The Unconscious Homunculus

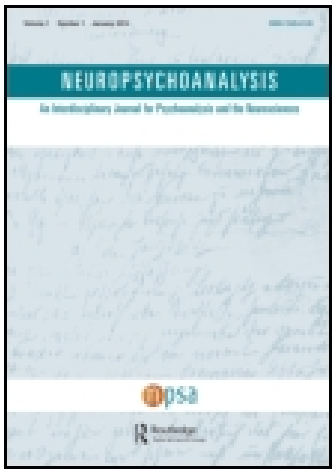
Article *in* Neuropsychoanalysis · January 2014

DOI: 10.1080/15294145.2000.10773273

2 authors, including:

Christof Koch
Allen Institute for Brain Science
919 PUBLICATIONS **95,225** CITATIONS

SEE PROFILE



Neuropsychoanalysis: An Interdisciplinary Journal for Psychoanalysis and the Neurosciences
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rnpa20

The Unconscious Homunculus

Francis Crick[a] & Christof Koch[b]
[a] San Diego
[b] Christof Koch, Division of Biology, 139-74, Caltech, Pasadena, CA 91125, e-mail: ,
Phone: 626-395-6855, Fax: 626-796-8876, Web: klab.caltech.edu
Published online: 09 Jan 2014.

To cite this article: Francis Crick & Christof Koch (2000) The Unconscious Homunculus, Neuropsychoanalysis: An Interdisciplinary Journal for Psychoanalysis and the Neurosciences, 2:1, 3-11, DOI: 10.1080/15294145.2000.10773273

To link to this article: http://dx.doi.org/10.1080/15294145.2000.10773273

The Unconscious Homunculus

Francis Crick (San Diego) and
Christof Koch (Pasadena)

Abstract: *We briefly introduce our approach to understanding the neuronal correlates of consciousness and ask what can be said about the nature of qualia from an introspective, first-person account. We discuss Jackendoff's "Intermediate Level Theory of Consciousness" (1987) as well as related work of others, that implies that we are not directly conscious of our thoughts. We apply this hypothesis to the visual system of the macaque monkey and discuss possible experimental tests.*

Introduction

It is universally agreed that it is not completely obvious how the activity of the brain produces our sensory experiences; more generally, how it produces consciousness. This is what Chalmers has dubbed "The Hard Problem" (Chalmers, 1995). Philosophers are divided about the likely nature of the solution to this problem and whether it is, indeed, a problem at all. For very readable accounts of the nature of some of their discussions and disagreements the reader should consult the book by Searle (1997) with contributions from Chalmers and Dennett or the edited anthology by Shear (1997).

Our own view is that it is a plausible working assumption that some activity of the brain is all that is necessary to produce consciousness, and that this is the best line to follow unless and until there is clear, decisive evidence to the contrary (as opposed to arguments from ignorance). We suspect that our present ideas about how the brain works are likely to turn out to be inadequate; that radically new ideas may be necessary, and that well-formulated suggestions (even way-out ones) should be carefully considered. However, we also believe that, while *Gedanken* experiments are useful devices for generating new ideas or for suggesting difficulties with existing ideas, they do not lead, in general, to trustworthy conclusions. The problem is one that should be approached scientifically and not merely logically. That is, that any theoretical scheme should be pitted against at least one alternative theory, and that *real* experiments should be designed to test between them. (As an example, see our hypothesis that primates are not directly aware of the neural activity in cortical area VI, the primary visual cortex [Crick and Koch, 1995].)

The important first step is to find the neural correlate of consciousness (NCC), for at least one type of consciousness. It is plausible that the NCC involves a very specific set of neurons that are active in some special way. These are distinguished from all other neurons by one or more unique features, such as a particularly strong type of synaptic interconnection, a unique cellular morphology, a particular set of ionic channels or neuromodulators conferring some privileged cellular property, and so on. Note that we are not implying (and have never done so) that consciousness can be found at the level of individual neurons but that consciousness emerges out of the firing behavior of a specific and identifiable subset of all neurons (and that this subset of neurons might be quite small). Alternatively, it is possible that any neuron can, in

---

This article is modified from a chapter that will appear in *The Neuronal Correlates of Consciousness*, ed. T. Metzinger. Cambridge, MA: MIT Press. We refer the reader to this excellent conference volume for much of the relevant philosophical and scientific literature.

Acknowledgments: We thank Dave Chalmers, Patricia Churchland, Ray Jackendoff, Thomas Metzinger, Graeme Mitchinson, Roger Penrose, David Perrett, Tomaso Poggio, Mark Solms, and Richard Stevens. We thank the J. W. Kieckhefer Foundation, the National Institute of Mental Health, the Office of Naval Research, the Vede Foundation, and the National Science Foundation.

