenomenally conscious is its availability to HOT simpliciter. And given that there is nothing in the account to force us to take the next step up in the regress each time, there can be no danger of a regress by vicious descent here, either.

References

Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10.

Carruthers, P. 1996. *Language, Thought and Consciousness: an essay in philosophical psychology*. Cambridge University Press.

Carruthers, P. 2000. *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Clegg, S. 2001. PhD Proposal, University of Sheffield.

Dennett, D. 1978. Why you can't make a computer that feels pain. In his *Brainstorms*, Harvester Press.

Dennett, D. 1991. *Consciousness Explained*. Penguin Press.

Loar, B. 1981. *Mind and Meaning*. Cambridge University Press.

Loar, B. 1982. Conceptual role and truth-conditions. *Notre Dame Journal of Formal Logic*, 23.

MacIntyre, A. 1981. *After Virtue*. Duckworth.

McGinn, C. 1982. The structure of content. In A. Woodfield, ed., *Thought and Object*, Oxford University Press.

McGinn, C. 1989. *Mental Content.* Blackwell.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories.* MIT Press.

Millikan, R. 1986. Thoughts without laws: cognitive science with content. *Philosophical Review*, 95.

Millikan, R. 1989. Biosemantics. *Journal of Philosophy*, 86.

Milner, D. and Goodale, M. 1995. *The Visual Brain in Action*. Oxford University Press.

Papineau, D. 1987. *Reality and Representation.* Blackwell.

Papineau, D. 1993. *Philosophical Naturalism.* Blackwell.

Peacocke, C. 1986. *Thoughts*. Blackwell.

Peacocke, C. 1992. *A Study of Concepts.* MIT Press.

Ramachandran, V. and Blakeslee, S. 1998. *Phantoms in the Brain*. Fourth Estate.

Weiskrantz, L. 1986. *Blindsight.* Oxford University Press.

Young, J. 1986. *Philosophy and the Brain*. Oxford University Press.
