united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1343/15

EU ADVANCED COURSE IN COMPUTATIONAL NEUROSCIENCE

An IBRO Neuroscience School

(30 July - 24 August 2001)

"Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework"

presented by:

Stanislas DEHAENE
Unité INSERM 334
Service Hospitalier Frédéric Joliot
CEA/DRM/DSV
4 Place du Général Leclerc
91401 Orsay Cedex
FRANCE

ELSEVIER

Cognition 79 (2001) 1–37

COGNITION

www.elsevier.com/locate/cognit

Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework

Stanislas Dehaene*, Lionel Naccache

Unité INSERM 334, Service Hospitalier Frédéric Joliot, CEA/DRM/DSV, 4, Place du Général Leclerc, 91401 Orsay Cedex, France

Received 8 February 2000; accepted 27 September 2000

Abstract

This introductory chapter attempts to clarify the philosophical, empirical, and theoretical bases on which a cognitive neuroscience approach to consciousness can be founded. We isolate three major empirical observations that any theory of consciousness should incorporate, namely (1) a considerable amount of processing is possible without consciousness, (2) attention is a prerequisite of consciousness, and (3) consciousness is required for some specific cognitive tasks, including those that require durable information maintenance, novel combinations of operations, or the spontaneous generation of intentional behavior. We then propose a theoretical framework that synthesizes those facts: the hypothesis of a global neuronal workspace. This framework postulates that, at any given time, many modular cerebral networks are active in parallel and process information in an unconscious manner. An information becomes conscious, however, if the neural population that represents it is mobilized by top-down attentional amplification into a brain-scale state of coherent activity that involves many neurons distributed throughout the brain. The long-distance connectivity of these 'workspace neurons' can, when they are active for a minimal duration, make the information available to a variety of processes including perceptual categorization, long-term memorization, evaluation, and intentional action. We postulate that this global availability of information through the workspace is what we subjectively experience as a conscious state. A complete theory of consciousness should explain why some cognitive and cerebral representations can be permanently or temporarily inaccessible to consciousness, what is the range of possible conscious contents, how they map onto specific cerebral circuits, and whether a generic neuronal mechanism underlies all of them. We confront the workspace model with those issues and identify novel experimental predictions. Neurophysiological, anatomical, and brain-imaging data strongly argue for a major role of prefrontal cortex,

* Corresponding author. Tel.: +33-1-69-86-78-73; fax: +33-1-69-86-78-16.
E-mail address: dehaene@shfj.cea.fr (S. Dehaene).

PII: S0010-0277(00)00123-2

anterior cingulate, and the areas that connect to them, in creating the postulated brain-scale workspace.

Keywords: Consciousness; Awareness; Attention; Priming

1. Introduction

The goal of this volume is to provide readers with a perspective on the latest contributions of cognitive psychology, neuropsychology, and brain imaging to our understanding of consciousness. For a long time, the word 'consciousness' was used only reluctantly by most psychologists and neuroscientists. This reluctance is now largely overturned, and consciousness has become an exciting and quickly moving field of research. Thanks largely to advances in neuropsychology and brain imaging, but also to a new reading of the psychological and neuropsychological research of the last decades in domains such as attention, working memory, novelty detection, or the body schema, a new comprehension of the neural underpinnings of consciousness is emerging. In parallel, a variety of models, pitched at various levels in neural and/or cognitive science, are now available for some of its key elements.

Within this fresh perspective, firmly grounded in empirical research, the problem of consciousness no longer seems intractable. Yet no convincing synthesis of the recent literature is available to date. Nor do we know yet whether the elements of a solution that we currently have will suffice to solve the problem, or whether key ingredients are still missing. By grouping some of the most innovative approaches together in a single volume, this special issue aims at providing the readers with a new opportunity to see for themselves whether a synthesis is now possible.

In this introduction, we set the grounds for subsequent papers by first clarifying what we think should be the aim of a cognitive neuroscience approach to consciousness. We isolate three major findings that are explored in greater detail in several chapters of this volume. Finally, we propose a synthesis that integrates them into what we view as a promising theoretical framework: the hypothesis of a global neuronal workspace. With this framework in mind, we look back at some of the remaining empirical and conceptual difficulties of consciousness research, and examine whether a clarification is in sight.

2. Nature of the problem and range of possible solutions

Let us begin by clarifying the nature of the problem that a cognitive neuroscience of consciousness should address. In our opinion, this problem, though empirically challenging, is conceptually simple. Human subjects routinely refer to a variety of conscious states. In various daily life and psychophysical testing situations, they use phrases such as 'I was not conscious of X', 'I suddenly realized that Y', or 'I knew that Z, therefore I decided to do X'. In other words, they use a vocabulary of psychological attitudes such as believing, pretending, knowing, etc., that all involve to various extents the concept of 'being conscious'. In any given situation, such conscious phenomenological reports can be very consistent both within and across subjects. The task of cognitive neuroscience is to identify which mental representations and, ultimately, which brain states are associated with such reports. Within a materialistic framework, each instance of mental activity is also a physical brain state.[1] The cognitive neuroscience of consciousness aims at determining whether there is a systematic form of information processing and a reproducible class of neuronal activation patterns that systematically distinguish mental states that subjects label as 'conscious' from other states.[2]

From this perspective, the problem of the cognitive neuroscience of consciousness does not seem to pose any greater conceptual difficulty than identifying the cognitive and cerebral architectures for, say, motor action (identifying what categories of neural and/or information-processing states are systematically associated with moving a limb). What is specific to consciousness, however, is that the object of our study is an introspective phenomenon, not an objectively measurable response. Thus, the scientific body of consciousness calls for a specific attitude which departs from the 'objectivist' or 'behaviorist' perspective often adopted in behavioral and neural experimentation. In order to cross-correlate subjective reports of consciousness with neuronal or information-processing states, the first crucial step is *to take seriously introspective phenomenological reports.* Subjective reports are the key phenomena that a cognitive neuroscience of consciousness purport to study. As such, they constitute primary data that need to be measured and recorded along with other psychophysiological observations (Dennett, 1992; Weiskrantz, 1997; see also Merikle, Smilek, & Eastwood, this volume).

The idea that introspective reports must be considered as serious data in search of a model does not imply that introspection is a privileged mode of access to the inner workings of the mind. Introspection can be wrong, as is clearly demonstrated, for instance, in split-brain subjects whose left-hemispheric verbal 'interpreter' invents a plausible but clearly false explanation for the behavior caused by their right hemisphere (Gazzaniga, LeDoux, & Wilson, 1977). We need to find a scientific explanation for subjective reports, but we must not assume that they always constitute accurate descriptions of reality. This distinction is clearest in the case of hallucinations. If someone claims to have visual hallucinations of floating faces, or 'out-of-body' experiences, for instance, it would be wrong to take these reports as unequi-

[1] We use the word 'state' in the present context to mean any configuration of neural activity, whether stable (a fixed point) or dynamic (a trajectory in neural space). It is an open question as to whether neural states require stability over a minimal duration to become conscious, although the workspace model would predict that some degree of stable amplification over a period of at least about 100 ms is required.

[2] One should also bear in mind the possibility that what naive subjects call 'consciousness' will ultimately be parceled into distinct theoretical constructs, each with its own neural substrate, just like the naive concept of 'warmth' was ultimately split into two distinct physical parameters, temperature and heat.

vocal evidence for parapsychology, but it would be equally wrong to dismiss them as unverifiable subjective phenomena. The correct approach is to try to explain how such conscious states can arise, for instance by appealing to an inappropriate activation of face processing or vestibular neural circuits, as can indeed be observed by brain-imaging methods during hallucinations (Ffytche et al., 1998; Silbersweig et al., 1995).

The emphasis on subjective reports as data does not mean that the resulting body of knowledge will be inherently subjective and therefore non-scientific. As noted by Searle (1998), a body of knowledge is scientific ('epistemically objective') inasmuch as it can be verified independently of the attitudes or preferences of the experimenters, but there is nothing in this definition that prevents a genuinely scientific approach of domains that are inherently subjective because they exist only in the experience of the subject ('ontologically subjective' phenomena). "The requirement that science be objective does not prevent us from getting an epistemically objective science of a domain that is ontologically subjective." (Searle, 1998, p. 1937).

One major hurdle in realizing this program, however, is that "we are still in the grip of a residual dualism" (Searle, 1998, p. 1939). Many scientists and philosophers still adhere to an essentialist view of consciousness, according to which conscious states are ineffable experiences of a distinct nature that may never be amenable to a physical explanation. Such a view, which amounts to a Cartesian dualism of substance, has led some to search for the bases of consciousness in a different form of physics (Penrose, 1990). Others make the radical claim that two human brains can be identical, atom for atom, and yet one can be conscious while the other is a mere 'zombie' without consciousness (Chalmers, 1996).

Contrary to those extreme statements, contributors to the present volume share the belief that the tools of cognitive psychology and neuroscience may suffice to analyze consciousness. This need not imply a return to an extreme form of direct psychoneural reductionism. Rather, research on the cognitive neuroscience of consciousness should clearly take into account the many levels of organization at which the nervous system can be studied, from molecules to synapses, neurons, local circuits, large scale networks, and the hierarchy of mental representations that they support (Changeux & Dehaene, 1989). In our opinion, it would be inappropriate, and a form of 'category error', to attempt to reduce consciousness to a low level of neural organization, such as the firing of neurons in thalamocortical circuits or the properties of NMDA receptors, without specifying in functional terms the consequences of this neural organization at the cognitive level. While characterization of such neural bases will clearly be indispensable to our understanding of consciousness, it cannot suffice. A full theory will require many more 'bridging laws' to explain how these neural events organize into larger-scale active circuits, how those circuits themselves support specific representations and forms of information processing, and how these processes are ultimately associated with conscious reports. Hence, this entire volume privileges cognitive neuroscientific approaches to consciousness that seem capable of addressing both the cognitive architecture of mental representations and their neural implementation.

3. Three fundamental empirical findings on consciousness

In this section, we begin by providing a short review of empirical observations that we consider as particularly relevant to the cognitive neuroscience of consciousness. We focus on three findings: the depth of unconscious processing; the attention-dependence of conscious perception; and the necessity of consciousness for some integrative mental operations.

3.1. Cognitive processing is possible without consciousness

Our first general observation is that a considerable amount of processing can occur without consciousness. Such unconscious processing is open to scientific investigation using behavioral, neuropsychological and brain-imaging methods. By increasing the range of cognitive processes that do not require consciousness, studies of unconscious processing contribute to narrowing down the cognitive bases of consciousness. The current evidence indicates that many perceptual, motor, semantic, emotional and context-dependent processes can occur unconsciously.

A first line of evidence comes from studies of brain-lesioned patients. Pöppel, Held, and Frost (1973) demonstrated that four patients with a partial blindness due to a lesion in visual cortical areas (hemianopsic scotoma) remained able to detect visual stimuli presented in their blind field. Although the patients claimed that they could not see the stimuli, indicating a lack of phenomenal consciousness, they nevertheless performed above chance when directing a visual saccade to them. This 'blindsight' phenomenon was subsequently replicated and extended in numerous studies (Weiskrantz, 1997). Importantly, some patients performed at the same level as control subjects, for instance in motor pointing tasks. Thus, unconscious processing is not limited to situations in which information is degraded or partially available. Rather, an entire stream of processing may unfold outside of consciousness.

Dissociations between accurate performance and lack of consciousness were subsequently identified in many categories of neuropsychological impairments such as visual agnosia, prosopagnosia, achromatopsia, callosal disconnection, aphasia, alexia, amnesia, and hemineglect (for reviews, see Köhler & Moscovitch, 1997; Schacter, Buckner, & Koutstaal, 1998; see also Driver & Vuilleumier, this volume). The current evidence suggests that, in many of these cases, unconscious processing is possible at a perceptual, but also a semantic level. For instance, Renault, Signoret, Debruille, Breton, and Bolgert (1989) recorded event-related potentials to familiar and unknown faces in a prosopagnosic patient. Although the patient denied any recognition of the familiar faces, an electrical waveform indexing perceptual processing, the P300, was significantly shorter and more intense for the familiar faces. Similar results were obtained by recording the electrodermal response, an index of vegetative processing of emotional stimuli, in prosopagnosic patients (Bauer, 1984; Tranel & Damasio, 1985). Even clearer evidence for semantic-level processing comes from studies of picture–word priming in neglect patients (McGlinchey-Berroth, Milberg, Verfaellie, Alexander, & Kilduff, 1993). When two images are presented simultaneously in the left and right visual fields, neglect patients deny seeing the one on the left, and indeed cannot report it beyond chance level. Nevertheless, when having to perform a lexical decision task on a subsequent foveal word, which can be related or unrelated to the previous image, they show the same amount of semantic priming from both hemifields, indicating that even the unreportable left-side image was processed to a semantic level.

Similar priming studies indicate that a considerable amount of unconscious processing also occurs in normal subjects. Even a very brief visual stimulus can be perceived consciously when presented in isolation. However, the same brief stimulus can fail to reach consciousness when it is surrounded in time by other stimuli that serve as masks. This lack of consciousness can be assessed objectively using signal detection theory (for discussion, see Holender, 1986; Merikle, 1992; see also Merikle et al.).[3] Crucially, the masked stimulus can still have a measurable influence on the processing of subsequent stimuli, a phenomenon known as masked priming. There are now multiple demonstrations of perceptual, semantic, and motor processing of masked stimuli. For instance, in various tasks, processing of a conscious target stimulus can be facilitated by the prior masked presentation of the same stimulus (repetition priming; e.g. Bar & Biederman, 1999). Furthermore, masked priming also occurs when the relation between prime and target is a purely semantic one, such as between two related words (Dehaene, Naccache et al., 1998; Klinger & Greenwald, 1995; Marcel, 1983; see also Merikle et al.). We studied semantic priming with numerical stimuli (Dehaene, Naccache et al., 1998; Koechlin, Naccache, Block, & Dehaene, 1999). When subjects had to decide whether target numbers were larger or smaller than five, the prior presentation of another masked number accelerated the response in direct proportion to its amount of similarity with the target, as measured by numerical distance (Koechlin et al., 1999). Furthermore, the same number-comparison experiment also provided evidence that processing of the prime occurs even beyond this semantic stage to reach motor preparation systems (Dehaene, Naccache et al., 1998). When the instruction specified that targets larger than five should be responded to with the right hand, for instance, primes that were larger than five facilitated a right-hand response, and measures of brain activation demonstrated a significant covert activation of motor cortex prior to the main overt response (see also Eimer & Schlaghecken, 1998; Neumann & Klotz, 1994). Thus, an entire stream of perceptual, semantic and motor processes, specified by giving arbitrary verbal instructions to a normal subject, can occur outside of consciousness.

The number priming experiment also illustrates that it is now feasible to visualize

[3] Unfortunately, signal detection theory provides an imperfect criterion for consciousness. If subjects exhibit a d′ measure that does not differ significantly from zero in a forced-choice stimulus detection or discrimination task, one may conclude that no information about the stimulus was available for conscious processing. Conversely, however, a non-zero d′ measure need not imply consciousness, but may result from both conscious and unconscious influences. Experimental paradigms that partially go beyond this limitation have been proposed (e.g. Jacoby, 1991; Klinger & Greenwald, 1995). We concur with Merikle et al. (this volume), however, in thinking that subjective reports remain the crucial measure when assessing the degree of consciousness (see also Weiskrantz, 1997).

directly the brain areas involved in unconscious processing, without having to rely exclusively on indirect priming measures (Dehaene, Naccache et al., 1998; Morris, Öhman, & Dolan, 1998; Sahraie et al., 1997; Whalen et al., 1998; see also Driver & Vuilleumier and Kanwisher, this volume). In the Whalen et al. (1998) experiment, for instance, subjects were passively looking at emotionally neutral faces throughout. Yet the brief, unconscious presentation of masked faces bearing an emotional expression of fear, relative to neutral masked faces, yielded an increased activation of the amygdala, a brain structure known to be involved in emotional processing. We expect such brain-imaging studies to play an important role in mapping the cerebral networks implicated in unconscious processing, and therefore isolating the neural substrates of consciousness.

3.2. Attention is a prerequisite of consciousness

Experiments with masked primes indicate that some minimal duration and clarity of stimulus presentation are necessary for it to become conscious. However, are these conditions also sufficient? Do all stimuli with sufficient intensity and duration automatically gain access to consciousness? Evidence from brain-lesioned patients as well as normal subjects provides a negative answer. Conditions of stimulation, by themselves, do not suffice to determine whether a given stimulus is or is not perceived consciously. Rather, conscious perception seems to result from an interaction of these stimulation factors with the attentional state of the observer. The radical claim was even made that "there seems to be no conscious perception without attention" (Mack & Rock, 1998, p. ix).

Brain-lesioned patients suffering from hemineglect provide a striking illustration of the role of attentional factors in consciousness (Driver & Mattingley, 1998; see also Driver & Vuilleumier, 2001, this issue). Hemineglect frequently results from lesions of the right parietal region, which is thought to be involved in the orientation of attention towards locations and objects. Neglect patients fail to attend to stimuli located in contralesional space, regardless of their modality of their presentation. The focus of attention seems permanently biased toward the right half of space, and patients behave as if the left half had become unavailable to consciousness. This is seen most clearly in the extinction phenomenon: when two visual stimuli are presented side by side left and right of fixation, the patients report only seeing the stimulus on the right, and appear completely unconscious of the identity or even the presence of a stimulus on the left. Nevertheless, the very same left-hemifield stimulus, when presented in isolation at the same retinal location, is perceived normally. Furthermore, even during extinction, priming measures indicate a considerable amount of covert processing of the neglected stimulus at both perceptual and semantic levels (e.g. McGlinchey-Berroth et al., 1993). Hence, although the cortical machinery for bottom-up processing of left-lateralized stimuli seems to be largely intact and activated during extinction, this is clearly not sufficient to produce a conscious experience; a concomitant attentional signal seems compulsory.

In normal subjects, the role of attention in conscious perception has been the subject of considerable research (see Merikle et al., 1995). While there remains controversy concerning the depth of processing of unattended stimuli, there is no doubt that attention serves as a filter prior to conscious perception (see Driver & Vuilleumier, 2001, this issue). Visual search experiments indicate that, given an array of items, the orienting of attention plays a critical role in determining whether a given item gains access to consciousness (Sperling, 1960; Treisman & Gelade, 1980). Objects that do not fall in an attended region of the visual field cannot be consciously reported. Furthermore, there are systematic parallels between the fate of unattended stimuli and the processing of masked primes. Merikle and Joordens (1997) describe three phenomena (Stroop priming, false recognition, and exclusion failure) in which qualitatively similar patterns of performance are observed in divided attention and in masked priming experiments. They conclude that "perception with and without awareness, and perception with and without attention, are equivalent ways of describing the same underlying process distinction" (p. 219).