Francis Crick is the co-discoverer, with James Watson, of the double helical structure of DNA. Since 1976, he has been at the Salk Institute for Biological Studies in San Diego.

Christof Koch was awarded his Ph.D. in biophysics at the University of Tübingen in Germany (with a minor in philosophy). He joined the California Institute of Technology in 1986, where he is a Professor of Computation and Neural Systems.

principle, at some time or another, contribute to consciousness. A more extreme view would be that consciousness can't be localized to any one subset of neurons and arises out of the holistic interaction of all cells making up the nervous system.

Historically, one of the most interesting formulations of the hypothesis that a particular subset of neurons is responsible for generating conscious experience are the ω neurons introduced by Sigmund Freud in 1895 within his then unpublished "Project for a Scientific Psychology." In this insightful essay of 100 handwritten sheets, Freud attempts to derive a psychology on the basis of the newly formulated neurone theory to which he contributed himself in his thesis work on the neuroanatomy of the stomatogastric ganglion in the crawfish (Shepherd, 1991). He introduces three classes of cells, ϕ, φ, and ω neurons. The first class is involved in mediating perception and the second class in mediating memory; indeed, he postulates that memory is represented by the facilitations existing between the φ neurons at their contact-barriers (i.e., the synapses), a prescient formulation of the idea that memory is encoded in the synaptic weights. The last class of neurons is responsible for mediating consciousness and quality (that is, qualia), even though he admits, "No attempt, of course, can be made to explain how it is that excitatory processes in the ω neurones bring consciousness along with them. It is only a question of establishing a coincidence between the characteristics of consciousness that are known to us and processes in the ω neurones which vary in parallel with them" (p. 311). Frustratingly, a century later, we are still unable to go beyond the correlate. When reading the remainder of the essay, it becomes patently obvious why Freud was unsatisfied with his attempt to link the mind to the brain. At the time, almost nothing was known concerning the biophysics of neurons and the manner in which they communicate, the existence of Broca's area had barely been established and the localization of visual function to the occipital lobe was still controversial. Subsequently, Freud abandoned neurology in favor of pursuing pure psychological theories (see also Solms, 1998).

In approaching the problem, we made the tentative assumption (Crick and Koch, 1998) that all the different aspects of consciousness (for example, pain, visual awareness, self-consciousness, and so on) employ a basic common mechanism or perhaps a few such mechanisms. If one could understand the mechanism for one aspect, then, we hope, we will have gone most of the way toward understanding them all.

For tactical reasons, we believe it is best that several topics should be set aside or merely stated without further discussion, for experience has shown us that otherwise valuable time can be wasted arguing about them without coming to any conclusion.

1. Everyone has a rough idea of what is meant by being conscious. For now, it is better to avoid a precise definition of consciousness because of the dangers of premature definition. Until the problem is understood much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both. In a general sense, consciousness appears to involve attention and some form of short-term memory.

2. It is plausible that some species of animals—in particular the higher mammals—possess some of the essential features of consciousness, but not necessarily all. For this reason, appropriate experiments on such animals are relevant to finding the mechanisms underlying consciousness. It follows that a language system (of the type found in humans) is not essential for consciousness—that is, one can have the key features of consciousness without language. This is not to say that language does not enrich consciousness considerably.

3. It is not profitable at this stage to argue about whether simpler animals (such as octopus, fruit flies, nematodes) are conscious. It is probable that consciousness correlates to some extent with the degree of complexity of any nervous system (Koch and Laurent, 1999). For the same reason, we won't ask whether some parts of our nervous system have a special, isolated consciousness of their own; nor will we spend time discussing whether a digital computer could be conscious.

4. There are many forms of consciousness, such as those associated with seeing, thinking, emotion, pain, and so on. Self-consciousness—that is, the self-referential aspect of consciousness—is probably a special case of consciousness. In our view, it is better left to one side for the moment, especially as it would be difficult to study self-consciousness in a monkey. Various rather unusual states, such as the hypnotic state, lucid dreaming, and sleepwalking, will not be considered here, since they do not seem to us to have special features that would make them experimentally advantageous.

5. Lastly, we personally choose to concentrate on visual perception and visual consciousness. More is known about vision than about any other sensory system. Fortunately, the visual system of primates appears fairly similar to our own (Tootell, Dale, Sereno,

and Malach, 1996), and many experiments on vision have already been done on animals such as the macaque monkey. It is, of course, important to work on alert animals. Very light anesthesia may not make much difference to the response of neurons in macaque VI, but it certainly does to neurons in cortical areas like V4 or IT (inferotemporal).

In this paper we wish to venture a step further by asking what can be said about the precise nature of qualia from an introspective, first-person perspective. Another way to look at the matter is to emphasize that it is qualia that are at the root of the hard problem, and that one needs to have a clear idea of under what exact circumstances qualia occur.