Mack and Rock (1998) have investigated a phenomenon called inattentional blindness that clearly illustrates this point. They asked normal subjects to engage in a demanding visual discrimination task at a specific location in their visual field. Then on a single trial, another visual stimulus appeared at a different location. This stimulus clearly had sufficient contrast and duration (typically 200 ms) to be perceptible in isolation, yet the use of a single critical trial and of a distracting task ensured that it was completely unattended and unexpected. Under these conditions a large percentage of subjects failed to report the critical stimulus and continued to deny its presence when explicitly questioned about it. In some experimental conditions, even a large black circle presented for 700 ms in the fovea failed to be consciously perceived! Yet priming measures again indicated that the unseen stimulus was processed covertly. For instance, a word extinguished by inattentional blindness yielded strong priming in a subsequent stem completion task. Such evidence, together with similar observations that supra-threshold visual stimuli fail to be reported during the 'attentional blink' (Luck, Vogel, & Shapiro, 1996; Raymond, Shapiro, & Arnell, 1992; Vogel, Luck, & Shapiro, 1998), and that large changes in a complex visual display fail to be noticed unless they are attended ('change blindness'; e.g. O'Regan, Rensink, & Clark, 1999), support the hypothesis that attention is a necessary prerequisite for conscious perception.[4]

3.3. Consciousness is required for specific mental operations

Given that a considerable amount of mental processing seems to occur uncon-

[4] The notion that attention is required for conscious perception seems to raise a potential paradox: if we can only perceive what we attend to, how do we ever become aware of unexpected information? In visual search experiments, for instance, a vertical line 'pops out' of the display and is immediately detected regardless of display size. How is this possible if that location did not receive prior attention? Much of this paradox dissolves, however, once it is recognized that some stimuli can automatically and unconsciously capture attention (Yantis & Jonides, 1984, 1996). Although we can consciously orient our attention, for instance to search through a display, orienting of attention is also determined by unconscious bottom-up mechanisms that have been attuned by evolution to quickly orient us to salient new features of our environment. Pop-out experiments can be reinterpreted as revealing a fast attraction of attention to salient features.

sciously, one is led to ask what are the computational benefits associated with consciousness. Are there any specific mental operations that are feasible only when one is conscious of performing them? Are there sharp limits on the style and amount of unconscious computation? This issue is obviously crucial if one is to understand the computational nature and the evolutionary advantages associated with consciousness. Yet little empirical research to date bears on this topic. In this section, which is clearly more speculative than previous ones, we tentatively identify at least three classes of computations that seem to require consciousness: durable and explicit information maintenance, novel combinations of operations, and intentional behavior (see also Jack & Shallice, this volume for a similar attempt to identify 'Type-C' processes specifically associated with consciousness).

3.3.1. Durable and explicit information maintenance

The classical experiment by Sperling (1960) on iconic memory demonstrates that, in the absence of conscious amplification, the visual representation of an array of letters quickly decays to an undetectable level. After a few seconds or less, only the letters that have been consciously attended remain accessible. We suggest that, in many cases, the ability to maintain representations in an active state for a durable period of time in the absence of stimulation seems to require consciousness. By 'in an active state', we mean that the information is encoded in the firing patterns of active populations of neurons and is therefore immediately available to influence the systems they connect with. Although sensory and motor information can be temporarily maintained by passive domain-specific buffers such as Sperling's iconic store, with a half-life varying from a few hundreds of milliseconds to a few seconds (auditory information being possibly held for a longer duration than visual information), exponential decay seems to be the rule whenever information is not attended (e.g. Cohen & Dehaene, 1998; Tiitinen, May, Reinikainen, & Naatanen, 1994).

Priming studies nicely illustrate the short-lived nature of unconscious representations. In successful masked priming experiments, the stimulus onset asynchrony (SOA) between prime and target is typically quite short, in the order of 50–150 ms. Experiments that have systematically varied this parameter indicate that the amount of priming drops sharply to a non-significant value within a few hundreds of milliseconds (Greenwald, 1996). Thus, the influence of an unconscious prime decays very quickly, suggesting that its mental representation vanishes dramatically as time passes.[5] This interpretation is supported by single-unit recordings in the monkey infero-temporal (IT) cortex during masked and unmasked presentations of faces (Rolls & Tovee, 1994; Rolls, Tovee, & Panzeri, 1999). A very short and masked visual presentation yields a short-lasting burst of firing (~50 ms) in face-

[5] Some incidental learning and mere exposure experiments have reported unconscious priming effects at a long duration (e.g. Bar & Biederman, 1999; Bornstein & D'Agostino, 1992; Elliott & Dolan, 1998). We interpret these findings as suggesting that even unseen, short-lived stimuli may leave long-lasting latent traces, for instance in the form of alterations in synaptic weights in the processing network. This is not incompatible, however, with our postulate that no active, explicit representation of a prime can remain beyond a few hundred milliseconds in the absence of conscious attentional amplification.

selective cells. However, an unmasked face presented for the same short duration yields a long burst whose duration (up to 350 ms) far exceeds the stimulation period. Physiological and behavioral studies in both humans and monkeys suggest that this ability to maintain information on-line independently of the stimulus presence depends on a working memory system associated with dorsolateral prefrontal regions (Fuster, 1989; Goldman-Rakic, 1987). By this argument, then, the working memory system made available by prefrontal circuitry must be tightly related to the durable maintenance of information in consciousness (e.g. Fuster, 1989; Kosslyn & Koenig, 1992; Posner, 1994; see below).

Another remarkable illustration of the effect of time delays on the ability to maintain active and accurate unconscious representation is provided by studies of the impact of visual illusions on reaching behavior (Aglioti, DeSouza, & Goodale, 1995; Daprati & Gentilucci, 1997; Gentilucci, Chieffi, Deprati, Saetti, & Toni, 1996; Hu, Eagleson, & Goodale, 1999). In the Müller-Lyer and Tichener illusions, although two objects have the same objective length, one of them is perceived as looking shorter due to the influence of contextual cues. Nevertheless, when subjects make a fast reaching movement toward the objects, their finger grip size is essentially unaffected by the illusion and is therefore close to objective size. Hence, the motor system is informed of an objective size parameter which is not available to consciousness, providing yet another instance of unconscious visuo-motor processing. Crucially, however, when one introduces a short delay between the offset of stimuli and the onset of the motor response, grip size becomes less and less accurate and is now influenced by the subjective illusion (Gentilucci et al., 1996). In this situation, subjects have to bridge the gap between stimulus and response by maintaining an internal representation of target size. The fact that they now misreach indicates that the accurate but unconscious information cannot be maintained across a time delay. Again, active information survives a temporal gap only if it is conscious.

3.3.2. Novel combinations of operations

The ability to combine several mental operations to perform a novel or unusual task is a second type of computation that seems to require consciousness. Conflict situations, in which a routine behavior must be inhibited and superseded by a non-automatized strategy, nicely illustrate this point. Merikle, Joordens, and Stolz (1995) studied subjects' ability to control inhibition in a Stroop-like task as a function of the conscious perceptibility of the conflicting information. Subjects had to classify a colored target string as green or red. Each target was preceded by a prime which could be the word GREEN or RED. In this situation, the classical Stroop effect was obtained: responses were faster when the word and color were congruent than when they were incongruent. However, when the prime–target relations were manipulated by presenting 75% of incongruent trials, subjects could strategically take advantage of the predictability of the target from the prime, and became faster on incongruent trials than on congruent trials, thus inverting the Stroop effect. Crucially, this strategic inversion only occurred when the prime was consciously perceptible. No strategic effect was observed when the word prime was masked (Merikle et al.,

1995) or fell outside the focus of attention (Merikle & Joordens, 1997). In this situation, only the classical, automatic Stroop effect prevailed. Thus, the ability to inhibit an automatic stream of processes and to deploy a novel strategy depended crucially on the conscious availability of information.

We tentatively suggest, as a generalization, that the strategic operations which are associated with planning a novel strategy, evaluating it, controlling its execution, and correcting possible errors cannot be accomplished unconsciously. It is noteworthy that such processes are always associated with a subjective feeling of 'mental effort', which is absent during automatized or unconscious processing and may therefore serve as a selective marker of conscious processing (Dehaene, Kerszberg, & Changeux, 1998).[6]

3.3.3. Intentional behavior

A third type of mental activity that may be specifically associated with consciousness is the spontaneous generation of intentional behavior. Consider the case of blindsight patients. Some of these patients, even though they claim to be blind, show such an excellent performance in pointing to objects that some have suggested them as a paradigmatic example of the philosopher's 'zombie' (a hypothetical human being who would behave normally, but lacks consciousness). As noted by Dennett (1992) and Weiskrantz (1997), however, this interpretation fails to take into account a fundamental difference with normal subjects: blindsight patients never spontaneously initiate any visually-guided behavior in their impaired field. Good performance can be elicited only by forcing them to respond to stimulation.

All patients with preserved implicit processing seem to have a similar impairment in using the preserved information to generate intentional behavior. The experimental paradigms that reveal above-chance performance in these patients systematically rely on automatizable tasks (stimulus–response associations or procedural learning) with forced-choice instructions. This is also true for normal subjects in subliminal processing tasks. As noted above, masked priming experiments reveal the impossibility for subjects to strategically use the unconscious information demonstrated by priming effects. Given the large amount of information that has been demonstrated to be available with consciousness, this limitation on subliminal processing is not trivial. Intentionally driven behaviors may constitute an important class of processes accessible only to conscious information.

Introspective speech acts, in which the subject uses language to describe his/her mental life, constitute a particular category of intentional behaviors that relate to conscious processing. Consciousness is systematically associated with the potential ability for the subject to report on his/her mental state. This property of *reportabilityis* so exclusive to conscious information that it is commonly used as an empirical

[6] An important qualification is that even tasks that involve complex series of operations and that initially require conscious effort may become progressively automatized after some practice (e.g. driving a car). At this point, such tasks may proceed effortlessly and without conscious control. Indeed, many demonstrations of unconscious priming involve acquired strategies that required a long training period, such as word reading.

criterion to assess the conscious or unconscious status of an information or a mental state (Gazzaniga et al., 1977; Weiskrantz, 1997).

4. A theoretical framework for consciousness

Once those three basic empirical properties of conscious processing have been identified, can a theoretical framework be proposed for them? Current accounts of consciousness are founded on extraordinarily diverse and seemingly incommensurate principles, ranging from cellular properties such as thalamocortical rhythms to purely cognitive constructions such as the concept of a 'central executive'. Instead of attempting a synthesis of those diverse proposals, we isolate in this section three theoretical postulates that are largely shared, even if they are not always explicitly recognized. We then try to show how these postulates, taken together, converge onto a coherent framework for consciousness: the hypothesis of a global neuronal workspace.

4.1. The modularity of mind

A first widely shared hypothesis is that automatic or unconscious cognitive processing rests on multiple dedicated processors or 'modules' (Baars, 1989; Fodor, 1983; Shallice, 1988). There are both functional and neurobiological definitions of modularity. In cognitive psychology, modules have been characterized by their information encapsulation, domain specificity, and automatic processing. In neuroscience, specialized neural circuits that process only specific types of inputs have been identified at various spatial scales, from orientation-selective cortical columns to face-selective areas. The breakdown of brain circuits into functionally specialized subsystems can be evidenced by various methods including brain imaging, neuropsychological dissociation, and cell recording.

We shall not discuss here the debated issue of whether each postulated psychological module can be identified with a specific neural circuit. We note, however, that the properties of automaticity and information encapsulation postulated in psychology are partially reflected in modular brain circuits. Specialized neural responses, such as face-selective cells, can be recorded in both awake and anesthetized animals, thus reflecting an automatic computation that can proceed without attention. Increasingly refined analyses of anatomical connectivity reveal a channeling of information to specific targeted circuits and areas, thus supporting a form of information encapsulation (Felleman & Van Essen, 1991; Young et al., 1995).

As a tentative theoretical generalization, we propose that *a given process, involving several mental operations, can proceed unconsciously only if a set of adequately interconnected modular systems is available to perform each of the required operations*. For instance, a masked fearful face may cause unconscious emotional priming because there are dedicated neural systems in the superior colliculus, pulvinar, and right amygdala associated with the attribution of emotional valence to faces (Morris, Öhman, & Dolan, 1999). Our hypothesis implies that multiple unconscious operations can proceed in parallel, as long as they do not simultaneously appeal to the same modular systems in contradictory ways. Note that unconscious processing may not be limited to low-level or computationally simple operations. High-level processes may operate unconsciously, as long as they are associated with functional neural pathways either established by evolution, laid down during development, or automatized by learning. Hence, there is no systematic relation between the objective complexity of a computation and the possibility of its proceeding unconsciously. For instance, face processing, word reading, and postural control all require complex computations, yet there is considerable evidence that they can proceed without attention based on specialized neural subsystems. Conversely, computationally trivial but non-automatized operations, such as solving $21 - 8$, require conscious effort.

4.2. The apparent non-modularity of the conscious mind

It was recognized early on that several mental activities cannot be explained easily by the modularity hypothesis (Fodor, 1983). During decision making or discourse production, subjects bring to mind information conveyed by many different sources in a seemingly non-modular fashion. Furthermore, during the performance of effortful tasks, they can temporarily inhibit the automatic activation of some processors and enter into a strategic or 'controlled' mode of processing (Posner, 1994; Schneider & Shiffrin, 1977; Shallice, 1988). Many cognitive theories share the hypothesis that controlled processing requires a distinct functional architecture which goes beyond modularity and can establish flexible links amongst existing processors. It has been called the central executive (Baddeley, 1986), the supervisory attentional system (Shallice, 1988), the anterior attention system (Posner, 1994; Posner & Dehaene, 1994), the global workspace (Baars, 1989; Dehaene, Kerszberg, & Changeux, 1998) or the dynamic core (Tononi & Edelman, 1998).

Here we synthesize those ideas by postulating that, besides specialized processors, the architecture of the human brain also comprises *a distributed neural system or 'workspace' with long-distance connectivity that can potentially interconnect multiple specialized brain areas in a coordinated, though variable manner* (Dehaene, Kerszberg, & Changeux, 1998). Through the workspace, modular systems that do not directly exchange information in an automatic mode can nevertheless gain access to each other's content. The global workspace thus provides a common 'communication protocol' through which a particularly large potential for the combination of multiple input, output, and internal systems becomes available (Baars, 1989).

If the workspace hypothesis is correct, it becomes an empirical issue to determine which modular systems make their contents globally available to others through the workspace. Computations performed by modules that are not interconnected through the workspace would never be able to participate in a conscious content, regardless of the amount of introspective effort (examples may include the brainstem systems for blood pressure control, or the superior colliculus circuitry for gaze control). The vast amounts of information that we can consciously process suggests that at least five main categories of neural systems must participate in the workspace: perceptual circuits that inform about the present state of the environment; motor circuits that allow the preparation and controlled execution of actions; long-term memory circuits that can reinstate past workspace states; evaluation circuits that attribute them a valence in relation to previous experience; and attentional or top-down circuits that selectively gate the focus of interest. The global interconnection of those five systems can explain the subjective unitary nature of consciousness and the feeling that conscious information can be manipulated mentally in a largely unconstrained fashion. In particular, connections to the motor and language systems allow any workspace content to be described verbally or non-verbally ('reportability'; Weiskrantz, 1997).

4.3. Attentional amplification and dynamic mobilization

A third widely shared theoretical postulate concerns the role of attention in gating access to consciousness. As reviewed earlier, empirical data indicate that considerable processing is possible without attention, but that attention is required for information to enter consciousness (Mack & Rock, 1998). This is compatible with Michael Posner's hypothesis of an attentional amplification (Posner, 1994; Posner & Dehaene, 1994), according to which the orienting of attention causes increased cerebral activation in attended areas and a transient increase in their efficiency.

Dehaene, Kerszberg, and Changeux (1998) have integrated this notion within the workspace model by postulating that *top-down attentional amplification is the mechanism by which modular processes can be temporarily mobilized and made available to the global workspace, and therefore to consciousness*. According to this theory, the same cerebral processes may, at different times, contribute to the content of consciousness or not. To enter consciousness, it is not sufficient for a process to have on-going activity; this activity must also be amplified and maintained over a sufficient duration for it to become accessible to multiple other processes. Without such 'dynamic mobilization', a process may still contribute to cognitive performance, but only unconsciously.

A consequence of this hypothesis is the absence of a sharp anatomical delineation of the workspace system. In time, the contours of the workspace fluctuate as different brain circuits are temporarily mobilized, then demobilized. It would therefore be incorrect to identify the workspace, and therefore consciousness, with a fixed set of brain areas. Rather, many brain areas contain workspace neurons with the appropriate long-distance and widespread connectivity, and at any given time only a fraction of these neurons constitute the mobilized workspace. As discussed below, workspace neurons seem to be particularly dense in prefrontal cortices (PFCs) and anterior cingulate (AC), thus conferring those areas a dominant role. However, we see no need to postulate that any single brain area is systematically activated in all conscious states, regardless of their content. It is the style of activation (dynamic long-distance mobilization), rather than its cerebral localization, which characterizes consciousness. This hypothesis therefore departs radically from the notion of a single central 'Cartesian theater' in which conscious information is displayed

(Dennett, 1992). In particular, information that is already available within a modular process does not need to be re-represented elsewhere for a 'conscious audience': dynamic mobilization makes it directly available in its original format to all other workspace processes.

The term 'mobilization' may be misinterpreted as implying the existence of an internal homunculus who decides to successively amplify and then suppress the relevant processes at will. Our view, however, considers this mobilization as a collective dynamic phenomenon that does not require any supervision, but rather results from the spontaneous generation of stochastic activity patterns in workspace neurons and their selection according to their adequacy to the current context (Dehaene, Kerszberg, & Changeux, 1998). Stochastic fluctuations in workspace neurons would result, at the collective neuronal assembly level, in the spontaneous activation, in a sudden, coherent, exclusive and 'auto-catalytic' (self-amplifying) manner, of a subset of workspace neurons, the rest being inhibited. This active workspace state is not completely random, but is heavily constrained and selected by the activation of surrounding processors that encode the behavioral context, goals, and rewards of the organism. In the resulting dynamics, transient self-sustained workspace states follow one another in a constant stream, without requiring any external supervision. Explicit, though still elementary, computer simulations of such 'neuronal Darwinism' are available, illustrating its computational feasibility (Changeux & Dehaene, 1989; Dehaene & Changeux, 1997; Dehaene, Kerszberg, & Changeux, 1998; Friston, Tononi, Reeke, Sporns, & Edelman, 1994).

5. Empirical consequences, reinterpretations, and predictions

The remainder of this paper is devoted to an exploration of the empirical consequences of this theoretical framework. We first examine the predicted structural and dynamical conditions under which information may become conscious. We then consider the consequences of our views for the exploration of the neural substrates of consciousness and its clinical or experimental disruption.

5.1. Structural constraints on the contents of consciousness

An important scientific goal regarding consciousness is to explain why some representations that are encoded in the nervous system are permanently impervious to consciousness. In the present framework, the conscious availability of information is postulated to be determined by two structural criteria which are ultimately grounded in brain anatomy. First, the information must be represented in an active manner in the firing of one or several neuronal assemblies. Second, bidirectional connections must exist between these assemblies and the set of workspace neurons, so that a sustained amplification loop can be established. Cerebral representations that violate either criteria are predicted to be permanently inaccessible to consciousness.

The first criterion – active representation – excludes from the contents of consciousness the enormous wealth of information which is present in the nervous system only in latent form, for instance in the patterns of anatomical connections or in strengthened memory traces. As an example, consider the wiring of the auditory system. We can consciously attend to the spatial location of a sound, but we are oblivious to the cues that our nervous system uses to compute it. One such cue is the small time difference between the time of arrival of sound in the two ears (interaural delay). Interaural delay is coded in a very straightforward manner by the neural connection lengths of the medial superior olive (Smith, Joris, & Yin, 1993). Yet such connectivity information, by hypothesis, cannot reach consciousness.