The Intermediate-Level Theory of Consciousness

In earlier publications about the visual system of primates (Crick and Koch, 1995) we suggested that the biological usefulness of visual consciousness in humans is to produce the best current interpretation of the visual scene in the light of past experience, either of ourselves or of our ancestors (embodied in our genes), and to make this interpretation directly available—for a sufficient amount of time—to the parts of the brain that plan possible voluntary motor outputs of one sort or another, including speech.

Philosophers have invented a creature they call a "zombie," who is supposed to act just as normal people do but to be completely unconscious (Chalmers, 1995). While strictly logically possible, this seems to us to be an untenable scientific idea, but there is now suggestive evidence that *part* of the brain does behave like a zombie. That is, in some cases, a person uses current visual input to produce a relevant motor output, without being able to say what was seen. Milner and Goodale (1995) point out that a frog has at least two independent systems for action. These may well be unconscious. One is used by the frog to snap at small, preylike objects, and the other for jumping away from large, looming objects. Why does our brain not consist simply of a series of such specialized zombie systems? We proposed (Crick and Koch, 1995) that such an arrangement is inefficient when very many such systems are required. Better to produce a single but complex representation and make it available for a sufficient time to the parts of the brain that make a choice among many different but possible plans for action. This, in our view, is what seeing is about.

Milner and Goodale (1995) suggest that in primates there are two systems, which we have called the on-line and the seeing systems. The latter is conscious, while the former, acting more rapidly, is not. If a bundle of such unconscious specialized on-line systems could do everything more efficiently than our present arrangement, we would not be conscious of anything.

We decided to reexamine the ideas of Ray Jackendoff (1987) as expressed in his book entitled *Consciousness and the Computational Mind* is which he put forward the Intermediate-Level Theory of Consciousness. Jackendoff's book, which is based on a detailed knowledge of cognitive science, is a closely argued defense of the at-first-sight paradoxical idea that we are not directly conscious of our thoughts, but only of a representation of them in sensory terms. His argument is based on a deep knowledge of modern linguistics and the structure of music, though he also makes some suggestions about the visual system.

Let us first consider Jackendoff's overall view of the mind–brain problem. His analysis postulates three very different domains. These are: (1) the brain, (2) the computational mind, (3) the phenomenological mind.

The brain domain includes both the neurons (and associated cells) and their activities. The computational mind handles information by doing a whole series of "computations" on it. The level of the computational mind is not concerned with exactly how these computations are implemented—this is the standard AI view—but takes for granted that neural instantiation will eventually play an important role in constraining the theory. The domain of the phenomenological mind consists of qualia such as blueness, saltiness, painfulness, and so on. Jackendoff confesses he has no idea how to get blueness and the other experiences to arise out of computation (Chalmers's hard problem). What he is concerned with is what type of computations have qualia associated with them. He is less concerned with the main problem that interests us, which is how some activities of the brain correlate with qualia, though he would agree with us that, roughly speaking, it is the transient results of the computations that correlate with qualia; most of the computations leading up to those results are likely to be *un*conscious. But since computations are implemented in neuronal hardware, these two questions can be connected by asking which parts of the brain are responsible for which computations.

Jackendoff remarks that common sense seems to tell us that awareness and thought are inseparable and that introspection can reveal the contents of the mind. He argues at length that both these beliefs are untrue.

They contrast strongly with his conclusion that thinking is largely unconscious. What is conscious about thoughts is visual or other images, or talking to oneself. He maintains that visual and verbal images are associated with intermediate-level sensory representations, which are in turn generated from thoughts by the fast processing mechanisms in short-term memory. Both the process of thought *and its content* are not directly accessible to awareness.

An example may make this clearer. A bilingual person can express a thought in either language, but the thought itself, which generates the verbal activity or imagery, is not *directly* accessible to him but only in these sensory forms.

Another way of stating these ideas is to say that most of what we are directly aware of falls under two broad headings: (1) a representation of the outer world (including our bodies); and (2) a representation of the inner world, that is, of our thoughts.

This implies that we are neither *directly* aware of the outer world nor of the inner world, although we have the persistent illusion that we are. Curiously enough, this idea, which seems very appealing to us, has attracted rather little attention from brain scientists though it dates back to at least as early as Immanuel Kant. In addition: (3) both of these representations are expressed solely in sensory terms.

To appreciate these arguments, the reader should consult Jackendoff (1987) as well as some updates to these ideas in Jackendoff (1996). For the visual system he proposed ideas based on the theories of David Marr. Marr argued in his posthumous book *Vision* (1982) that it would be almost certainly impossible for the brain to arrive at a visual representation, corresponding to what we consciously see, in only one step. He therefore suggested a hypothetical series of stages. In his analysis he concentrated on the documentation of shape, though he realized that a fuller treatment would include movement, texture, and color.