More generally, the 'active representation' criterion may explain the observation that we can never be conscious of the inner workings of our cerebral processes, but only of their outputs. A classical example is syntax: we can become conscious that a sentence is not grammatical, but we have no introspection on the inner workings of the syntactical apparatus that underlies this judgment, and which is presumably encoded in connection weights within temporal and frontal language areas.

There is, however, one interesting exception to this limit on introspection. Subjects' verbal reports do provide a reliable source of information on a restricted class of processes that are slow, serial, and controlled, such as those involved in solving complex arithmetic problems or the Tower of Hanoi task (Ericcson & Simon, 1993). We propose that what distinguishes such processes is that they are not encoded in hardwired connectivity, but rather are generated dynamically through the serial organization of active representations of current goals, intentions, decisions, intermediate results, or errors. The firing of many prefrontal neurons in the monkey encodes information about the animal's current goals, behavioral plans, errors and successes (e.g. Fuster, 1989). According to the workspace model, such active representations of on-going performance can become available for conscious amplification and communication to other workspace components, explaining, for instance, that we can consciously report the strategic steps that we adopted. The model implies that this is the only situation where we can have reliable conscious access to our mental algorithms. Even then, such access is predicted to be limited. Indeed, when multiplying 32 by 47, we are conscious of our goals, subgoals, main steps (multiplying 2 by 7, then 3 by 7, etc.), and possible errors, but we have no introspection as to how we solve each individual problem.

Our second criterion – bidirectional connectivity with the workspace – implies that some representations, even though they are encoded by an active neuronal assembly, may permanently evade consciousness. This may occur if the connectivity needed to establish a reverberating loop with workspace units is absent or damaged. Consider, for example, the minimal contrast between patients with visual neglect, patients with a retinal scotoma, and normal subjects who all have a blind spot in their retina. Superficially, these conditions have much in common. In all of them, subjects fail to consciously perceive visual stimuli presented at a certain location. Yet the objective processing abilities and subjective reports associated with those visual impairments are strikingly different (see Table 1). Patients with visual neglect typically cannot see stimuli in the neglected part of space, but are not conscious that they are lacking this information. Patients with a retinal scotoma also cannot see in a specific region of their visual field, but they are conscious of their

Table 1
Three classical perceptual conditions in which conscious vision is affected, and their proposed theoretical interpretation. A plus sign indicates an available ability, while a minus sign indicates an absent or deteriorated ability

Condition	Symptoms			Theoretical interpretation	
	Consciousness of visual stimulus	Consciousness of visual impairment	Capacity for unconscious processing	Conscious amplification	Modular processing
Visual neglect	–	–	+	–	+
Retinal scotoma	–	+	–	+	–
Blind spot in normal subjects	–	–	–	–	–

blindness in this region. Finally, all of us have a blind region devoid of photoreceptors in the middle of our retinas, the blind spot, yet we are not conscious of having a hole in our vision.

Table 1 shows how the hypothesis of a conscious mobilization of visual processes through top-down amplification can explain these phenomena. In the case of parietal neglect patients, it has been shown that considerable information about the neglected stimulus is still being actively processed (cf. supra). The lesion, however, is thought to affect parietal circuits involved in spatial attention. According to our framework, this may have the effect of disrupting a crucial component in the top-down amplification of visual information, and therefore preventing this information from being mobilized into the workspace. This predicts that recordings of activity evoked by a neglected stimulus in the intact occipito-temporal visual pathway would reveal a significant short-lived activation (sufficient to underlie residual unconscious processing), but without attentional amplification and with an absence of cross-correlation with other distant areas (correlating with the subjects' inability to bring this information to consciousness). Very recently, data compatible with those predictions have been described (Rees et al., 2000; see also Driver & Vuilleumier, 2001, this issue).

The case of retinal scotomas is essentially symmetrical:[7] subjects lack peripheral visual input, but they have an intact network of cortical areas supporting the attentional amplification of visual information into consciousness. Thus, the information

[7] We do not attempt here a full theoretical treatment of visual scotomas of cortical origin, which are more complex than those of retinal origin. In both types of pathologies, patients are conscious of their visual impairment, presumably because their intact attentional system allows them to detect the absence of visual inputs. However, patients with cortical scotomas sometimes exhibit residual unconscious processing abilities in their blind field (blindsight). One possibility is that those residual abilities rely on subcortical circuits such as the superior colliculus (Sahraie et al., 1997) which, for lack of workspace neurons, would remain permanently inaccessible to consciousness. Alternatively, they may be supported by cortical activity (e.g. in area V5; Zeki & Ffytche, 1998), but of a weakened and transient nature insufficient to establish a sustained closed loop with the workspace and therefore to enter consciousness.

that visual inputs are no longer available in their scotoma can be made available to the workspace and, from there, contact long-term memory and motor intention circuits. Retinal patients can therefore recognize that they are impaired relative to an earlier time period, and they can report it verbally or non-verbally. According to this account, becoming conscious that one is blind occurs when a discrepancy is detected, within the preserved cortical visual representations, between the activity elicited when attending to long-term memory circuits, and the absence of activity elicited when attempting to attend to the outside world. (Note that this would predict that a person blinded from birth would not experience blindness in the same form; being unable to elicit memories of prior seeing, he/she would have a more 'intellectual' or verbal understanding of blindness, perhaps similar to the one that sighted people have.)

Finally, normal subjects' lack of consciousness of their blind spot can be explained by the lack of both perceptual *and* attentional resources for this part of the visual field (Dennett, 1992). Our visual cortex receives no retinal information from the blind spot, but then, neither can we orient attention to this absence of information. We therefore remain permanently unaware of it. It is noteworthy that the presence of the blind spot can only be demonstrated indirectly: closing one eye, we attend to a small object and move it on the retina until we suddenly see it disappear as it passes over the blind spot. In this situation, object-oriented attention is used to detect an anomalous object disappearance. Spatial attention, however, is permanently unable to let us perceive the blind spot as a hole in our visual field.[8]

As should be clear from the above discussion, whether or not a given category of information is accessible to consciousness cannot be decided a priori, but must be submitted to an empirical investigation. Indeed, recent research has begun to reveal brain circuits that seem to be permanently inaccessible to consciousness. For instance, psychophysical experiments indicate that some information about visual gratings, though extracted by V1 neurons, cannot be consciously perceived (He, Cavanagh, & Intriligator, 1996). Likewise, the dorsal occipito-parietal route involved in guiding hand and eye movements makes accurate use of information about object size and shape, of which subjects can be completely unaware (Aglioti et al., 1995; Daprati & Gentilucci, 1997; Gentilucci et al., 1996; Hu et al., 1999). Although such experiments bear on the contents, rather than on the mechanisms, of consciousness, they may provide crucial tests of theories of consciousness (Crick & Koch, 1995). If the present hypothesis is correct, they should always reveal that the unconscious information is either not explicitly encoded in neural firing, or is encoded by neural populations that lack bidirectional connectivity with the workspace.

5.2. *Dynamical constraints on consciousness*

The previous section dealt with permanently inaccessible information. However, information can also be temporarily inaccessible to consciousness for purely dynamical reasons, as seen in masking paradigms. A masked stimulus presented for a very short duration, even if subject to considerable scrutiny, cannot be brought to consciousness; a slightly longer duration of presentation, however, is sufficient to render the prime easily visible.

According to workspace theory, conscious access requires the temporary dynamical mobilization of an active processor into a self-sustained loop of activation: active workspace neurons send top-down amplification signals that boost the currently active processor neurons, whose bottom-up signals in turn help maintain workspace activity. Establishment of this closed loop requires a minimal duration, thus imposing a temporal 'granularity' to the successive neural states that form the stream of consciousness. This dynamical constraint suggests the existence of two thresholds in human information processing, one that corresponds to the minimal stimulus duration needed to cause any differentiated neural activity at all, and another, the 'consciousness threshold', which corresponds to the significantly longer duration needed for such a neural representation to be mobilized in the workspace through a self-sustained long-distance loop. Stimuli that fall in between those two thresholds cause transient changes in neuronal firing and can propagate through multiple circuits (subliminal processing), but cannot take part in a conscious state.

Fig. 1 shows an elementary neural network that illustrates this idea (though it is obviously too simplistic to represent more than a mere aid to intuition). A cascading series of feed-forward networks initially receives a short burst of firing, similar to the phasic response that can be recorded from IT neurons in response to a masked face (Rolls & Tovee, 1994). As seen in Fig. 1, this burst can then propagate through a large number of successive stages, which might be associated with a transformation of the information into semantic, mnemonic or motor codes. At all these levels, however, only a short-lived burst of activity is seen. Although a closed loop involving a distant workspace network is present in the circuit, it takes a longer or stronger stimulus to reliably activate it and to place the network in a long-lasting self-sustained state. Short stimuli cause only a weak, transient and variable activation of the workspace. It is tempting to view such transient firing as a potential neuronal basis for the complex phenomenology of masking paradigms. At intermediate prime durations (40–50 ms), subjects never characterize their perception of the masked prime as 'conscious', but many of them report that they occasionally experience a glimpse of the stimulus that seems to immediately recede from their grasp, thus leaving them unable to describe what they saw.

In a more realistic situation, the chain of subliminal processing need not be as rigid as suggested in Fig. 1. In humans at least, verbal instructions can induce a rapid reorganization of existing processors into a novel chain through top-down amplification and selection (Fig. 2). Could such a dynamic chain also be traversed by unconscious information? The model leads us to answer positively as long as the instruction or context stimulus used to guide top-down selection itself is conscious.

[8] A more thorough discussion of unawareness of the blind spot would require mention of filling-in experiments (e.g. Ramachandran, 1992). The parts of objects or textures that fall in the blind spot seem to be partially reconstructed or 'filled in' further on in the visual system, and the results of this reconstruction process can then be consciously attended. None of those experiments, however, refute the basic fact that we are not conscious of having a hole in our retinas.

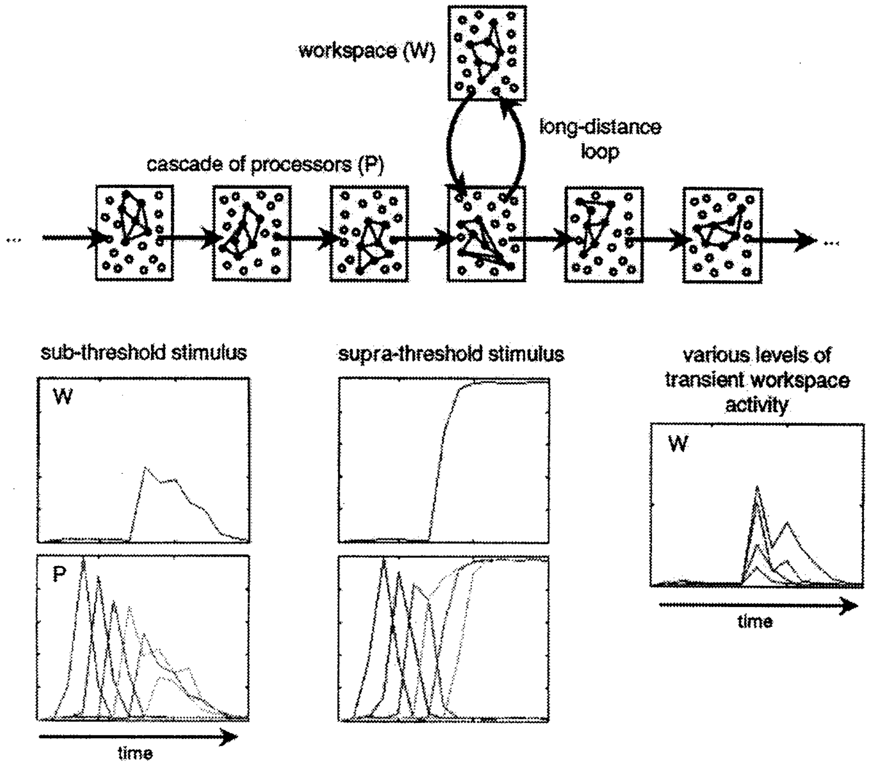

Fig. 1. A simple neural network exhibits dynamical activation patterns analogous to the processing of visual stimuli below and above the consciousness threshold in human subjects. In this minimal scheme, a series of processors are organized in a feed-forward cascade. One of them can also enter in a self-sustained reciprocal loop with a distant workspace network. For simplicity, the evolution of the activation of each assembly (processors and workspace) is modeled by a single McCulloch–Pitts equation, which assumes that activation grows as a non-linear sigmoidal function of the sum of inputs to the assembly, including a small self-connection term, a noise term, and a threshold. Across a wide range of parameters, the same basic findings are reproduced. A short, transient input burst propagates through the processors while causing only a minimal, transient workspace activation (bottom left panels, plots of activation as a function of time). This illustrates how a masked prime, which causes only a transient burst of activation in perceptual neurons of the ventral visual stream (Rolls & Tovee, 1994), can launch an entire stream of visual, semantic and motor processes (Dehaene, Naccache et al., 1998) while failing to establish the sustained coherent workspace activation necessary for consciousness. A slightly more prolonged input causes a sharp transition in activation, with the sudden establishment of a long-lasting activation of both workspace and processor units (middle panels). Thus, the system exhibits a perceptual threshold that stimuli must exceed in order to evoke sustained workspace activation. The right panel illustrates how the very same subthreshold stimulus, on different trials, can evoke transient workspace activity of variable intensity and duration. Similarly, subjects presented with masked primes report a variable phenomenology that ranges from total blindness to a transient feeling that the prime may be on the brink of reportability.

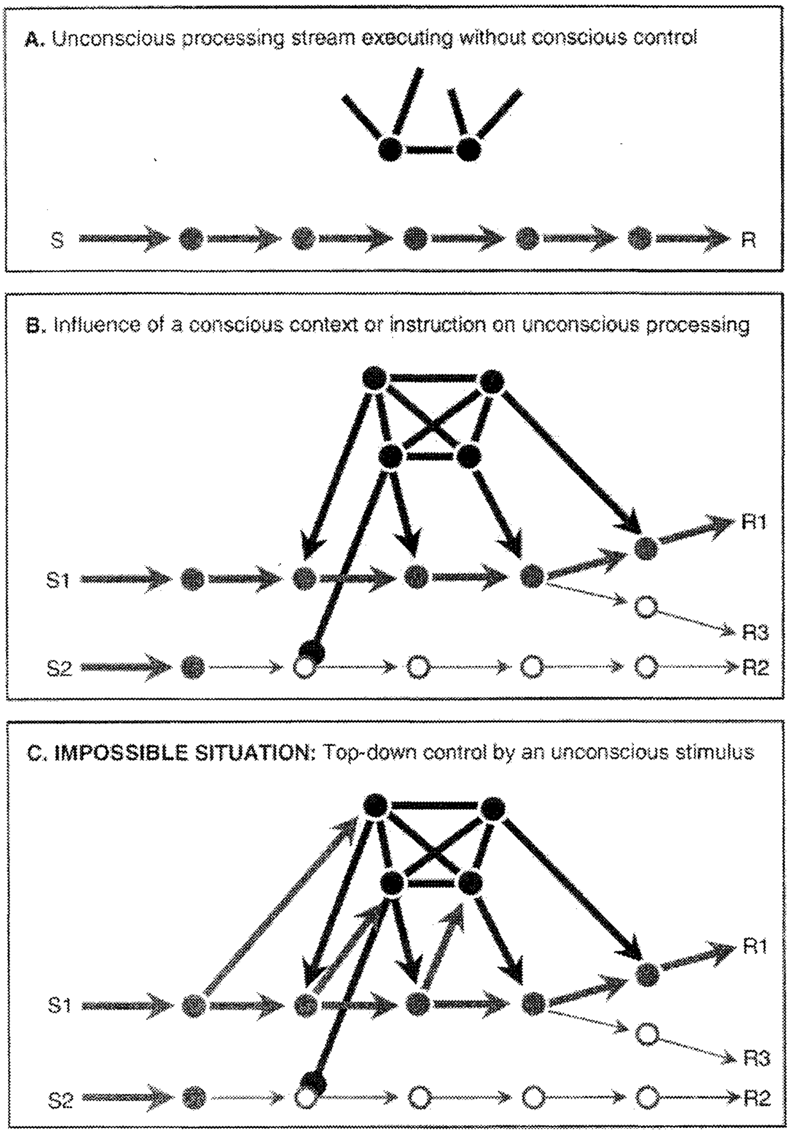
A. Unconscious processing stream executing without conscious control
S
R
B. Influence of a conscious context or instruction on unconscious processing
S1
S2
R1
R3
R2
C. IMPOSSIBLE SITUATION: Top-down control by an unconscious stimulus
S1
S2
R1
R3
R2

Fig. 2. Which tasks may or may not proceed unconsciously? In these schemas, the gray lines represent the propagation of neural activation associated with the unconscious processing of some information, and the black lines the activation elicited by the presently active conscious workspace neurons. The workspace model predicts that one or several automated stimulus–response chains can be executed unconsciously while the workspace is occupied elsewhere (A). Even tasks that require stimulus and processor selection may be executed unconsciously once the appropriate circuit has been set up by a conscious instruction or context (B). However, it should be impossible for an unconscious stimulus to modify processing on a trial-by-trial basis through top-down control (C). A stimulus that contacts the workspace for a duration sufficient to alter top-down control should always be globally reportable.

Thus, the workspace model makes the counter-intuitive prediction that even a complex task that calls for the setting up of a novel non-automatized pathway, once prepared consciously, can be applied unconsciously. Support for this prediction comes from the above-cited masked priming experiments, which indicate that a novel and arbitrary task instruction can be applied to masked primes (Dehaene, Naccache et al., 1998), even when the instruction changes on every trial (Neumann & Klotz, 1994). The same argument leads us to expect that neglect and blindsight patients should be able to selectively attend to stimuli in their blind field and even to react differentially to them as a function of experimenter instructions, all the while denying seeing them.

What should *not* be possible, however, is for an unconscious stimulus itself to control top-down circuit selection (Fig. 2C). Thus, tasks in which the instruction stimulus itself is masked, or is presented in the blind field of a neglect or blindsight patient, should not be applicable unconsciously. Although this prediction has not been tested explicitly, the finding that normal subjects cannot perform exclusion and strategic inversion tasks with masked primes (Merikle et al., 1995) fits with this idea, since those tasks require an active inhibition that varies on a trial-by-trial basis.

5.3. Neural substrates of the contents of consciousness

The new tools of cognitive neuroscience, particularly brain-imaging methods, now make it possible to explore empirically the neural substrates of consciousness. Rather than attempting a exhaustive review of this fast-growing literature, we use the workspace framework to help organize existing results and derive new predictions. The framework predicts that multiple processors encode the various possible contents of consciousness, but that all of them share a common mechanism of coherent brain-scale mobilization. Accordingly, brain-imaging studies of consciousness can be coarsely divided into two categories: studies of the various contents of consciousness, and studies of its shared mechanisms.