Marr proposed four possible stages. The first one he called "Image" (there might be several of such steps). This simply represents the light intensity value at each point in the visual image. The second he called the "Primal sketch." This makes explicit important information about the two-dimensional image, such as edge segments, terminations, etc. The third stage was the "2-1/2 sketch." This makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities, in a *viewer-centered* coordinate frame. The fourth and final step he called the 3D model representation. This describes shapes and their special organization in an *object-centered* frame.

Work on the visual system of the macaque does indeed suggest that it consists of a series of stages (Felleman and Van Essen, 1991) and that these follow one another along the broad lines suggested by Marr, but the system probably does not display the exact stages he suggested, and is probably considerably more complicated. For the sake of convenience, though, we will continue to use his nomenclature.

Jackendoff proposes that we are directly conscious of an extended version of something corresponding roughly to Marr's 2-½D sketch but not of his 3D model. For instance, when we look at a person's face we are directly conscious of the shape, color, movement, and so on, of the front of her face (like the 2-½D sketch), but not of the back of her head, though we can imagine what the back of her head might look like, deriving this image from a 3D model of the head of which we are not *directly* conscious.

The experimental evidence shows that the higher levels of the visual system, in the various inferotemporal regions, have neurons that do appear to respond mainly to something like an enriched 2-½D sketch, and show a certain amount of size, position, and rotation invariance. This has been especially studied for faces and, more recently, for artificial bent-wire 3D shapes (Perrett, Oram, Hietanen, and Benson, 1994; Logothetis, Pauls, and Poggio, 1995; Logothetis and Pauls, 1995; Booth and Rolls, 1998). We will discuss these results more fully in a later section.

Similar Suggestions

We have located several other suggestions along similar lines. There are probably more (for a philosophical perspective, see Metzinger [1995]; for a dissenting view, see Siewert [1998]).

An idea somewhat similar to Jackendoff's was put forward by Sigmund Freud. Consider this statement (Freud, 1900): "What part is there left to be played in our scheme by consciousness, which was once so omnipotent and hid all else from view? Only that of a sense-organ for the perception of psychical qualities" (p. 615). Or this quotation from Freud's essay on "The Unconscious," published in 1915: "In psycho-analysis there is no choice but for us to assert that mental processes are in themselves unconscious, and to liken the perception of them by means of consciousness to the perception of the external world by means of sense-organs" (p. 171).

(The quotation was brought to our attention in a paper by Mark Solms [1997] who states that Freud probably derived the idea from Kant, either directly or indirectly.) As is well-known, Freud was driven to this idea by his studies on disturbed patients. "He found that without making this assumption he was unable to explain or even describe a large variety of phenomena which he came across" (Solms, 1997). Or the following, perhaps even more direct quote from Freud's writings (1923): "It dawns upon us like a new discovery that only something which has once been a perception can become conscious, and that anything arising from within (apart from feelings) that seeks to become conscious must try to transform itself into external perception" (p. 19).

There is also the well-known claim by Karl Lashley. In his provocative *Cerebral Organization and Behaviour* (1956) he wrote:

> *No activity of mind is ever conscious.* [Lashley's italics] This sounds like a paradox, but it is nonetheless true. There are order and arrangement, but there is no experience of the creation of that order. I could give numberless examples, for there is no exception to the rule. A couple of illustrations should suffice. Look at a complicated scene. It consists of a number of objects standing out against an indistinct background: desk, chairs, faces. Each consists of a number of lesser sensations combined in the object, but there is no experience of putting them together. The objects are immediately present. When we think in words, the thoughts come in grammatical form with subject, verb, object, and modifying clauses falling into place without our having the slightest perception of how the sentence structure is produced. . . . Experience clearly gives no clue as to the means by which it is organized [p. 7].

In other words, Lashley believed that the processes underlying thoughts, imagery, silent speech, and so on are unconscious while only their content may be accessible to consciousness. However, it is not clear that Lashley was suggesting that all conscious thoughts are expressed solely in sensory terms.

Finally, we discovered a suggestion almost identical to Jackendoff's in the work of Stevens (1997). He concludes from periods of closely observed introspection that, "Conscious awareness is essentially perceptual. It consists entirely of perceptual images. These may be directly stimulated by outside events or internally generated in the more elusive and less well defined form of remembered or imagined percepts" (p. 107).

Among perceptual images he includes unspoken speech. This is in striking agreement with Jackendoff's ideas, which were then largely unknown to him. Stevens (1997) also makes the point that consciousness is necessary for certain forms of evaluations, "because it is only when thoughts and possibilities are conscious in the form of words and/or images that we can begin to compare and contrast them" (p. 117).