We first examine studies of the cerebral substrates of specific contents of consciousness. Brain-imaging studies provide striking illustrations of a one-to-one mapping between specific brain circuits and categories of conscious contents. For instance, Kanwisher (this volume) describes an area, the fusiform face area (FFA), that seems to be active whenever subjects report a conscious visual percept of a face. This includes non-trivial cases, such as rivalry and hallucinations. When subjects are presented with a constant stimulus consisting of a face in one eye and a house in the other, they report alternatively seeing a face, then a house, but not both (binocular rivalry). Likewise, the FFA does not maintain a constant level of activity, but oscillates between high and low levels of activation in tight synchrony with the subjective reports (Tong, Nakayama, Vaughan, & Kanwisher, 1998). Furthermore, when patients hallucinate faces, the FFA activates precisely when subjects report seeing the hallucination (Ffytche et al., 1998). Although the FFA may not be entirely *specific* for faces, as suggested by its involvement in visual expertise for cars or birds

(Gauthier, Skudlarski, Gore, & Anderson, 2000), its activation certainly correlates tightly with conscious face perception.[9]

Another classical example is motion perception. The activation of the human area V5 (or MT) correlates systematically with the conscious perception of motion (Watson et al., 1993), even in non-trivial cases such as visual illusions. For instance, V5 is active when subjects are presented with static paintings of 'kinetic art' that elicit a purely subjective impression of motion (Zeki, Watson, & Frackowiak, 1993). V5 is also active when subjects report experiencing a motion-aftereffect to a static stimulus, and the duration of the activation matches the duration of the illusion (Tootell et al., 1995).

Such experiments indicate a tight correlation between the activation of a specific neural circuit (say, V5) and the subjective report of a conscious content (motion). However, correlation does not imply causation. To establish that a given brain state represents the *causal substrate* of a specific conscious content, rather than a mere *correlate* of consciousness, causality must be established by demonstrating that alterations of this brain state systematically alter subjects' consciousness. In the case of face or motion perception, this can be verified with several methods. Some patients happen to suffer from brain lesions encompassing the FFA or area V5. As predicted, they selectively lose the conscious visual perception of faces or motion (for review, see Young, 1992; Zeki, 1993). Transcranial magnetic stimulation can also be used to temporarily disrupt brain circuits. When applied to area V5, it prevents the conscious perception of motion (Beckers & Zeki, 1995; Walsh, Ellison, Battelli, & Cowey, 1998). Finally, implanted electrodes can be used, not only to disrupt consciousness, but even to change its contents. In the monkey, microstimulation of small populations of neurons in area V5 biases the perception of motion towards the neurons' preferred direction (Salzman, Britten, & Newsome, 1990). While the interpretation of this particular study raises the difficult issue of animal consciousness, similar experiments can also be performed in conscious humans in whom electrodes are implanted for therapeutic purposes. It is known since Penfield that human brain stimulation can elicit a rich conscious phenomenology, including dream-like states. Stimulation can even induce highly specific conscious contents such as feelings of profound depression (Bejjani et al., 1999) or hilarity (Fried, Wilson, MacDonald, & Behnke, 1998). Such experiments, together with the others reported in this section, begin to provide evidence for a form of 'type–type physicalism', in which the major categories of contents of consciousness are causally related, in a systematic manner, to categories of physical brain states that can be reproducibly identified in each subject.

[9] The workspace theory predicts that the same modular processors are involved in both conscious and unconscious processing. Hence, the systematic correlation between FFA activation and conscious face perception is predicted to break down under conditions of subliminal face perception or inattentional blindness for faces. In those situations, there should be a small but significant FFA activation (relative to non-face stimuli) without consciousness. Driver and Vuilleumier (this volume) report results compatible with this prediction (see Rees et al., 2000).

5.4. Neural substrates of the mechanisms of consciousness: prefrontal cortex (PFC), anterior cingulate (AC), and the workspace hypothesis

While the various contents of consciousness map onto numerous, widely distributed brain circuits, the workspace model predicts that all of these conscious states share a common mechanism. The mobilization of any information into consciousness should be characterized by the simultaneous, coherent activation of multiple distant areas to form a single, brain-scale workspace. Areas rich in workspace neurons should be seen as 'active' with brain-imaging methods whenever subjects perform a task which is feasible only in a conscious state, such as one requiring a novel combination of mental operations. Finally, conscious processing should be accompanied by a temporary top-down amplification of activity in neural circuits encoding the current content of consciousness. The cognitive neuroscience literature contains numerous illustrations of these principles, and many of them point to PFC and AC as playing a crucial role in the conscious workspace (Posner, 1994).

5.4.1. Brain imaging of conscious effort

Although this was not their goal, many early brain-imaging experiments studied complex, effortful tasks that presumably cannot be performed without conscious guidance. A common feature of those tasks is the presence of intense PFC and AC activation (Cohen et al., 1997; Pardo, Pardo, Janer, & Raichle, 1990; Paus, Koski, Caramanos, & Westbury, 1998). Importantly, PFC and AC activations do not seem needed for automatized tasks, but appear suddenly whenever an automatized task suddenly calls for conscious control. In the verb generation task, for instance, Raichle et al. (1994) demonstrated that PFC and AC activation is present during initial task performance, vanishes after the task has become automatized, but immediately recovers when novel items are presented. Furthermore, in a variety of tasks, AC activates immediately after errors and, more generally, whenever conflicts must be resolved (Carter et al., 1998; Dehaene, Posner, & Tucker, 1994). In the Wisconsin card sorting test, PFC activates suddenly when subjects have to invent a new behavioral rule (Konishi et al., 1998). Both PFC and AC possess the ability to remain active in the absence of external stimulation, such as during the delay period of a delayed-response task (Cohen et al., 1997), or during internally driven activities such as mental calculation (Chochon, Cohen, van de Moortele, & Dehaene, 1999; Rueckert et al., 1996). Finally, concomitant to PFC and AC activation, a selective attentional amplification is seen in relevant posterior areas during focused-attention tasks (Corbetta, Miezin, Dobmeyer, Smulman, & Petersen, 1991; Posner & Dehaene, 1994).

In those experiments, whose goal was not to study consciousness, conscious control is correlated with a variety of factors such as attention, difficulty, and effort. A stronger test of the neural substrates of consciousness requires contrasting two experimental conditions that differ minimally in all respects except for the subjects' state of consciousness (Baars, 1989). In the last few years, many such paradigms have been developed, contrasting, for instance, levels of anesthesia (Fiset et al., 1999) or implicit versus explicit memory tasks (Rugg et al., 1998). Here we discuss only two examples (but see Frith, Perry, & Lumer, 1999, for review).

5.4.2. Contrasting conscious and unconscious subjects

An elegant approach consists of using exactly the same stimuli and tasks, but in separating a posteriori subjects who became conscious of an aspect of the experimental situation and subjects who did not. Using this method, McIntosh, Rajah, and Lobaugh (1999) showed an increased activation in left PFC (and, to a lesser degree, in bilateral occipital cortices and left thalamus) only in subjects who became aware of a systematic relation between auditory and visual stimuli. Importantly, this activation was accompanied by a major increase in the functional correlation of left PFC with other distant brain regions including the contralateral PFC, sensory association cortices, and cerebellum. This long-distance coherence pattern appeared precisely when subjects became conscious and started to use their conscious knowledge to guide behavior. Using a sequence learning task, Grafton, Hazeltine, and Ivry (1995) also identified a large-scale circuit with a strong focus in the right PFC, which was only observed in subjects who became aware of the presence of a repeated sequence in the stimuli.

5.4.3. Binocular rivalry

In binocular rivalry, subjects are presented with two dissimilar images, one in each eye, and report seeing only one of them at a time. The dominant image, however, alternates with a period of a few seconds. This paradigm is ideal for the study of consciousness because the conscious content changes while the stimulus remains constant. One can therefore study the cerebral activity caused by the dominant image and, a few seconds later, contrast it with the activity when the very same image has become unconscious. Neuronal recordings in awake monkeys trained to report their perception of two rivaling stimuli indicate that early on in visual pathways (e.g. areas V1, V2, V4, and V5), many cells maintain a constant level of firing, indicating that they respond to the constant stimulus rather than to the variable percept (Leopold & Logothetis, 1999). As one moves up in the visual hierarchy, however, an increasing proportion of cells modulate their firing with the reported perceptual alternations. In IT, as many as 90% of the cells respond only to the perceptually dominant image. Brain imaging in humans indicates that IT activity evoked by the dominant image of a rivaling stimulus is indistinguishable from that evoked when the same image is presented alone in a non-rivaling situation (Tong et al., 1998). Importantly, however, IT activity is accompanied by a concomitant widespread increase in AC, prefrontal, and parietal activation (Lumer, Friston, & Rees, 1998). Thus, the point in time when a given image becomes dominant is characterized by a major brain-scale switch in many areas including AC and PFC. This is not merely a coincidental activation of several unrelated neural systems. Rather, posterior and anterior areas appear to transiently form a single large-scale coherent state, as revealed by increases in functional correlation in fMRI (Lumer & Rees, 1999), high-frequency coherences over distances greater than 10 cm in MEG (Srinivasan, Russell, Edelman, & Tononi, 1999), and transient synchronous neuronal firing in V1 neurons (Fries, Roelfsema, Engel, Konig, & Singer, 1997).

In humans, similar increases in coherence and phase synchrony in the EEG and MEG gamma band (30–80 Hz) have been evidenced in a variety of conscious perception paradigms besides rivalry (Rodriguez et al., 1999; Tallon-Baudry & Bertrand, 1999). According to the workspace model, such a long-range coherence should be systematically observed whenever distant areas are mobilized into the conscious workspace. Conversely, however, it cannot be excluded that temporal coding by synchrony is also used by modular processors during non-conscious processing. Thus, increased high-frequency coherence and synchrony are predicted to be a necessary but not sufficient neural precondition for consciousness.

5.4.4. Anatomy and neurophysiology of the conscious workspace

Is the workspace hypothesis compatible with finer-grained brain anatomy and physiology? The requirements of the workspace model are simple. Neurons contributing to this workspace should be distributed in at least five categories of circuits (high-level perceptual, motor, long-term memory, evaluative and attentional networks). During conscious tasks, they should, for a minimal duration, enter into coherent self-sustained activation patterns in spite of their spatial separation. Therefore, they must be tightly interconnected through long axons. Again, all three criteria point to PFC, AC, and areas interconnected with them as playing a major role in the conscious workspace (Fuster, 1989; Posner, 1994; Shallice, 1988).

Consider first the requirement for long-distance connectivity. Dehaene, Kerszberg, and Changeux (1998) noted that long-range cortico-cortical tangential connections, including interhemispheric connections, mostly originate from the pyramidal cells of layers 2 and 3. This suggests that the extent to which an area contributes to the global workspace might be simply related to the fraction of its pyramidal neurons that belong to layers 2 and 3. Those layers, though present throughout the cortex, are particularly thick in dorsolateral prefrontal and inferior parietal cortical structures. A simple prediction, then, is that the activity of those layers may be tightly correlated with consciousness. This could be tested using auto-radiography and other future high-resolution functional imaging methods in primates and in humans, for instance in binocular rivalry tasks.

In monkeys, Goldman-Rakic (1988) and her collaborators have described a dense network of long-distance reciprocal connections linking dorsolateral PFC with premotor, superior temporal, inferior parietal, anterior and posterior cingulate cortices as well as deeper structures including the neostriatum, parahippocampal formation, and thalamus (Fig. 3). This connectivity pattern, which is probably also present in humans, provides a plausible substrate for fast communication amongst the five categories of processors that we postulated contribute primarily to the conscious workspace.[10] Temporal and parietal circuits provide a variety of high-level perceptual categorizations of the outside world. Premotor, supplementary

[10] For her studies of monkey anatomy, Goldman-Rakic (1988) proposes a strictly modular view of PFC, according to which multiple such circuits run in parallel, each of them specialized for a stimulus attribute. The non-modularity of the workspace, however, leads us to postulate that humans may differ from monkeys in showing greater cross-circuit convergence towards common prefrontal and cingulate projection sites as well as heavier reciprocal projections between various sectors of PFC. The degree of cross-circuit convergence in the monkey PFC might also be greater than was initially envisaged (e.g. Rao, Rainer & Miller, 1997)

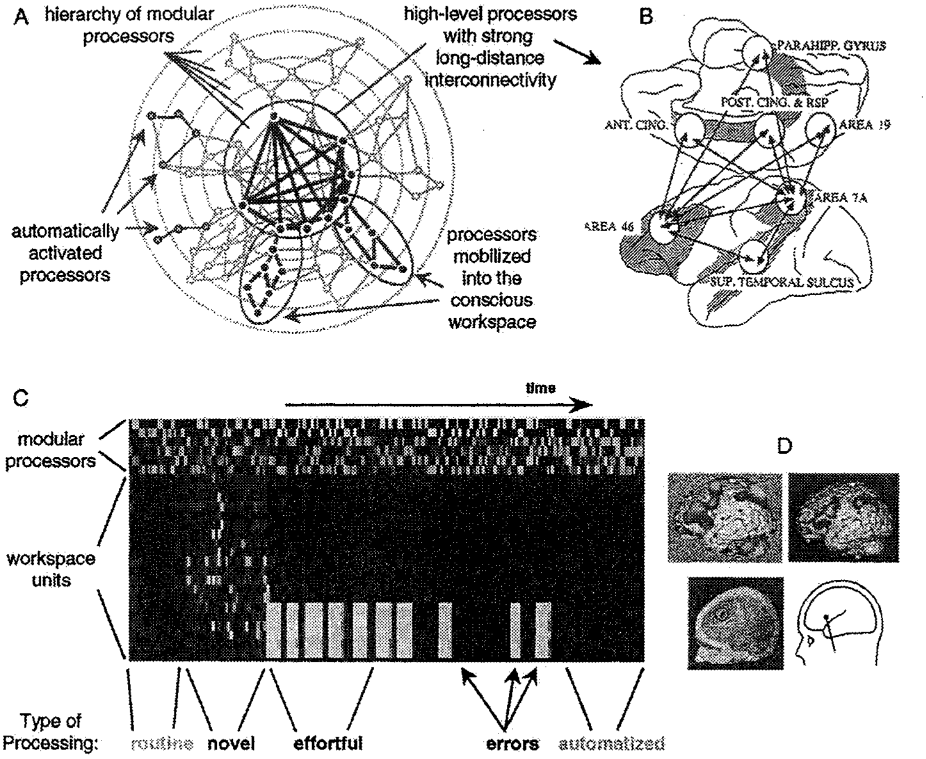

Fig. 3. Neural substrates of the proposed conscious workspace. (A) Symbolic representation of the hierarchy of connections between brain processors (each symbolized by a circle) (after Dehaene, Kerszberg, & Changeux, 1998). Higher levels of this hierarchy are assumed to be widely interconnected by long-distance interconnections, thus forming a global neuronal workspace. An amplified state of workspace activity, bringing together several peripheral processors in a coherent brain-scale activation pattern (black circles), can coexist with the automatic activation of multiple local chains of processors outside the workspace (gray circles). (B) Possible anatomical substrate of the proposed workspace: long-distance network identified in the monkey and linking dorsolateral prefrontal, parietal, temporal, and AC areas with other subcortical targets (from Goldman-Rakic, 1988). (C) Neural dynamics of the workspace, as observed in a neural simulation of a connectivity pattern simplified from (A) (see Dehaene et al., 1998 for details). The matrix shows the activation level of various processor units (top lines) and workspace units (bottom lines) as a function of time. Increased workspace unit activity is seen whenever a novel effortful task is introduced and after errors. Although processor unit activity is continuously present, selective amplification can also be seen when workspace activity is present. (D) Examples of functional neuroimaging tasks activating the postulated workspace network: generation of a novel sequence of random numbers (top left, from Artiges et al., 2000), effortful arithmetic (Chochon et al., 1999), and error processing (left, fMRI data from Carter et al., 1998; right, ERP dipole from Dehaene et al., 1994).

motor and posterior parietal cortices, together with the basal ganglia (notably the caudate nucleus), the cerebellum, and the speech production circuits of the left inferior frontal lobe, allow for the intentional guidance of actions, including verbal reports, from workspace contents. The hippocampal region provides an ability to store and retrieve information over the long term. Direct or indirect connections with orbitofrontal cortex, AC, hypothalamus, amygdala, striatum, and mesencephalic neuromodulatory nuclei may be involved in computing the value or relevance of current representations in relation to previous experience. Finally, parietal and cingulate areas contribute to the attentional gating and shifting of the focus of interest. Although each of these systems, in isolation, can probably be activated without consciousness, we postulate that their coherent activity, supported by their strong interconnectivity, coincides with the mobilization of a conscious content into the workspace.[11]

Physiologically, the neural dynamics of those areas are compatible with the role of consciousness in the durable and explicit maintenance of information over time. The dynamics of prefrontal activity are characterized by periods of long-lasting self-sustained firing, particularly obvious when the animal is engaged in a delayed response task (Fuster, 1989). Sustained firing can also be observed in most cortical regions belonging to the above-mentioned circuit, and prefrontal cooling abolishes those distant sustained responses, suggesting that self-sustained states are a functional property of the integral circuit (Chelazzi, Duncan, Miller, & Desimone, 1998; Fuster, Bauer, & Jervey, 1985; Miller, Erickson, & Desimone, 1996). Lesion studies in monkeys and humans indicate that prefrontal lesions often have little effect on the performance of automatized tasks, but strongly impact on exactly the three types of tasks that were listed earlier as crucially dependent on consciousness: the durable maintenance of explicit information (frontal patients suffer from impairments in delayed response and other working memory tests); the elaboration of novel combinations of operations (frontal patients show perseveration and impaired performance in tasks that call for the invention of novel strategies, such as the Tower of London test; Shallice, 1982); and the spontaneous generation of intentional behavior (patients with frontal or cingulate lesions may perform unintended actions that are induced by the experimenter or the context, as seen in utilization and imitation behavior; Lhermitte, 1983; Shallice, Burgess, Schon, & Baxter, 1989).

Given that workspace neurons must be distributed in widespread brain areas, it should not be surprising that prefrontal lesions do not altogether suppress consciousness, but merely interfere, with variable severity, with some of its functions. However, workspace neurons must be specifically targeted by diffuse neuromodulator systems involved in arousal, the sleep/wake cycle, and reward (Dehaene,

[11] Some have postulated that the hippocampus plays a central role in consciousness (e.g. Clark & Squire, 1998). However, recent evidence suggests that the hippocampus also contributes to implicit learning (Chun & Phelps, 1999) and can be activated by subliminal stimuli such as novel faces (Elliott & Dolan, 1998). Thus, it seems more likely that the hippocampus and surrounding cortices support a modular system that at any given time may or may not be mobilized into the workspace. Gray (1994) proposes that one possible role for the hippocampus, in relation to consciousness, is to serve as a 'novelty detector' that automatically draws attention when the organism is confronted with an unpredicted situation.

Kerszberg, & Changeux, 1998). Impairments of those ascending systems may thus cause a global alteration of conscious workspace activity. For instance, upper brainstem lesions affecting reticular ascending systems frequently cause vigilance impairments or coma (see Parvizi & Damasio, this volume). Interestingly, brain imaging of patients in this vegetative state has revealed a partial preservation of cortical activation, for instance to speech stimuli (de Jong, Willemsen, & Paans, 1997; Menon et al., 1998), indicating that modular processors may still be partially functional at a subliminal level.

6. Final remarks

The present chapter was aimed at introducing the cognitive neuroscience of consciousness and proposing a few testable hypotheses about its cerebral substrates. While we think that a promising synthesis is now emerging, based on the concepts of global workspace, dynamic mobilization, attentional amplification, and frontal circuitry, some readers may feel that those ideas hardly scratch the surface. What about the so-called 'hard problems' posed by concepts such as voluntary action, free will, qualia, the sense of self, or the evolution of consciousness? Our personal view is that, in the present state of our methods, trying to address those problems head-on can actually impede rather than facilitate progress (but see Block, this volume, for a different view). We believe that many of these problems will be found to dissolve once a satisfactory framework for consciousness is achieved. In this conclusion, we examine how such a dissolution might proceed.