It is worth noting that all three authors appeared to have arrived at broadly the same conclusion from significantly different evidence. The exclamation, "How do I know what I think till I hear what I say?" shows that the idea is not unknown to ordinary people.

Let us assume, therefore, that qualia are associated with sensory percepts, and make a few rather obvious points about them. Apart from the fact that they differ from each other (red is quite different from blue, and both from a pain or a sound), qualia also differ in intensity and duration. Thus the qualia associated with the visual world, in a good light, are more vivid than a recollection of the same visual scene (vivid visual recollections are usually called hallucinations). A quale can be very transient, and pass so quickly that we may have little or no recollection of it. Neither of these two properties is likely to cause any special difficulties when we consider the behavior of neurons, since neurons can easily express intensity and duration.

However, there is a class of conscious percepts which have a rather different character from straightforward sensory percepts. Jackendoff originally used the term *affect* to describe them, though more recently he has substituted the term *valuation* (Jackendoff, 1996). Examples would be a feeling of familiarity, or novelty, or the tip-of-the-tongue feeling, and all the various emotions. It is not clear whether these feelings exist in their own right, or are merely certain mixtures of various bodily sensations. Both Freud (1923) as well as Stevens (1997) treated "feelings" as a special case. We propose to leave these more diffuse percepts on one side for the moment, though eventually they, too, will have to be explained in neural terms. Some of these may merely express simple relationships (such as "same as" or "different from") between sensory qualia.

The Homunculus

The homunculus (note that we are here not referring to the map of the human body surface sensibilities as

mapped onto the cortex; Penfield and Boldrey, [1937]) is usually thought of as a "little man inside the head," who perceives the world through the senses, thinks and plans, and executes voluntary actions. In following up this idea we came across a "Comment" by Fred Attneave (1961), entitled "In Defense of Homunculi." He lists two kinds of objections to a homunculus. The first is an aversion to dualism, since it might involve "a fluffy kind of nonmatter . . . quite beyond the pale of scientific investigation" (p. 777). The second has to do with the supposed regressive nature of the concept; who is looking at the brain states of the homunculus? Attneave notes, "that we fall into a regress only if we try to make the homunculus do everything. The moment we specify certain processes that occur outside the homunculus, we are merely classifying or portioning psychoneural functions; the classification may be crude but it is not itself as regressive" (p. 778). He puts forward a very speculative overall block diagram of the brain, involving hierarchical sensory processing, an affect system, a motor system, and a part he calls H, the homunculus. It is reciprocally connected to the perceptual machinery at various levels in the hierarchy, not merely the higher ones. It receives an input from the affective centers and projects to the motor machinery (there are other details about reflexes, skills, proprioception, and so on). He emphasizes that his scheme avoids the difficulty of an infinite regress.

Attneave tentatively locates the homunculus in a subcortical location, such as the reticular formation, and he considers it to be conscious. Yet his basic idea is otherwise very similar to the one discussed above. We all have this illusion of a homunculus inside the brain (that's what "I" am), so this illusion needs an explanation. The problem of the infinite regress is avoided in our case, since the true homunculus is *un*conscious, and only a representation of it enters consciousness. This puts the problem of consciousness in a somewhat new light. We have therefore named this type of theory as one postulating an *unconscious homunculus,* wherever it may be located in the brain. The unconscious homunculus receives information about the world through the senses and thinks, plans, and executes voluntary actions. What becomes conscious then is a representation of some of the activities of the unconscious homunculus in the form of various kinds of imagery and spoken and unspoken speech. Notice that this idea does not, by itself, explain how qualia arise.

The concept of the unconscious homunculus is not a trivial one. It does throw a new light on certain other theoretical approaches. For example, it may make Penrose's worries about consciousness unnecessary. Penrose (1989, 1997) has argued that present-day physics is not capable of explaining how mathematicians think, but if all such thinking is necessarily *un*conscious—as mathematicians and scientists have testified (Hadamard, 1945) that certainly some of it is—then although something such as quantum gravity may be needed for certain types of thinking, it may not be required to explain consciousness as such. Penrose has given no argument that sensory experiences themselves are difficult to explain in terms of present-day physics.

Possible Experimental Approaches

In approaching a system as complex as the brain, it is important to have some idea, however provisional, as to what to look for. Let us therefore follow these authors and adopt the idea of the unconscious homunculus as a tentative working hypothesis, and ask what experiments might be done to support it. For the moment we will concentrate on the visual system.

What we are trying to identify is the activity of the brain that produces visual qualia. We have argued (Crick and Koch, 1995, 1998) that whatever other properties are involved we should expect to find neurons whose firing is in some way correlated with the type of qualia being perceived. So it is not unreasonable to ask which are the neurons whose activity is correlated with Marr's 2-½D sketch (roughly speaking, the visual features of which we are directly aware) and which are the neurons whose activity is correlated with Marr's 3D model (of which we are only indirectly aware). For the moment we will assume that this latter activity is represented somewhere in the cortex and leave aside other less likely possibilities, such as in the reticular formation or the claustrum.