6.1. Voluntary action and free will

The hypothesis of an attentional control of behavior by supervisory circuits including AC and PFC, above and beyond other more automatized sensorimotor pathways, may ultimately provide a neural substrate for the concepts of voluntary action and free will (Posner, 1994). One may hypothesize that subjects label an action or a decision as 'voluntary' whenever its onset and realization are controlled by higher-level circuitry and are therefore easily modified or withheld, and as 'automatic' or 'involuntary' if it involves a more direct or hardwired command pathway (Passingham, 1993). One particular type of voluntary decision, mostly found in humans, involves the setting of a goal and the selection of a course of action through the serial examination of various alternatives and the internal evaluation of their possible outcomes. This conscious decision process, which has been partially simulated in neural network models (Dehaene & Changeux, 1991, 1997), may correspond to what subjects refer to as 'exercising one's free will'. Note that under this hypothesis free will characterizes a certain type of decision-making algorithm and is therefore a property that applies at the cognitive or systems level, not at the neural or implementation level. This approach may begin to address the old philosophical issue of free will and determinism. Under our interpretation, a physical system whose successive states unfold according to a deterministic rule can still be described as having free will, if it is able to represent a goal and to estimate the outcomes of its actions before initiating them.[12]

6.2. *Qualia and phenomenal consciousness*

According to the workspace hypothesis, a large variety of perceptual areas can be mobilized into consciousness. At a microscopic scale, each area in turn contains a complex anatomical circuitry that can support a diversity of activity patterns. The repertoire of possible contents of consciousness is thus characterized by an enormous combinatorial diversity: each workspace state is 'highly differentiated' and of 'high complexity', in the terminology of Tononi and Edelman (1998). Thus, the flux of neuronal workspace states associated with a perceptual experience is vastly beyond accurate verbal description or long-term memory storage. Furthermore, although the major organization of this repertoire is shared by all members of the species, its details result from a developmental process of epigenesis and are therefore specific to each individual. Thus, the contents of perceptual awareness are complex, dynamic, multi-faceted neural states that cannot be memorized or transmitted to others in their entirety. These biological properties seem potentially capable of substantiating philosophers' intuitions about the 'qualia' of conscious experience, although considerable neuroscientific research will be needed before they are thoroughly understood.

To put this argument in a slightly different form, the workspace model leads to a distinction between three levels of accessibility. Some information encoded in the nervous system is permanently inaccessible (set I_1). Other information is in contact with the workspace and could be consciously amplified if it was attended to (set I_2). However, at any given time, only a subset of the latter is mobilized into the workspace (set I_3). We wonder whether these distinctions may suffice to capture the intuitions behind Ned Block's (Block, 1995; see also Block, this volume) definitions of phenomenal (P) and access (A) consciousness. What Block sees as a difference in essence could merely be a qualitative difference due to the discrepancy between the size of the potentially accessible information (I_2) and the paucity of information that can actually be reported at any given time (I_3). Think, for instance, of Sperling's experiment in which a large visual array of letters seems to be fully visible, yet only a very small subset can be reported. The former may give rise to the intuition of a rich phenomenological world – Block's P-consciousness – while the latter corresponds to what can be selected, amplified, and passed on to other processes (A-consciousness). Both, however, would be facets of the same underlying phenomenon.

6.3. *Sense of self and reflexive consciousness*

Among the brain's modular processors, some do not extract and process signals

[12] This argument goes back to Spinoza: "men are mistaken in thinking themselves free; their opinion is made up of consciousness of their own actions, and ignorance of the causes by which they are conditioned. Their idea of freedom, therefore, is simply their ignorance of any cause for their actions." (Ethics, II, 35).

from the environment, but rather from the subject's own body and brain. Each brain thus contains multiple representations of itself and its body at several levels (Damasio, 1999). The physical location of our body is encoded in continuously updated somatic, kinesthetic, and motor maps. Its biochemical homeostasis is represented in various subcortical and cortical circuits controlling our drives and emotions. We also represent ourselves as a person with an identity (presumably involving face and person-processing circuits of the inferior and anterior temporal lobes) and an autobiography encoded in episodic memory. Finally, at a higher cognitive level, the action perception, verbal reasoning, and 'theory of mind' modules that we apply to interpret and predict other people's actions may also help us make sense of our own behavior (Fletcher et al., 1995; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Weiskrantz, 1997). All of those systems are modular, and their selective impairment may cause a wide range of neuropsychological deficits involving misperception of oneself and others (e.g. delusions of control, Capgras and Fregoli syndrome, autism). We envisage that the bringing together of these modules into the conscious workspace may suffice to account for the subjective sense of self. Once mobilized into the conscious workspace, the activity of those 'self-coding' circuits would be available for inspection by many other processes, thus providing a putative basis for reflexive or higher-order consciousness.

6.4. *The evolution of consciousness*

Any theory of consciousness must address its emergence in the course of phylogenesis. The present view associates consciousness with a unified neural workspace through which many processes can communicate. The evolutionary advantages that this system confers to the organism may be related to the increased independence that it affords. The more an organism can rely on mental simulation and internal evaluation to select a course of action, instead of acting out in the open world, the lower are the risks and the expenditure of energy. By allowing more sources of knowledge to bear on this internal decision process, the neural workspace may represent an additional step in a general trend towards an increasing internalization of representations in the course of evolution, whose main advantage is the freeing of the organism from its immediate environment.

This evolutionary argument implies that 'having consciousness' is not an all-or-none property. The biological substrates of consciousness in human adults are probably also present, but probably in partial form in other species (or in young children or brain-lesioned patients). It is therefore a partially arbitrary question as to whether we want to extend the use of the term 'consciousness' to them. For instance, several mammals, and possibly even young human children, exhibit greater brain modularity than human adults (Cheng & Gallistel, 1986; Hermer & Spelke, 1994). Yet they also show intentional behavior, partially reportable mental states, some working memory ability – but perhaps no theory of mind. Do they have consciousness, then? Our hope is that once a detailed cognitive and neural theory of the various aspects of consciousness is available, the vacuity of this question will become obvious.

References

Aglioti, S., DeSouza, J. F., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology*, *5* (6), 679–685.

Artiges, E., Salame, P, Recasens, C., Poline, J. B., Attar-Levy, D., De La Raillere, A., Pailere-Martinot, M. L., Danion, J. M., & Martinot, J. L. (2000). Working memory control in patients with schizophrenia: a PET study during a random number generation task. *American Journal of Psychiatry*, *157* (9), 1517–1519.

Baars, B. J. (1989). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Bar, M., & Biederman, I. (1999). Localizing the cortical region mediating visual awareness of object identity. *Proceedings of the National Academy of Sciences USA*, *96* (4), 1790–1793.

Bauer, R. M. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the Guilty Knowledge Test. *Neuropsychologia*, *22* (4), 457–469.

Beckers, G., & Zeki, S. (1995). The consequences of inactivating areas V1 and V5 on visual motion perception. *Brain*, *118* (Pt. 1), 49–60.

Bejjani, B. P., Damier, P., Arnulf, I., Thivard, L., Bonnet, A. M., Dormont, D., Cornu, P., Pidoux, B., Samson, Y., & Agid, Y. (1999). Transient acute depression induced by high-frequency deep-brain stimulation. *New England Journal of Medicine*, *340* (19), 1476–1480.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18* (2), 227–287.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology*, *63* (4), 545–552.

Carter, C. S., Braver, T. S., Barch, D., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*, 747–749.

Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.

Changeux, J. P., & Dehaene, S. (1989). Neuronal models of cognitive functions. *Cognition*, *33*, 63–109.

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology*, *80* (6), 2918–2940.

Cheng, K., & Gallistel, C. R. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, *23*, 149–178.

Chochon, F., Cohen, L., van de Moortele, P. F., & Dehaene, S. (1999). Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience*, *11*, 617–630.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience*, *2* (9), 844–847.

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems: the role of awareness. *Science*, *280* (5360), 77–81.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*, 604–608.

Cohen, L., & Dehaene, S. (1998). Competition between past and present. Assessing and explaining verbal perseverations. *Brain*, *121*, 1641–1659.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Smulman, G. L., & Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape color and speed: functional anatomy by positron emission tomography. *Journal of Neuroscience*, *11*, 2383–2402.

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

Damasio, A. (1999). *The feeling of what happens*. New York: Harcourt Brace.

Daprati, E., & Gentilucci, M. (1997). Grasping an illusion. *Neuropsychologia*, *35* (12), 1577–1582.

Dehaene, S., & Changeux, J. P. (1991). The Wisconsin Card Sorting Test: theoretical analysis and modelling in a neuronal network. *Cerebral Cortex*, *1*, 62–79.

Dehaene, S., & Changeux, J. P. (1997). A hierarchical neuronal network for planning behavior. *Proceedings of the National Academy of Sciences USA*, *94*, 13293–13298.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences USA*, *95*, 14529–14534.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, *5*, 303–305.

de Jong, B. M., Willemsen, A. T., & Paans, A. M. (1997). Regional cerebral blood flow changes related to affective speech presentation in persistent vegetative state. *Clinical Neurology and Neurosurgery*, *99* (3), 213–216.

Dennett, D. C. (1992). *Consciousness explained.* London: Penguin.

Driver, J., & Mattingley, J. B. (1998). Parietal neglect and visual awareness. *Nature Neuroscience*, *1* (1), 17–22.

Driver, J., & Vuilleumier, P. (2001 this issue). Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition*, *79*, 39–88.

Eimer, M., & Schlaghecken, F. (1998). Effects of masked stimuli on motor activation: behavioral and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24* (6), 1737–1747.

Elliott, R., & Dolan, R. J. (1998). Neural response during preference and memory judgments for subliminally presented stimuli: a functional neuroimaging study. *Journal of Neuroscience*, *18* (12), 4697–4704.

Ericcson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data.* Cambridge, MA: MIT Press.

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

Ffytche, D. H., Howard, R. J., Brammer, M. J., David, A., Woodruff, P., & Williams, S. (1998). The anatomy of conscious vision: an fMRI study of visual hallucinations. *Nature Neuroscience*, *1* (8), 738–742.

Fiset, P., Paus, T., Daloze, T., Plourde, G., Meuret, P., Bonhomme, V., Hajj-Ali, N., Backman, S. B., & Evans, A. C. (1999). Brain mechanisms of propofol-induced loss of consciousness in humans: a positron emission tomographic study. *Journal of Neuroscience*, *19* (13), 5506–5513.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Donlan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of theory of mind in story comprehension. *Cognition*, *57*, 109–128.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fried, I., Wilson, C. L., MacDonald, K. A., & Behnke, E. J. (1998). Electric current stimulates laughter. *Nature*, *391* (6668), 650.

Fries, P., Roelfsema, P. R., Engel, A. K., Konig, P., & Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences USA*, *94* (23), 12699–12704.

Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O., & Edelman, G. M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience*, *59*, 229–243.

Frith, C., Perry, R., & Lumer, E. (1999). The neural correlates of conscious experience: an experimental framework. *Trends in Cognitive Science*, *3*, 105–114.

Fuster, J. M. (1989). *The prefrontal cortex.* New York: Raven Press.

Fuster, J. M., Bauer, R. H., & Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research*, *330* (2), 299–307.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119* (Pt. 2), 593–609.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3* (2), 191–197.

Gazzaniga, M. S., LeDoux, J. E., & Wilson, D. H. (1977). Language, praxis, and the right hemisphere: clues to some mechanisms of consciousness. *Neurology*, *27* (12), 1144–1147.

Gentilucci, M., Chieffi, S., Deprati, E., Saetti, M. C., & Toni, I. (1996). Visual illusion and action. *Neuropsychologia, 34* (5), 369–376.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational knowledge. In F. Plum, & V. Mountcastle (Eds.), *Handbook of physiology* (Vol. 5, pp. 373–417). Bethesda, MD: American Physiological Society.

Goldman-Rakic, P. S. (1988). Topography of cognition: parallel distributed networks in primate association cortex. *Annual Review of Neuroscience, 11*, 137–156.

Grafton, S. T., Hazeltine, E., & Ivry, R. (1995). Functional mapping of sequence learning in normal humans. *Journal of Cognitive Neuroscience, 7*, 497–510.

Gray, J. A. (1994). The contents of consciousness: a neuropsychological conjecture. *Behavioral and Brain Sciences, 18*, 659–722.

Greenwald, A. G. (1996). Three cognitive markers of unconscious semantic activation. *Science, 273* (5282), 1699–1702.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383* (6598), 334–337.

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature, 370*, 57–59.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision and visual masking: a survey and appraisal. *Behavioral and Brain Sciences, 9*, 1–23.

Hu, Y., Eagleson, R., & Goodale, M. A. (1999). The effects of delay on the kinematics of grasping. *Experimental Brain Research, 126* (1), 109–116.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

Klinger, M. R., & Greenwald, A. G. (1995). Unconscious priming of association judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21* (3), 569–581.

Koechlin, E., Naccache, L., Block, E., & Dehaene, S. (1999). Primed numbers: exploring the modularity of numerical representations with masked and unmasked semantic priming. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1882–1905.

Köhler, S., & Moscovitch, M. (1997). Unconscious visual processing in neuropsychological syndromes: a survey of the literature and evaluation of models of consciousness. In M. D. Rugg (Ed.), *Cognitive neuroscience* (pp. 305–373). Hove: Psychology Press.

Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K., & Miyashita, Y. (1998). Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience, 1*, 80–84.

Kosslyn, S. M., & Koenig, O. (1992). *Wet mind: the new cognitive neuroscience.* New York: Macmillan.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends in Cognitive Science, 3*, 254–264.

Lhermitte, F. (1983). "Utilization behaviour" and its relation to lesions of the frontal lobe. *Brain, 106*, 237–255.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature, 383* (6601), 616–618.

Lumer, E. D., Friston, K. J., & Rees, G. (1998). Neural correlates of perceptual rivalry in the human brain. *Science, 280*, 1930–1934.

Lumer, E. D., & Rees, G. (1999). Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proceedings of the National Academy of Sciences USA, 96* (4), 1669–1673.

Mack, A., & Rock, I. (1998). *Inattentional blindness.* Cambridge, MA: MIT Press.

Marcel, A. J. (1983). Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognitive Psychology, 15*, 197–237.

McGlinchey-Berroth, R., Milberg, W. P., Verfaellie, M., Alexander, M., & Kilduff, P. (1993). Semantic priming in the neglected field: evidence from a lexical decision task. *Cognitive Neuropsychology, 10*, 79–108.

McIntosh, A. R., Rajah, M. N., & Lobaugh, N. J. (1999). Interactions of prefrontal cortex in relation to awareness in sensory learning. *Science, 284* (5419), 1531–1533.

Menon, D. K., Owen, A. M., Williams, E. J., Minhas, P. S., Allen, C. M., Boniface, S. J., & Pickard, J. D. (1998). Cortical processing in persistent vegetative state. *Lancet*, *352* (9123), 200.

Merikle, P. M. (1992). Perception without awareness: critical issues. *American Psychologist*, *47*, 792–796.

Merikle, P. M., & Joordens, S. (1997). Parallels between perception without attention and perception without awareness. *Consciousness and Cognition*, *6* (2–3), 219–236.

Merikle, P. M., Joordens, S., & Stolz, J. A. (1995). Measuring the relative magnitude of unconscious influences. *Consciousness and Cognition*, *4*, 422–439.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16* (16), 5154–5167.

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, *393*, 467–470.

Morris, J. S., Öhman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences USA*, *96* (4), 1680–1685.

Neumann, O., & Klotz, W. (1994). Motor responses to non-reportable, masked stimuli: where is the limit of direct motor specification. In C. Umiltà, & M. Moscovitch (Eds.), *Conscious and non-conscious information processingAttention and performance* (Vol. XV, pp. 123–150). Cambridge, MA: MIT Press.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature*, *398* (6722), 34.

Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences USA*, *87*, 256–259.

Passingham, R. (1993). *The frontal lobes and voluntary action* (Vol. 21). New York: Oxford University Press.

Paus, T., Koski, L., Caramanos, Z., & Westbury, C. (1998). Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: a review of 107 PET activation studies. *NeuroReport*, *9*, R37–R47.

Penrose, R. (1990). *The emperor's new mind. Concerning computers, minds, and the laws of physics.* London: Vintage Books.

Pöppel, E., Held, R., & Frost, D. (1973). Residual visual function after brain wounds involving the central visual pathways in man. *Nature*, *243* (405), 295–296.

Posner, M. I. (1994). Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Sciences USA*, *91*, 7398–7403.

Posner, M. I., & Dehaene, S. (1994). Attentional networks. *Trends in Neuroscience*, *17*, 75–79.

Raichle, M. E., Fiez, J. A., Videen, T. O., MacLeod, A. K., Pardo, J. V., Fox, P. T., & Petersen, S. E. (1994). Practice-related changes in human brain functional anatomy during non-motor learning. *Cerebral Cortex*, *4*, 8–26.

Ramachandran, V. S. (1992). Filling in the blind spot. *Nature*, *356* (6365), 115.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Intergration of what and where in the primate prefrontal cortex. *Science*, *276* (5313), 821–824.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18* (3), 849–860.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., & Driver, J. (2000). Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain*, *123* (Pt. 8), 1624–1633.

Renault, B., Signoret, J. L., Debruille, B., Breton, F., & Bolgert, F. (1989). Brain potentials reveal covert facial recognition in prosopagnosia. *Neuropsychologia*, *27* (7), 905–912.

Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B., & Varela, F. J. (1999). Perception's shadow: long-distance synchronization of human brain activity. *Nature*, *397* (6718), 430–433.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *257* (1348), 9–15.

Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *Journal of Cognitive Neuroscience*, *11* (3), 300–311.

Rueckert, L., Lange, N., Partiot, A., Appollonio, I., Litvar, I., Le Bihan, D., & Grafman, J. (1996). Visualizing cortical activation during mental calculation with functional MRI. *NeuroImage*, *3*, 97–103.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, *392* (6676), 595–598.

Sahraie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C. R., & Brammer, M. J. (1997). Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences USA*, *94*, 9406–9411.

Salzman, C. D., Britten, K. H., & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, *346* (6280), 174–177.

Schacter, D. L., Buckner, R. L., & Koutstaal, W. (1998). Memory, consciousness and neuroimaging. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *353* (1377), 1861–1878.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. 1. Detection, search, and attention. *Psychological Review*, *84*, 1–66.

Searle, J. R. (1998). How to study consciousness scientifically. *Philosophical Transactions of the Royal Society of London, Series B*, *353*, 1935–1942.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London, Series B*, *298*, 199–209.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shallice, T., Burgess, P. W., Schon, F., & Baxter, D. M. (1989). The origins of utilization behaviour. *Brain*, *112*, 1587–1598.

Silbersweig, D. A., Stern, E., Frith, C. D., Cahill, C., Holmes, A., Grootoonk, S., Seaward, J., McKenna, P., Chua, S. E., Schnoor, L., Jones, T., & Frackowiak, R. S. J. (1995). A functional neuroanatomy of hallucinations in schizophrenia. *Nature*, *378*, 176–179.