As far as we know, there are only a few sets of relevant experimental results. One of the earliest is due to Perrett and his coworkers in their study on the neurons in the alert macaque monkey that respond to faces (Perrett, Hietanen, Oram, and Benson, 1992). Most of the neurons in the higher levels of the visual system that respond to faces fire only to one aspect of the face, usually a specific view. The firing is somewhat independent of scale, of small translations, and of some degree of rotation (Pauls, Bricolo, and Logothetis, 1996). These neurons look as if they are members of a distributed representation of a particular view of a face, as suggested by the theoretical work of Pog-

gio (1990; see also Poggio and Edelman, 1990; Logothetis, Pauls, Bülthoff, and Poggio, 1994) and supported (on a lightly anesthetized macaque) by Young and Yamane (1992).

However, Perrett, Hietanen, et al. (1992) do report 6 neurons (4% of the total) that respond to *all* horizontal views of a head: that is, they are view-invariant. These might be taken to be part of a 3D model representation. However, it is known that some of the circuits in the hidden layers of a 3-level feed-forward neural network, trained by back-projection, often have somewhat unusual properties (Sejnowski and Rosenberg, 1987), so one could argue that these apparent 3D model neurons are really only a small accidental part of a 2-½D sketch. Against this interpretation, Perrett, Hietanen, et al. (1992) claim that these 6 neurons have a significantly longer latency (130 msecs against 119 msecs), suggesting that they are one step higher in the visual hierarchy. The crucial question is whether these minority of neurons are of a different *type* from the view-specific face neurons (for example, project to a different place), and this is not known. Here lies one possibly fruitful direction for future research.

Another example comes from the experiments of Logothetis and Pauls (1995) on the responses of the neurons in an alert macaque, again in the higher levels of the visual hierarchy, to artificial paper-clip-like models. Again, a minority of neurons (8 of the 773 cells analyzed) respond in a view-independent manner, but in these experiments the latencies were not measured, nor was it known exactly which type of neuron was being recorded (N. Logothetis, personal communication). More recently, Booth and Rolls (1998) also report the existence of view-independent cells in the inferior temporal cortex of the macaque.

A naive interpretation of our general idea would be that the face representations in the macaque prefrontal cortex reported by O Scalaidhe, Wilson, and Goldman-Rakic (1997), would be implemented solely by view-independent neurons, and without any view-dependent ones. As far as we know, this has not yet been studied. Note that while the activity of view-independent neurons should always be unconscious, it does not follow that the activities of all view-dependent ones must always be conscious. Our unconscious thoughts may well involve neurons of this latter type.

We think this simple guess at the location of these two types of neurons is rather unlikely, though we would not be surprised if the percentage of neurons showing view-invariance turns out to be higher in prefrontal areas than the very small fractions reported in inferotemporal cortex. One might also find a higher percentage in such areas as the parahippocampal gyrus and the perihinal cortex, leading to the hippocampus. Whether they will also be found in parts of the thalamus and in the amygdala remains an empirically open question.

Another possibility is that, contrary to Jackendoff's suggestion, there is no true, object-centered (3D) visual representation in an explicit form in the brain. That is, object-centered information is never made explicit at the level of individual neurons, being coded instead in an implicit manner across a distributed set of neurons. While there are still unconscious computations that lead up to thoughts, the results of the computations are expressed directly in sensory, viewer-centered terms. If this were true, the search for view-invariant neurons in prefrontal cortex would be unsuccessful.

We have briefly considered the visual system, but though they are outside the scope of this paper, the same analysis should be applied to the other sensory systems, such as audition, touch, olfaction, and pain. It may not always be completely obvious what the difference is between (unconscious) thoughts and the (conscious) sensory representations of these thoughts in these systems. The crucial test to distinguish between these two is whether any qualia are involved beyond mental imagery and unspoken speech (e.g., the putative *noniconic* thoughts of Siewert [1998]). We leave this to the future.

Another problem concerns our guess that unconscious thought processes may be located in some places in the prefrontal cortex. First, it is not clear exactly where prefrontal cortex ends as one proceeds posteriorly, especially in the general region of the insula. Second, the selection of "prefrontal" cortex (or a subset thereof) in this way seems rather arbitrary. It would be more satisfactory if there were a more operational definition, such as those cortical areas receiving a projection from the basal ganglia, via the thalamus (usually thalamic area MD). It is conceivable that the rather rapid sequential winner-take-all operations performed by the basal ganglia may not be compatible with consciousness, but are frequently used by more rapid, unconscious thought processes.

Conclusion

It cannot be overstated that Chalmers's hard problem of consciousness is unlikely to yield to a purely logical

or philosophical attack. Rather, it needs to be approached in a reductionist, scientific manner.

The picture that emerges from a review of the existing, fragmentary evidence is quite surprising. As has often been assumed, we are not directly aware of the outer world of sensory events. Instead, we are conscious of the results of some of the computations performed by the nervous system on the various neural representations of this sensory world. These results are expressed in various cortical areas (excluding primary visual cortex; Crick and Koch [1995]). Nor are we directly aware of our inner world of thoughts, intentions, and planning (that is, of our unconscious homunculus) but—and this is the surprising part—only of the sensory representations associated with these mental activities. What remains is the sobering realization that our subjective world of qualia—what distinguishes us from zombies and fills our life with color, music, smells, and other vivid sensations—is probably caused by the activity of a small fraction of all the neurons in the brain, located strategically between the outer and the inner worlds. How this activity acts to produce the subjective world that is so dear to us is still a complete mystery.

References

Attneave, F. (1961), In defense of homunculi. In: *Sensory Communication,* ed. W. A. Rosenblith. Cambridge, MA: MIT Press, pp. 777–782.

Booth, M. C. A., & Rolls, E. T. (1998), View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex,* 8:510–523.

Chalmers, D. (1995), *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press.

Crick, F., & Koch, C. (1995), Are we aware of neural activity in primary visual cortex? *Nature,* 375:121–123.

——— ——— (1998), Consciousness and neuroscience. *Cereb. Cortex,* 8:97–107.

Felleman, D. J., & Van Essen, D. (1991), Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex,* 1:1–47.

Freud, S. (1895), Project for a scientific psychology. *Standard Edition,* 1:281–391. London: Hogarth Press, 1966.

——— (1900), The Interpretation of Dreams. *Standard Edition,* 4&5. London: Hogarth Press, 1953.

——— (1915), The unconscious. *Standard Edition,* 14:159–204. London: Hogarth Press, 1957.

——— (1923), The Ego and the Id. *Standard Edition,* 19:1–59. London: Hogarth Press, 1961.

Hadamard, J. (1945), *The Mathematician's Mind.* Princeton, NJ: Princeton University Press.

Jackendoff, R. (1987), *Consciousness and the Computational Mind.* Cambridge, MA: MIT Press.

——— (1996), How language helps us think. *Pragmatics & Cognition,* 4:1–34.

Koch, C., & Laurent, G. (1999), Complexity and the nervous system. *Science,* 289:96–98.

Lashley, K. (1956), Cerebral organization and behavior. In: *The Brain and Human Behavior, Proc. Assn. Nervous & Mental Disease.* New York: Halner, pp. 1–18.

Logothetis, N. K., & Pauls, J. (1995), Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex,* 3:270–288.

——— ——— Bülthoff, H. H., & Poggio, T. (1994), View-dependent object recognition by monkeys. *Curr. Biol.,* 4:401–414.

——— ——— Poggio, T. (1995), Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.,* 5:552–563.

Marr, D. (1982), *Vision.* San Francisco, CA: W. H. Freeman.

Metzinger, T. (1995), Einleitung: Das Problem des Bewußtseins (Introduction: The problem of consciousness). In: *Bewußtsein,* ed. T. Metzinger. Germany: Paderborn.

——— (2000), *The Neuronal Correlates of Consciousness.* Cambridge, MA: MIT Press.

Milner, D., & Goodale, M. (1995), *The Visual Brain in Action.* Oxford: Oxford University Press.

O Scalaidhe, S. P., Wilson, F. A. W., & Goldman-Rakic, P. S. (1997), Areal segregation of face-processing neurons in prefrontal cortex. *Science,* 278:1135–1138.

Pauls, J., Bricolo, E., & Logothetis, N. (1996), View invariant representations in monkey temporal cortex: Position, scale, and rotational invariance. In: *Early Visual Learning,* ed. S. K. Nayar & T. Poggio. New York: Oxford University Press, pp. 9–41.

Penfield, W., & Boldrey, E. (1937), Somatic motor and sensory representations in the cerebral cortex of man as studied by electrical stimulation. *Brain,* 60:389–443.

Penrose, R. (1989), *The Emperor's New Mind.* Oxford: Oxford University Press.

——— (1997), *The Large, the Small and the Human Mind.* Cambridge, U.K.: Cambridge University Press.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992), Organization and functions of cells responsive to faces in the temporal cortex. *Philosoph. Trans. Roy. Soc. London B,* 335:23–30.