Smith, P. H., Joris, P. X., & Yin, T. C. (1993). Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat: evidence for delay lines to the medial superior olive. *Journal of Comparative Neurology*, *331* (2), 245–260.

Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs*, *74*, 1–29.

Srinivasan, R., Russell, D. P., Edelman, G. M., & Tononi, G. (1999). Increased synchronization of neuromagnetic responses during conscious perception. *Journal of Neuroscience*, *19* (13), 5435–5448.

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Science*, *3*, 151–162.

Tiitinen, H., May, P., Reinikainen, K., & Naatanen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature*, *372* (6501), 90–92.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, *21* (4), 753–759.

Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science*, *282* (5395), 1846–1851.

Tootell, R. B. H., Reppas, J. B., Dale, A. M., Look, R. B., Sereno, M. I., Malach, R., Brady, T. J., & Rosen, B. R. (1995). Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. *Nature*, *375*, 139–141.

Tranel, D., & Damasio, A. R. (1985). Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. *Science*, *228* (4706), 1453–1454.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *24* (6), 1656–1674.

Walsh, V., Ellison, A., Battelli, L., & Cowey, A. (1998). Task-specific impairments and enhancements induced by magnetic stimulation of human visual area V5. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *265* (1395), 537–543.

Watson, J. D. G., Myers, R., Frackowiak, R. S. J., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S., & Zeki, S. (1993). Area V5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, *3*, 79–94.

Weiskrantz, L. (1997). *Consciousness lost and found: a neuropsychological exploration*. New York: Oxford University Press.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, *18*, 411–418.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10* (5), 601–621.

Yantis, S., & Jonides, J. (1996). Attentional capture by abrupt onsets: new perceptual objects or visual masking? *Journal of Experimental Psychology: Human Perception and Performance*, *22* (6), 1505–1513.

Young, A. W. (1992). Face recognition impairments. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *335* (1273), 47–54.

Young, M. P., Scannell, J. W., O'Neill, M. A., Hilgetag, C. C., Burns, G., & Blakemore, C. (1995). Non-metric multidimensional scaling in the analysis of neuroanatomical connection data and the organization of the primate cortical visual system. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *348* (1325), 281–308.

Zeki, S. (1993). *A vision of the brain*. London: Blackwell.

Zeki, S., & Ffytche, D. H. (1998). The Riddoch syndrome: insights into the neurobiology of conscious vision. *Brain*, *121* (Pt. 1), 25–45.

Zeki, S., Watson, J. D., & Frackowiak, R. S. (1993). Going beyond the information given: the relation of illusory visual motion to brain activity. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *252* (1335), 215–222.

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14529–14534, November 1998
Neurobiology

A neuronal model of a global workspace in effortful cognitive tasks

STANISLAS DEHAENE*†, MICHEL KERSZBERG‡, AND JEAN-PIERRE CHANGEUX‡

*Institut National de la Santé et de la Recherche Médicale, Unité 334, Service hospitalier Frédéric Joliot, Commissariat à l'énergie atomique, 4 Place du Général Leclerc, 91401 Orsay, France; and ‡Centre National de la Recherche Scientifique, Unité de Recherche Associée 1284, Neurobiologie Moléculaire, Institut Pasteur, 25 Rue du Dr Roux, 75015 Paris, France

Contributed by Jean-Pierre Changeux, September 23, 1998

ABSTRACT **A minimal hypothesis is proposed concerning the brain processes underlying effortful tasks. It distinguishes two main computational spaces: a unique global workspace composed of distributed and heavily interconnected neurons with long-range axons, and a set of specialized and modular perceptual, motor, memory, evaluative, and attentional processors. Workspace neurons are mobilized in effortful tasks for which the specialized processors do not suffice. They selectively mobilize or suppress, through descending connections, the contribution of specific processor neurons. In the course of task performance, workspace neurons become spontaneously coactivated, forming discrete though variable spatio-temporal patterns subject to modulation by vigilance signals and to selection by reward signals. A computer simulation of the Stroop task shows workspace activation to increase during acquisition of a novel task, effortful execution, and after errors. We outline predictions for spatio-temporal activation patterns during brain imaging, particularly about the contribution of dorsolateral prefrontal cortex and anterior cingulate to the workspace.**

We propose a simple hypothesis concerning the neural basis of "making a conscious mental effort." Why are some cognitive tasks performed effortlessly, whereas others require focused attention and conscious control? Mental effort is clearly unrelated to objective measures of computational difficulty: we routinely perform vision and motor control tasks without awareness of the complex underlying information processing, whereas elementary tasks such as solving 37 − 9 call for our attention and conscious effort.

Neurophysiological, anatomical, and brain-imaging studies have revealed that tasks that can be performed effortlessly mobilize well-defined modular cerebral systems specialized for various aspects of sensory-motor processing (1, 2). On the other hand, humans exhibit the capacity to go beyond modularity and flexibly, though effortfully, recombine these specialized cerebral processes in novel ways (3, 4). Once we are conscious of an item, we can readily perform a large variety of operations on it, including evaluation, memorization, action guidance, and verbal report. This impressive ability must be reconciled with the neurobiological fact that there is no single "cardinal area" to which all areas project (5–8).

Here, we propose a formal architecture of distributed neurons with long-distance connectivity that provides a "global workspace" that can potentially interconnect multiple distributed and specialized brain areas in a coordinated, though variable manner, and whose intense mobilization might be associated with a subjective feeling of conscious effort. This minimal scheme extends former attempts to modelize effortful tasks of delayed response (9), card sorting (10), number-processing (11), and planning (12) on the basis of plausible molecular, anatomical, and functional features of the brain. Here, we present simulations of another task, the Stroop task, to explicitly specify a common architectural principle underlying the effortful character of all these tasks, thus providing empirically testable predictions.

THEORETICAL PREMISES

Two Main Computational Spaces. We distinguish two main computational spaces within the brain (Fig. 1). The first is a processing network, composed of a set of parallel, distributed and functionally specialized processors (5) or modular subsystems (6) ranging from primary sensory processors (such as area V1) or unimodal processors (such as area V4), which combine multiple inputs within a given sensory modality, up to heteromodal processors (such as the visuo-tactile neurons in area LIP or the "mirror" neurons in area F5) that extract highly processed categorical or semantic information. Each processor is subsumed by topologically distinct cortical domains with highly specific local or medium-range connections that "encapsulate" information relevant to its function (13).

The second computational space is a global workspace, consisting of a distributed set of cortical neurons characterized by their ability to receive from and send back to homologous neurons in other cortical areas horizontal projections through long-range excitatory axons (which may impinge on either excitatory or inhibitory neurons). Our view is that this population of neurons does not belong to a distinct set of "cardinal" brain areas but, rather, is distributed among brain areas in variable proportions. It is known that long-range cortico-cortical tangential connections, including callosal connections, mostly originate from the pyramidal cells of layers 2 and 3, which give or receive the so-called "association" efferents and afferents. We therefore propose that the extent to which a given brain area contributes to the global workspace would be simply related to the fraction of its pyramidal neurons contributing to layers 2 and 3, which is particularly elevated in von Economo's type 2 (dorsolateral prefrontal) and type 3 (inferior parietal) cortical structures (14). In addition, these cortical neurons establish strong vertical and reciprocal connections, via layer 5 neurons, with corresponding thalamic nuclei, thus contributing both to the stability of workspace activity, for instance via self-sustained circuits and to the direct access to the processing networks (15, 16).

Selective Gating of Workspace Inputs and Outputs. Although a variety of processor areas project to the interconnected set of neurons composing the global workspace, at any given time only a subset of inputs effectively accesses it. We postulate that this gating is implemented by descending modulatory projections from workspace neurons to more peripheral processor neurons. These projections may selectively amplify or extinguish the ascending inputs from processing

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: dlPFC, dorsolateral prefrontal cortex; AC, anterior cingulate.

†To whom reprint requests should be addressed. e-mail dehaene@shfj.cea.fr.

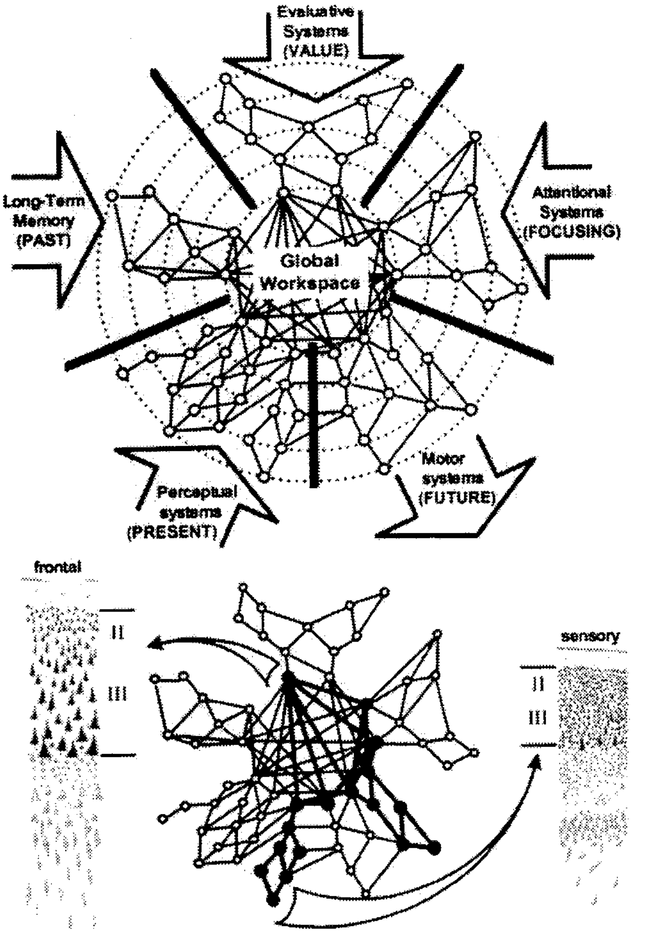

FIG. 1. (*Upper*) Schematic representation of the five main types of processors connected to the global workspace (inspired from ref. 13). (*Lower*) Sample activation during effortful processing; a coherent link between two informationally encapsulated processors is established through the activation of distributed workspace neurons. The long-range workspace connectivity, supported by layer II/III neurons, is more prominent in Von Economo's frontal-type cortex (left) than in sensory-type cortex (right) (14).

neurons, thus mobilizing, at a given time, a specific set of processors in the workspace while suppressing the contribution of others.

Spatio-Temporal Dynamics of Workspace Activity. The global workspace is the seat of a particular kind of "brain-scale" activity states characterized by the spontaneous activation, in a sudden, coherent and exclusive manner, of a subset of workspace neurons, the rest of workspace neurons being inhibited. The entire workspace is globally interconnected in such a way that only one such "workspace representation" can be active at any given time. This all-or-none invasive property distinguishes it from peripheral processors in which, due to local patterns of connections, several representations with different formats may coexist.

A representation that has invaded the workspace may remain active in an autonomous manner and resist changes in peripheral activity. If it is negatively evaluated, or if attention fails, it may however be spontaneously and randomly replaced by another discrete combination of workspace neurons. Functionally, this neural property implements an active "generator of diversity," which constantly projects and tests hypotheses (or prerepresentations) on the outside world (9–12). The dynamics of workspace neuron activity is thus characterized be a constant flow of individual coherent episodes of variable duration.

Content of the Global Workspace. Through their mutual projection to and from workspace neurons, five major categories of processors can be dynamically mobilized and multiply reconfigured (Fig. 1).

Perceptual circuits give the workspace access to the present state of the external world. In humans, perceptual circuits include the object-oriented ventral and lateral areas of the temporal lobes as well as the temporal and inferior parietal areas involved in language comprehension (including Wernicke's area) (13). Thus, the content of any attended object or discourse can access the global workspace.

Motor programming circuits allow the content of the workspace to be used to guide future intentional behavior. A hierarchy of nested circuits implements motor intentions, from the highest level of abstract plans to individual actions, themselves composed of gestures (12, 17). In humans, these circuits include premotor cortex, posterior parietal cortex, supplementary motor area, basal ganglia (notably the caudate nucleus), and cerebellum, as well as the high-level speech production circuits of the left inferior frontal lobe, including Broca's area. Connections of the workspace to motor and language circuits at the higher levels of this hierarchy endow any active representation in the workspace with the property of reportability (18), namely the fact that it can be described or commented upon using words or gestures.

Long-term memory circuits provide the workspace with an access to past percepts and events. Hippocampal and parahippocampal areas play a special role in mediating the storage in and retrieval from long-term memory stores, which are presumably distributed throughout the cortex according to their original content and modality (13).

Evaluation circuits (9, 10, 19, 20) allow representations in the workspace to be associated with a positive or negative value. The main anatomical systems in this respect include the orbitofrontal cortex, anterior cingulate (AC), hypothalamus, amygdala, and ventral striatum as well as the mesocortical catecholaminergic and cholinergic projections to prefrontal cortex. Reciprocal projections allow evaluation circuits to be internally activated by the current workspace content [auto-evaluation (10)] and, conversely, to selectively maintain or change workspace activity according to whether its value is predicted to be positive or negative (9–12, 20).

Attention circuits allow the workspace to mobilize its own circuits independently from the external world. Changes in workspace contents need not necessarily lead to changes in overt behavior but may result in covert attention switches to selectively amplify or attenuate the signals from a subset of processor neurons. Although all descending projections from workspace neurons to peripheral modular processors are important in this selective amplification process, a particular role is played by areas of the parietal lobe in visuo-spatial attention (7, 8, 13).

Global Modulation of Workspace Activation. The state of activation of workspace neurons is assumed to be under the control of global vigilance signals, for instance from mesencephalic reticular neurons. Some of these signals are powerful enough to control major transitions between the awake state (workspace active) and slow-wave sleep (workspace inactive). Others provide graded inputs that modulate the amplitude of workspace activation, which is enhanced whenever novel, unpredicted, or emotionally relevant signals occur, and conversely, drops when the organism is involved in a routine activity.

COMPUTER SIMULATION

To specify the above hypotheses in a computationally explicit manner, a minimal computer simulation of the workspace architecture and dynamics is presented. We are aware that it is necessarily partial and incomplete. We restrict it to the learning and execution of the well known Stroop task (21), which includes both an easy, automatic component and an effortful, attention-demanding component.

Network Architecture and Dynamics. Fig. 2 schematizes the proposed neuronal architecture, composed of excitatory and inhibitory units grouped into different assemblies: input systems, specialized processors, workspace neurons, vigilance, and reward systems. Each assembly is composed of multiple replicas of a basic element comprising an excitatory unit, a gating inhibitory unit, and a processing inhibitory unit. Gating and processing inhibitory units are classical McCulloch–Pitts units whose activity level S_{INH}, ranging from 0 to 1, obeys the update rule $S^i_{INH} = \text{sigmoid}(\Sigma\ w^{i,j}\ S^j)$, where the sigmoid function is defined as $\text{sigmoid}(x) = 1/(1 + e^{-x})$, and the $w^{i,j}$ are the synaptic weights of neurons contacting inhibitory unit i. For simplicity only excitatory units (both local and long-distance) are assumed to make synaptic contact onto inhibitory units.

The activity of excitatory units, S_{EXC}, obeys a modified update rule:

$S^i_{EXC} = \text{sigmoid}(\Sigma^i_{asc}\ \Phi(\Sigma^i_{desc}))$, where $\Sigma^i_{asc} = \Sigma w^{i,j}_{asc}\ S^j$ and $\Sigma^i_{desc} = \Sigma\ w^{i,j}_{desc}\ S^j$.

The weights $w^{i,j}_{asc}$ and $w^{i,j}_{desc}$ can be positive or negative, because inputs to excitatory units may come from excitatory as well as inhibitory units. The equation separates these inputs into two types: descending connections from hierarchically higher assemblies (subscript desc) and ascending or processing inputs (subscript asc), which represent all the other (nondescending) connections that give the neuron its specific functionality (Fig. 2, *Lower Inset*). The monotonic modulating function Φ is chosen as a sigmoid with $\Phi(x) \rightarrow 0$ when $x \rightarrow -\infty$, $\Phi(0) = 1$ and $\Phi(x) \rightarrow 2$ when $x \rightarrow +\infty$. This equation implies that descending signals have a gating effect on lower-level neuronal activity, with attentional amplification if $\Sigma_{desc} > 0$, normal unattended processing if $\Sigma_{desc} = 0$, and attentional suppression if $\Sigma_{desc} < 0$.

For simplicity, only the synaptic weights between two excitatory units are assumed to be modifiable according to a reward-modulated Hebbian rule $\Delta w^{post,pre} = \varepsilon\ R\ S^{pre}\ (2\ S^{post} - 1)$, where R is a reward signal provided after each network response ($R = +1$, correct; $R = -1$, incorrect), pre is the presynaptic unit and post the postsynaptic unit (9). Weights are bounded to remain between 0 and a maximum value (here arbitrarily fixed at 7).

Finally, workspace neuron activity is under the influence of both vigilance and reward signals. The vigilance signal V is treated as having a descending modulatory influence on all workspace neurons according to the above-described gating mechanism. It is updated after each response: if $R > 0$, then $\Delta V = -0.1\ V$, otherwise $\Delta V = 0.5\ (1 - V)$. This rule has the effect of a slowly decreasing vigilance with sharp increases on error trials. The reward signal R influences the stability of workspace activity through a short-term depression or potentiation of synaptic weights (9, 10, 12): if $R < 0$, $S^{pre} > 0.5$ and $S^{post} > 0.5$, then $\Delta w'^{post,pre} = -0.5\ w'^{post,pre}$, otherwise $\Delta w'^{post,pre} = 0.2\ (1 - w'^{post,pre})$, where w' is a short-term multiplier on the excitatory synaptic weight from unit pre to unit post. A plausible molecular implementation of this rule has been proposed in terms of allosteric receptors (9, 10) (Fig. 2, *Upper Inset*). It postulates that the time coincidence of a diffuse reward signal and of a postsynaptic marker of recent neuronal activity transiently shifts the allosteric equilibrium either toward, or from, a desensitized refractory conformation. Through this "chemical Hebb rule," negative reward destabilizes the self-sustaining excitatory connections between currently active workspace neurons, thus initiating a change in workspace activity.