——— Oram, M. W., Hietanen, J. K., & Benson, P. J. (1994), Issues of representation in object vision. In: *The Neuropsychology of High-Level Vision,* ed. M. J. Farah & G. Ratcliff. Hillsdale, NJ: Lawrence Erlbaum, pp. 33–61.

Poggio, T. (1990), A theory of how the brain might work. *Cold Spring Harbor Symp. Quant. Biol.,* 55:899–910.

——— Edelman, S. (1990), A network that learns to recognize three-dimensional objects. *Nature,* 343:263–266.

Searle, J. R. (1997), *The Mystery of Consciousness.* New York: New York Review of Books.

Sejnowski, T. J., & Rosenberg, C. R. (1987), Parallel networks that learn to pronounce English text. *Complex Systems,* 1:145–168.

Shear, J. (1997), *Explaining Consciousness: The Hard Problem.* Cambridge, MA: MIT Press.
Shepherd, G. M. (1991), *Foundations of the Neuron Doctrine.* New York: Oxford University Press.
Siewert, C. P. (1998), *The Significance of Consciousness.* Princeton, NJ: Princeton University Press.
Solms, M. (1997), What is consciousness? *J. Amer. Psychoanl. Assn.,* 45:681–703.
——— (1998), Before and after Freud's Project. In: *Neuroscience of the Mind on the Centennial of Freud's Project for a Scientific Psychology,* ed. R. M. Bilder & F. F. LeFever. *Ann. NY Acad. Sci.* 843:1–10.
Stevens, R. (1997), Western phenomenological approaches to the study of conscious experience and their implications. In: *Methodologies for the Study of Consciousness: A New Synthesis,* ed. J. Richardson & M. Velmans. Kalamazoo: Fetzer Institute, pp. 100–123.
Tootell, R. B. H., Dale, A. M., Sereno, M. I., & Malach, R. (1996), New images from human visual cortex. *Trends Neurosci.,* 19:481–489.
Young, M. P., & Yamane, S. (1992), Sparse population coding of faces in the inferotemporal cortex. *Science,* 256:1327–1331.

Christof Koch
Division of Biology, 139–74
Caltech, Pasadena, CA 91125
e-mail: koch@klab.caltech.edu
Phone: 626-395-6855
Fax: 626-796-8876
Web: klab.caltech.edu

Consciousness Cannot Be Limited to Sensory Qualities: Some Empirical Counterexamples: Commentary by Bernard J. Baars and Katharine A. McGovern (Berkeley, CA)

The idea proposed by Crick and Koch that conscious contents are confined to sensory events is attractive, in part because it is easier to study consciousness in the senses than anywhere else. The last 10 years have seen particularly good progress in studies of the visual cortex, where the question of visual consciousness has almost become normal science. This is an exceptional event in this period of scientific evasion of consciousness (and unconsciousness as well), and it bodes well for a better understanding of both of these essential concepts. Francis Crick and Christof Koch have made pioneering contributions to this emerging literature.

According to Crick and Koch, Freud wrote at times of consciousness as "a sense-organ for the perception of psychical qualities" (1900, p. 615). However, the expression "psychical qualities" would seem to extend beyond sensations to other mental states like thoughts, feelings, intuitions, concepts, beliefs, expectations, and intentions. Fifteen years later Freud wrote of this point as an analogy: "to liken the perception of (unconscious contents) by means of consciousness to the perception of the external world by means of sense-organs" (1915, p. 171). It is only in 1923 that he seems to take it literally: "It dawns upon us like a new discovery that only something which has once been a perception can become conscious, and that anything arising from within (apart from feelings) that seeks to become conscious must try to transform itself into external perception" (1923, p. 19). This is essentially Crick and Koch's claim, following Jackendoff. If we extend the notion of perception to events like mental images, inner speech, and somatically referred sensations, which are internally generated perceptlike experiences, it seems as if all of our mental lives can be understood as sensory in some way. It is in fact quite an old idea. Long before Freud, Plato and Aristotle made their claims upon it.

Shortly before 1900 a long and intractable controversy took place in Continental psychology about precisely this issue, in the so-called "imageless thought" debate. The Wurzburg School of empirical psychology asked, can thoughts exist without images, which are internally generated sense experiences? It was a difficult claim for the nineteenth century to resolve. The debate is therefore quite old. Note that this is not a neuroscientific question primarily, but a psychological one, dependent on the best information we can get about the actual experience of human beings. In the next section, therefore, we will provide some evidence the reader can assess experientially, to see whether his or her own conscious experience is fundamentally sensory.

Bernard J. Baars is Institute Faculty Professor at the Wright Institute in Berkeley, California, and author of a number of significant books and articles on the problem of consciousness.

Katharine A. McGovern is a cognitive psychologist and Adjunct Professor at the Wright Institute. She has a special interest in the problem of emotional feelings in the Jamesian "fringe" of consciousness.