Implementation of the Stroop Tasks. We submitted the network to several versions of the word-color Stroop tasks (21). For this purpose, four input units were dedicated to encoding four color words, four other input units encoded the color of the ink used to print the word, and four internal units corresponded to the four naming responses (Fig. 2). Routine task 1 (color naming) consisted in turning a single color unit on and rewarding the network for turning the corresponding naming unit on. Direct one-to-one connections between color

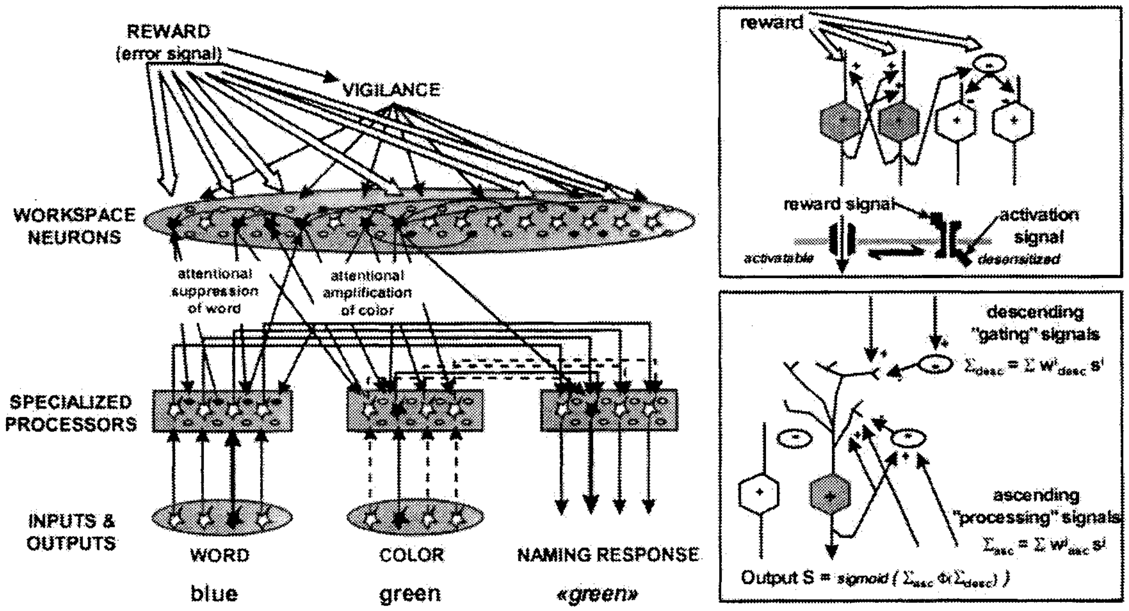
REWARD
(error signal)
VIGILANCE
WORKSPACE NEURONS
attentional suppression of word
attentional amplification of color
SPECIALIZED PROCESSORS
INPUTS & OUTPUTS
WORD
blue
COLOR
green
NAMING RESPONSE
«green»
reward
reward signal
activation signal
activatable
desensitized
descending "gating" signals
$\Sigma_{desc} = \Sigma\ w^i_{desc}\ s^i$
ascending "processing" signals
$\Sigma_{asc} = \Sigma\ w^i_{asc}\ s^i$
Output S = sigmoid ($\Sigma_{asc}\ \Phi(\Sigma_{desc})$)

FIG. 2. Architecture of the simulated network. (*Insets*) The proposed mechanisms for reward-dependent changes in workspace unit activity (*Upper*) and for the interaction of ascending and descending connections to a given area (*Lower*). Although each unit in the simulation presumably represents ≈100 neurons in an actual brain, our scheme epitomizes the basic organization of a cortical column, with intra-columnar recurrent excitation, intra-areal and mid-range excitatory connections providing excitation or inhibition (via intermediate processing inhibitory interneurons), and descending excitatory connections providing upward or downward modulation of activity (via intermediate gating inhibitory interneurons). The network is depicted in a state of activity typical of a correct trial in the effortful Stroop task. Attentional amplification reverses the relation between conflicting word and color inputs by amplifying the weaker color unit activity and suppressing the stronger word unit activity.

and naming units implemented a minimal version of the color naming process. Routine task 2 (word naming with color interference) consisted in turning a word unit on together with another incompatible ink color unit and rewarding the network for turning on the naming unit appropriate to the word, not the ink color. Again, word naming was implemented by direct one-to-one connections from word to naming units. As in previous models of the Stroop test (22, 23), stronger connections were used in the word-to-name pathway than in the color-to-name pathway, corresponding to the greater frequency of word naming in everyday use (21). Finally, the effortful task (color naming with word interference) consisted in providing conflicting word and color inputs, as in task 2, but rewarding the network for turning on the naming unit appropriate to the ink color, not the word.

Connections to and from workspace units were critical for the latter task. A random, patchy connection scheme was used, so that each processor had a Gaussian probability of contacting units in any given region of the workspace, and a similar Gaussian probability of receiving projections from units in the same region (with random initial weights). Note that ascending and descending connections in the model are reciprocal only in a statistical sense: any two processor and workspace units are generally not connected bidirectionally, but any region of the workspace that receives ascending projections from multiple processor units is highly likely to send back descending projections to the same units.

Simulation Results. When placed in routine task 1 (color naming, no interfering word) the network performs correctly with only processor unit activation, using the direct one-to-one connections from color units to name units. Although workspace activity is occasionally observed if vigilance is initially set high, it is clearly not needed. Hence, vigilance quickly drops without impacting on performance (Fig. 3).

Similar results are obtained when the network is submitted to routine task 2 (word naming with color interference). Even though there are now two conflicting inputs, word-to-name connections are stronger than color-to-name connections. Hence, the naming response appropriate to the word is activated faster and more strongly than the one appropriate to the color, which is quickly extinguished by lateral inhibition. Thus, workspace unit activity is not needed for this task either.

When the naive network is then switched to the effortful task (color naming with word interference), an initial series of errors takes place as the network perseverates in applying the routine task 2. The delivery of negative reward leads to an increase in vigilance and to the sudden activation of variable patterns amongst workspace units. The next ≈30 trials can be described as a "search phase." Workspace activation varies in a partly random manner as various response rules are explored. Workspace activation patterns that lead to activating the incorrect response unit are negatively rewarded and tend not to be repeated in subsequent trials. Eventually, the network settles into a stable activation pattern, with a fringe of variability that slowly disappears in subsequent trials. This stable pattern, which leads to correct performance, is characterized by (*i*) preferential descending projections to the excitatory units of the color processing network, thus causing an amplification of the color input, and its transmission to response units; (*ii*) preferential descending projections to the inhibitory gating units of the word processing network, thus causing a suppression of the word input; and (*iii*) strong long-distance excitatory connections amongst active workspace units maintaining the pattern active in the intertrial interval. Across multiple simulations with different initial connectivity, an

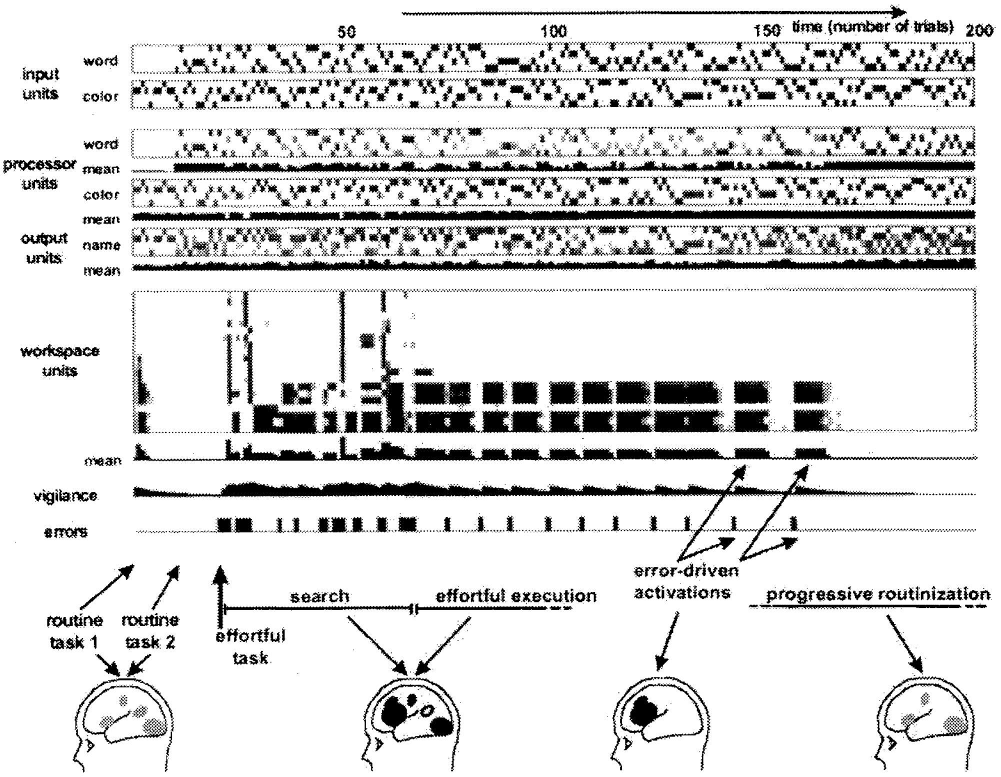

FIG. 3. Temporal dynamics of the simulation in the course of learning the effortful Stroop task; 200 trials were simulated. The Stroop task was introduced without warning after trial 20. Note the selective activation of workspace units with a simultaneous amplification of color processors and a suppression of word processors. Workspace activity is seen in the initial phase of searching for the appropriate response rule (with considerably inter-trial variability), during the effortful execution of the task and following each erroneous response. For illustration purposes, putative brain-imaging correlates of routine and workspace activation are shown (see refs. 24 and 25).

activation pattern with these characteristics was invariably found (after ≈5–50 trials), although its detailed composition varied. The crucial factor here is the patchy distribution of initial connections to and from the workspace, which ensures that sectors of the workspace with the appropriate preexisting connections do indeed exist in the initial state and can be selectively stabilized.

Following the search phase, the network goes through a phase of "effortful task execution" in which workspace activation remains indispensable to correct performance. During this phase, workspace activity remains high even on occasional trials in which the word and ink color information do not conflict. When performance is correct for a series of consecutive trials, vigilance tends to drop. However, any lapse in workspace activation is immediately sanctioned by an error. Each error is immediately followed by an intense reactivation of the workspace. Progressively though, the task becomes routinized as the Hebbian rule applied to processor units tended to increase the color-to-name connections and to decrease the word-to-name connections. Routinization is characterized by increasingly longer periods of correct performance without accompanying workspace activation. Eventually, workspace activation disappears, as the processor network now handles the routinized task by itself.

An interesting property of the network is its ability to maintain an active, sustained state of workspace and processor unit activity for some delay. This is due to the mutually reinforcing excitatory ascending and descending connections between processor and workspace units, together with the excitatory connections within the workspace itself. Once this self-sustained state of activity is established, the descending attentional amplification is often sufficient to maintain processor units active for some duration even when input units are turned off. Hence, the network architecture is adequate to pass delayed-response versions of the routine and effortful tasks in which the response must be postponed after the stimulus has been turned off. It is noteworthy that given this additional delayed-response requirement, even the routine task of color naming now requires workspace activity.

EMPIRICAL TESTS AND PREDICTIONS

Brain Imaging. The key empirical prediction of our hypothesis in the domain of brain imaging is the existence of a strong correlation between cortical areas that are found active in conscious effortful tasks, and areas that possess a strong long-distance cortico-cortical connectivity, presumably associated with dense cortical layers 2 and 3. Brain imaging techniques, once they resolve the transverse laminar distribution of brain activation, might show a differential laminar pattern of activity as a function of whether a given area is recruited for an automatic task or for an effortful task. The global activation of neurons dispersed in multiple cortical areas also might be visualized as a temporary increase in the long-distance coherence of brain activity in electro- and magneto-encephalography (26) or in studies of functional connectivity with functional MRI (27).

We also predict the conditions under which areas rich in workspace neurons should be seen as "active" by using brain-imaging techniques. In our simulation, workspace unit activation exhibits the following properties: (*i*) it is absent during routine tasks; (*ii*) it appears suddenly when a novel, nonroutine task is introduced; (*iii*) it varies semi-randomly during the initial learning of a novel task; (*iv*) it is high and stable during execution of a known but not yet routinized effortful task; (*v*) it decreases during routinization; (*vi*) it resumes sharply following an error; (*vii*) it is present during the delay period of a delayed-response task; and (*viii*) it temporarily mobilizes, in a descending manner, other units involved in specific task components.

Brain-imaging experiments indicate that dorsolateral prefrontal cortex (dlPFC) and AC possess these properties. Both are active in effortful cognitive tasks, including the Stroop test, with a graded level of activation as a function of task difficulty (28–30). With automatization, activation decreases in dlPFC and AC, but it immediately recovers if a novel, nonroutine situation occurs (31). AC activates in tight synchrony with subjects' errors (25, 32). In the Wisconsin card sorting test, dlPFC activates when subjects have to search for a new sorting rule (33). dlPFC and AC possess the ability to remain active in the absence of external stimulation, such as during the delay period of a delayed-response task (28), or during internally driven activities such as mental calculation (34). dlPFC and AC activity also has been found to correlate with subjective conscious perception in various situations in which carefully matched conscious and unconscious conditions were contrasted (35, 36). Finally, concomittent to dlPFC and AC activation, a selective attentional amplification is seen in relevant posterior areas during focused-attention tasks (7, 37).

Workspace activity in our model is concentrated in distinct, localized subsets of neurons that vary with the peripheral processors that must be amplified or suppressed. This is compatible with the evidence for specialization within subregions of AC and dlPFC (30, 38). Our model also posits that effortless or automatic processing should activate specialized processors throughout the cortex without requiring coordination by global workspace neurons. Recent images of brain activity during unconscious processing support this hypothesis (35, 39, 40). In particular, subliminal word stimuli have been shown to cause an entire stream of perceptual, semantic, and motor processes ending up in primary motor cortex (41).

Anatomy and Physiology. Consistent with a privileged contribution of horizontal, long-distance connections in establishing a coherent workspace, a dense network of connections linking dorso-lateral prefrontal and inferior parietal areas to anterior and posterior cingulate, temporal cortices, and parahippocampal cortices has been identified in the monkey (38). It may support the interconnection of the workspace to high-level perceptual, motor, memory, attentional, and evaluation circuits.

The model emphasizes the top-down mobilization of processor neurons by workspace neurons via excitatory descending connections. Such selective amplification or reduction of peripheral neuronal activity has been observed experimentally (42, 43). Because the descending projections are excitatory, they exert their modulatory effect in our model via intermediate connections to a special class of "gating" inhibitory interneurons that have a multiplicative effect on postsynaptic neuronal firing during effortful attentional suppression. These neurons differ from standard "processing inhibitory interneurons," which are the main targets of ascending and horizontal connections, have additive effects on postsynaptic firing, and are active during any type of processing in a given area, automatic as well as effortful. The differential behavior of these two categories of neurons could be established by electrophysiological recordings.

Pharmacology and Molecular Biology. Our theory predicts that workspace neurons are the specific targets of projections from neuronal structures that provide reward and vigilance inputs, presumably via specialize neurotransmitter pathways. Mesocortical dopaminergic neurons and cholinergic pathways, in particular, are known to differentially target prefrontal cortex (13, 44). The decoding of such signals by workspace neurons may be effected by specific subtypes of neurotransmitter receptors (45). Pathological mutations in humans and in genetically modified animals, in which the expression or the physiological properties of a specific subtype of receptor is altered, may thus help decipher the cerebral circuits involved in effortful tasks (46).

CONCLUSIONS

At variance with previous models (9, 10, 22, 23), the proposed neuronal architecture successfully learns the Stroop test without postulating prewired rule-coding units adequate for the task and on the basis of realistic neuronal processes. Our implementation of a global computational workspace operating under conditions of selection by reward does not aim at an exhaustive description of a "conscious workspace" (5). It is limited in scope to features characteristic of effortful tasks, for which it leads to a number of critical predictions, which can be experimentally tested, in particular, with brain-imaging techniques.

The model suffers from shortcomings that should be dealt with in future developments. Although workspace neurons are assumed to be heavily interconnected, they need not be functionally equivalent but rather may be organized in multiple hierarchically nested specialized circuits. An attempt at simulating these nested levels of internal planning was presented in a previous model of the Tower of London task (12). Other important issues include characterization of the variability in the initial connectivity needed to learn multiple tasks (47, 48); the inclusion of novelty detection mechanisms, presumably implemented in the hippocampus, which may serve as input to workspace units (49); and the connection to the workspace of self-representations that might allow the simulated organism to reflect on its own internal processes.

We thank M. Posner and M. Zoli for their comments and the Fondation pour la Recherche Médicale, the College de France, the Association Contre la Myopathie, and the European Union Biotech program for support.

1. Felleman, D. J. & Van Essen, D. C. (1991) *Cereb. Cortex* **1,** 1–47.
2. Cheng, K. & Gallistel, C. R. (1986) *Cognition* **23,** 149–178.
3. Hermer, L. & Spelke, E. S. (1994) *Nature (London)* **370,** 57–59.
4. Fodor, J. A. (1983) *The Modularity of Mind* (MIT Press, Cambridge, MA).
5. Baars, B. J. (1989) *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, Cambridge, MA).
6. Shallice, T. (1988) *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, Cambridge, MA).
7. Posner, M. I. & Dehaene, S. (1994) *Trends Neurosci.* **17,** 75–79.
8. Posner, M. I. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 7398–7403.
9. Dehaene, S. & Changeux, J. P. (1989) *J. Cognit. Neurosci.* **1,** 244–261.
10. Dehaene, S. & Changeux, J. P. (1991) *Cereb. Cortex* **1,** 62–79.
11. Dehaene, S. & Changeux, J. P. (1993) *J. Cognit. Neurosci.* **5,** 390–407.
12. Dehaene, S. & Changeux, J. P. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 13293–13298.
13. Mesulam, M. M. (1998) *Brain* **121,** 1013–1052.
14. Von Economo, C. (1929) *The Cytoarchitectonics of the Human Cerebral Cortex* (Oxford Univ. Press, London).
15. Llinas, R. R. & Paré, D. (1991) *Neuroscience* **44,** 521–535.
16. Munk, M. H., Roelfsema, P. R., Konig, P., Engel, A. K. & Singer, W. (1996) *Science* **272,** 271–274.
17. Jeannerod, M. (1997) *The Cognitive Neuroscience of Action* (Blackwell, Oxford).
18. Weiskrantz, L. (1997) *Consciousness Lost and Found: A Neuropsychological Exploration* (Oxford Univ. Press, New York).
19. Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O. & Edelman, G. M. (1994) *Neuroscience* **59,** 229–243.
20. Schultz, W., Dayan, P. & Montague, P. R. (1997) *Science* **275,** 1593–1599.
21. MacLeod, C. M. (1991) *Psychol. Bull.* **109,** 163–203.
22. Cohen, J. D., Dunbar, K. & McClelland, J. (1990) *Psychol. Rev.* **97,** 332–361.
23. Kimberg, D. Y. & Farah, M. J. (1993) *J. Exp. Psychol. Gen.* **122,** 411–428.
24. Raichle, M. E., Fiesz, J. A., Videen, T. O. & MacLeod, A. K. (1994) *Cereb. Cortex* **4,** 8–26.
25. Dehaene, S., Posner, M. I. & Tucker, D. M. (1994) *Psychol. Sci.* **5,** 303–305.
26. Tononi, G., Srinivasan, R., Russell, D. P. & Edelman, G. M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3198–3203.
27. Friston, K. J., Frith, C. D., Fletcher, P., Liddle, P. F. & Frackowiak, R. S. (1996) *Cereb. Cortex* **6,** 156–164.
28. Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J. & Smith, E. E. (1997) *Nature (London)* **386,** 604–608.
29. Pardo, J. V., Pardo, P. J., Janer, K. W. & Raichle, M. E. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 256–259.
30. Paus, T., Koski, L., Caramanos, Z. & Westbury, C. (1998) *NeuroReport* **9,** R37–R47.
31. Raichle, M. E., Fiez, J. A., Videen, T. O., MacLeod, A. K., Pardo, J. V., Fox, P. T. & Petersen, S. E. (1994) *Cereb. Cortex* **4,** 8–26.
32. Carter, C. S., Braver, T. S., Barch, D., Botvinick, M. M., Noll, D. & Cohen, J. D. (1998) *Science* **280,** 747–749.
33. Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K. & Miyashita, Y. (1998) *Nat. Neurosci.* **1,** 80–84.
34. Roland, P. E. & Friberg, L. (1985) *J. Neurophysiol.* **53,** 1219–1243.
35. Sahraie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C. R. & Brammer, M. J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 9406–9411.
36. Grafton, S. T., Hazeltine, E. & Ivry, R. (1995) *J. Cognit. Neurosci.* **7,** 497–510.
37. Corbetta, M., Miezin, F. M., Dobmeyer, S., Smulman, G. L. & Petersen, S. E. (1991) *J. Neurosci.* **11,** 2383–2402.
38. Goldman-Rakic, P. S. (1988) *Annu. Rev. Neurosci.* **11,** 137–156.
39. Morris, J. S., Öhman, A. & Dolan, R. J. (1998) *Nature (London)* **393,** 467–470.
40. Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B. & Jenike, M. A. (1998) *J. Neurosci.* **18,** 411–418.
41. Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F. & Le Bihan, D. (1998) *Nature (London)* **395,** 597–600.
42. Moran, J. & Desimone, R. (1985) *Science* **229,** 782–784.
43. Motter, B. C. (1994) *J. Neurosci.* **14,** 2178–2189.
44. Descarries, L., Gisiger, V. & Steriade, M. (1997) *Prog. Neurobiol.* **53,** 603–625.
45. Granon, S., Poucet, B., Thinus-Blanc, C., Changeux, J. P. & Vidal, C. (1995) *Psychopharmacology* **119,** 139–144.
46. Picciotto, M. R., Zoli, M., Rimondini, R., Lena, C., Marubio, L. M., Pich, E. M., Fuxe, K. & Changeux, J. P. (1998) *Nature (London)* **391,** 173–177.
47. Changeux, J. P., Courrège, P. & Danchin, A. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 2974–2978.
48. Edelman, G. (1987) *Neural Darwinism* (Basic Books, New York).
49. Gray, J. A. (1994) *Behav. Brain Sci.* **18,** 659–722.

These data show that Dam$^-$ *Salmonella* survive in Peyer's patches of the mouse small intestine for at least 5 days, providing an opportunity for elicitation of a host immune response. Dam$^-$ *Salmonella*, however, were unable to cause disease; they either were unable to invade systemic tissues or were able to invade but could not survive.

DNA adenine methylases are potentially excellent targets for both vaccines and antimicrobials. They are highly conserved in many pathogenic bacteria that cause significant morbidity and mortality, such as *Vibrio cholerae* (*21*), *Salmonella typhi* (*22*), pathogenic *E. coli* (*23*), *Yersinia pestis* (*22*), *Haemophilus influenzae* (*24*), and *Treponema pallidum* (*25*). In addition, because Dam is a global regulator of genes expressed during infection (Fig. 1), Dam$^-$ mutants may ectopically express multiple immunogens that are processed and presented to the immune system. Such ectopic expression could elicit a cross-protective immune response between related bacterial strains that share common epitopes. Finally, because the Dam methylase is essential for bacterial virulence, Dam inhibitors are likely to have broad antimicrobial action, hence Dam is a promising target for antimicrobial drug development.

References and Notes

1. M. G. Marinus, in Escherichia coli *and* Salmonella*: Cellular and Molecular Biology*, F. Neidhardt, Ed. (American Society for Microbiology, Washington, DC, ed. 2, 1996), pp. 782–791.
2. J. A. Roberts *et al.*, *J. Urol.* **133**, 1068 (1985).
3. M. van der Woude, B. Braaten, D. Low, *Trends Microbiol.* **4**, 5 (1996).
4. The *dam102*::Mud-Cm and *mutS121*::Tn*10* alleles (and additional alleles below) were transduced into virulent *S. typhimurium* strain 14028, resulting in strains MT2116 and MT2127, respectively. *dam*Δ*232* (MT2188) was constructed using internal oligonucleotides that serve as polymerase chain reaction primers designed to construct an in-frame 300-bp deletion of defined *dam* sequence. *dcm1*::Km (MT2198) was constructed according to (*26*); the Km resistance determinant is associated with an internal deletion of >600 bp of *dcm* sequence. *lrp31*::Km is a null insertion in the *lrp* gene (MT2126).
5. J. E. LeClerc, B. Li, W. L. Payne, T. A. Cebula, *Science* **274**, 1208 (1996).
6. E. B. Newman, R. T. Lin, R. D'Ari, in (*1*), pp. 1513–1525.
7. B. A. Braaten, X. Nou, L. S. Kaltenbach, D. A. Low, *Cell* **76**, 577 (1994).
8. D. M. Heithoff *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 934 (1997).
9. C. P. Conner, D. M. Heithoff, S. M. Julio, R. L. Sinsheimer, M. J. Mahan, *ibid.* **95**, 4641 (1998).
10. M. J. Mahan, J. M. Slauch, J. J. Mekalanos, *Science* **259**, 666 (1993).
11. M. J. Mahan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 669 (1995).
12. P. A. Gulig *et al.*, *Mol. Microbiol.* **7**, 825 (1993).
13. K. L. Roland, L. E. Martin, C. R. Esther, J. Spitznagel, *J. Bacteriol.* **75**, 4154 (1993).
14. E. Garcia Vescovi, F. C. Soncini, E. A. Groisman, *Cell* **84**, 165 (1996).
15. C. F. Earhart, in (*1*), pp. 1075–1090.
16. E. A. Groisman and F. Heffron, in *Two-Component Signal Transduction*, J. A. Hoch and T. J. Silhavy, Eds. (American Society for Microbiology, Washington, DC, 1995), pp. 319–332.
17. D. M. Heithoff and M. J. Mahan, unpublished data.
18. M. van der Woude, W. B. Hale, D. A. Low, *J. Bacteriol.* **180**, 5913 (1998).
19. W. B. Hale, M. W. van der Woude, D. A. Low, *ibid.* **176**, 3438 (1994).
20. S. Tavazoie and G. M. Church, *Nature Biotechnol.* **16**, 566 (1998).
21. R. Bandyopadhyay and J. Das, *Gene* **140**, 67 (1994).
22. Sanger Centre Web site (www.sanger.ac.uk).
23. F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
24. R. D. Fleischmann *et al.*, *ibid.* **269**, 496 (1995).
25. C. M. Fraser *et al.*, *ibid.* **281**, 375 (1998).
26. S. M. Julio, C. P. Conner, D. M. Heithoff, M. J. Mahan, *Mol. Gen. Genet.* **258**, 178 (1998).
27. D. M. Heithoff *et al.*, *J. Bacteriol.* **181**, 799 (1999).
28. J. M. Slauch and T. Silhavy, *ibid.* **173**, 4039 (1991).
29. C. L. Smith and C. R. Cantor, *Methods Enzymol.* **155**, 449 (1987).
30. M. G. Marinus, A. Poteete, J. A. Arraj, *Gene* **28**, 123 (1984).

31. We thank J. Roth for the *dam102*::Mud-Cm allele, T. Cebula for the *mutS121*::Tn*10* allele, R. Ballester for critically reading the manuscript, and D. Hillyard for constructing the *lrp31* mutant. Supported by NIH grant AI36373 and a Beckman Young Investigator Award (M.J.M.) and NIH grant AI23348 (D.A.L.).

27 January 1999; accepted 7 April 1999

Sources of Mathematical Thinking: Behavioral and Brain-Imaging Evidence

S. Dehaene,[1]* E. Spelke,[2] P. Pinel,[1] R. Stanescu,[1] S. Tsivkin[2]

Does the human capacity for mathematical intuition depend on linguistic competence or on visuo-spatial representations? A series of behavioral and brain-imaging experiments provides evidence for both sources. Exact arithmetic is acquired in a language-specific format, transfers poorly to a different language or to novel facts, and recruits networks involved in word-association processes. In contrast, approximate arithmetic shows language independence, relies on a sense of numerical magnitudes, and recruits bilateral areas of the parietal lobes involved in visuo-spatial processing. Mathematical intuition may emerge from the interplay of these brain systems.

> Will it ever happen that mathematicians will know enough about the physiology of the brain, and neurophysiologists enough of mathematical discovery, for efficient cooperation to be possible? [Jacques Hadamard (*1*)]

Until recently, the only source of information about the mental representations used in mathematics was the introspection of mathematicians. Eloquent support for the view that mathematics relies on visuo-spatial rather than linguistic processes came from Albert Einstein, who stated: "Words and language, whether written or spoken, do not seem to play any part in my thought processes. The psychological entities that serve as building blocks for my thought are certain signs or images, more or less clear, that I can reproduce and recombine at will" (*2*). Many mathematicians report similar experiences (*1*, *3*), but some have stressed the crucial role played by language and other formal symbol systems in mathematics (*4*). Still others have maintained that the critical processes giving rise to new mathematical insights are opaque to consciousness and differ from explicit thought processes (*1*, *3*, *5*).

We address the role of language and visuo-spatial representation in mathematical thinking using empirical methods in cognitive neuroscience. Within the domain of elementary arithmetic, current cognitive models postulate at least two representational formats for number: a language-based format is used to store tables of exact arithmetic knowledge, and a language-independent representation of number magnitude, akin to a mental "number line," is used for quantity manipulation and approximation (*6*, *7*). In agreement with these models, we now demonstrate that exact calculation is language-dependent, whereas approximation relies on nonverbal visuo-spatial cerebral networks.

We first used behavioral experiments in bilinguals to examine the role of language-based representations in learning exact and approximate arithmetic. In one experiment, Russian-English bilinguals were taught a set of exact or approximate sums of two two-digit numbers in one of their two languages (*8*). In the exact addition condition, subjects selected the correct sum from two numerically close numbers. In the approximate addition condition, they were asked to estimate the result and select the closest number. After training, subjects' response times for solving trained problems and novel problems were tested in their two languages. Performance in both tasks improved considerably with training (response times dropped, in

[1]Unité INSERM 334, Service Hospitalier Frédéric Joliot, CEA/DSV, 91401 Orsay Cedex, France. [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: dehaene@shfj.cea.fr

approximation, from 4423 to 2368 ms, and in exact calculation from 4285 to 2813 ms; both $P < 0.001$), regardless of the language in which a problem was trained (response times dropped from 4364 to 2644 ms in Russian and from 4344 to 2534 ms in English). Performance on exact and approximate tasks nevertheless showed different patterns of generalization during the test (Fig. 1). When tested on trained exact addition problems, subjects performed faster in the teaching language than in the untrained language, whether they were trained in Russian or English. This provided evidence that the arithmetic knowledge acquired during training with exact problems was stored in a language-specific format and showed a language-switching cost due to the required internal translation of the arithmetic problem. For approximate addition, in contrast, performance was equivalent in the two languages, providing evidence that the knowledge acquired by exposure to approximate problems was stored in a language-independent form.

Further evidence for contrasting representations underlying exact and approximate arithmetic came from comparisons of performance on trained problems and on novel problems involving similar magnitudes (Fig. 1). For exact addition, subjects performed faster on trained problems, suggesting that each new fact was stored independently of neighboring magnitudes, perhaps as a sequence of words. For approximate addition, performance generalized without cost to novel problems in the same range of magnitudes, providing evidence that new knowledge was stored using a number magnitude format (*9*).

A second experiment extended this phenomenon to more complex arithmetic tasks. A new group of bilinguals was taught two new sets of exact addition facts (two-digit addition with addend 54 or 63), two new exact operations (base 6 and base 8 addition), and two new sets of approximate facts (about cube roots and logarithms in base 2), with one task of each type trained in each of their languages (*10*). Over training, performance again showed large and comparable improvements for all tasks and for both languages. The exact tasks again exhibited large costs for language-switching and for generalization to novel problems for both languages of training, indicating language-specific learning, whereas the approximate tasks showed language- and item-independence (Fig. 1). These results suggest that the teaching of some advanced mathematical facts such as logarithms and cube roots can give rise to a language-independent conceptualization of their magnitude. Exact arithmetic, however, consistently relies on language-based representations (*11*).

To examine whether partly distinct cerebral circuits underlie the observed behavioral dissociation, two functional brain imaging techniques were used, one with high spatial resolution and one with high temporal resolution. Functional magnetic resonance images (fMRI) and event-related potentials (ERPs) were acquired while subjects performed tightly matched exact and approximate addition tasks (Fig. 2) (*12*).

In fMRI, the bilateral parietal lobes showed greater activation for approximation than for exact calculation. The active areas occupied the banks of the left and right intraparietal sulci, extending anteriorily to the depth of the postcentral sulcus and laterally into the inferior parietal lobule (Talaraich coordinates of main peaks: 44, –36, 52, Z = 6.37; 20, –60, 60, Z = 6.03; –56, –44, 52, Z = 5.96; –32, –68, 56, Z = 5.10) (Fig. 3). Activation was also found during approximation in the right precuneus (4, –60, 52, Z = 4.99), left and right precentral sulci (–56, 12, 24, Z = 5.81; 48, 16, 20, Z = 4.80), left dorsolateral prefrontal cortex (–44, 64, 12, Z = 4.46), left superior prefrontal gyrus (–32, 8, 64, Z = 4.75), left cerebellum (–48, –48, –28; Z = 4.74) and left and right thalami (12, –16, 16; Z = 4.43; –20, –8, 16, Z = 4.04).

Most of these areas fall outside of traditional perisylvian language areas (*13*), and are involved instead in various visuo-spatial and analogical mental transformations (*14–16*). Cortices in the vicinity of the intraparietal sulcus, in particular, are active during visually guided hand and eye movements (*15*), mental rotation (*16*), and attention orienting (*17*). Previous brain-imaging experiments also reported strong inferior parietal activation during calculation (*18*), although its functional significance could not be ascertained because of task-difficulty confounds. Here, the parietal activation cannot be attributed to eye movement, hand movement, and attentional or task difficulty artifacts because the approximate and exact tasks were matched in difficulty and in stimulus and response characteristics (*19*). Rather, it is compatible with the hypothesis that approximate

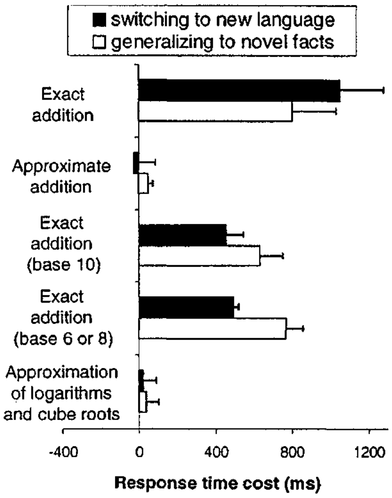

Fig. 1. Generalization of learning new exact or approximate number facts. Mean response times (RTs) to trained problems in the trained language are subtracted from RTs to trained problems in the untrained language (language cost: black bars) and from untrained problems in the trained language (generalization cost: gray bars). In experiment 1 (top two tasks), an analysis of variance on testing RTs indicated significant language-switching [$F(1,3) = 10.53$, $P < 0.05$] and generalization costs [$F(1,3) = 37.64$, $P < 0.01$] for the exact task, but no significant effect for the approximate task (both $Fs < 1$). The interactions of task (exact or approximate) on each cost measure were also significant [respectively, $F(1,6) = 11.10$, $P < 0.02$ and $F(1,6) = 24.71$, $P < 0.005$]. These effects were observed both with testing in English and with testing in Russian, and performance was similar in the two languages (for trained problems, mean RTs were 3445 ms in Russian and 3272 ms in English). In experiment 2 (bottom three tasks), similar analyses of variance indicated language-switching and generalization costs for base 10 addition, $F(1,7) = 24.23$, $P < 0.005$ and $F(1,7) = 28.61$, $P < 0.001$, and for addition in base 6 or 8, $F(1,7) = 304.06$, $P < 0.001$ and $F(1,7) = 71.10$, $P < 0.001$, but not for logarithm or cube root approximation (both $Fs < 1$). The interactions of task (exact or approximate) with each cost measure were also significant [respectively, $F(2,14) = 13.06$, $P < 0.001$ and $F(2,14) = 17.31$, $P < 0.001$]. Again, these effects were observed both with Russian and with English testing, and performance was similar in the two languages (for trained problems, mean RTs were 2639 ms in Russian and 2621 ms in English). Error rates were low in both experiments and were not indicative of speed-accuracy trade-offs.

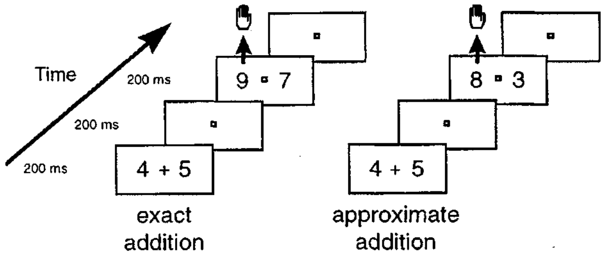

Fig. 2. Design of the tasks used during brain imaging. Subjects fixated continuously on a small central square. On each trial, an addition problem, then two candidate answers were flashed. Subjects selected either the correct answer (exact task) or the most plausible answer (approximate task) by depressing the corresponding hand-held button as quickly as possible. The same addition problems were used in both tasks (*12*).

calculation involves a representation of numerical quantities analogous to a spatial number line, which relies on visuo-spatial circuits of the dorsal parietal pathway.

The converse fMRI contrast of exact calculation relative to approximation revealed a large and strictly left-lateralized activation in the left inferior frontal lobe (–32, 64, 4, Z = 7.53) (*20*). Smaller activation was also found in the left cingulate gyrus (–8, 60, 16, Z = 6.14), left precuneus (–8, –56, 20, Z = 5.64), right parieto-occipital sulcus (20, –80, 28, Z = 5.27), left and right angular gyri (40, –76, 20, Z = 5.07; –44, –72, 36, Z = 4.99), and right middle temporal gyrus (48, –16, 8, Z = 4.68). Previous studies have found left inferior frontal activation during verbal association tasks, including generating a verb associated with a given noun (*21*). Together with the left angular gyrus and left anterior cingulate, these areas may constitute a network involved in the language-dependent coding of exact addition facts as verbal associations (*6*).

Because of their low temporal resolution, fMRI data are compatible with an alternative interpretation that does not appeal to dissociable representations underlying exact and approximate calculation. According to this alternative model, in both the exact and approximate tasks, subjects would compute the exact result using the same underlying representation of numbers. Differences in activation would be entirely due to a subsequent decision stage, during which subjects would select either an exact match or a proximity match to the addition result. The higher temporal resolution afforded by ERPs, however, shows that this alternative interpretation is not tenable. Crucially, ERP to exact and approximate trial blocks already differed significantly during the first 400 ms of a trial, when subjects were viewing strictly identical addition problems and had not yet received the choice stimuli (Fig. 3B). At 216 ms after the onset of the addition problem, ERPs first became more negative for exact rather than for approximate calculation over left inferior frontal electrodes, with a topography compatible with the fMRI activation seen in this same area. Previous ERP and intracranial recordings during the verb generation task also reported a latency of about 220 to 240 ms for the left inferior frontal activation (*22*). Later on in the epoch, starting at 272 ms after addition onset, ERPs became more negative for approximation over bilateral parietal electrodes, with a topography compatible with the bilateral parietal activation seen in fMRI. Thus, the recordings suggest that the two main components of the calculation circuits—the left inferior frontal activation for exact calculation and the bilateral intraparietal activation for approximation—are al-

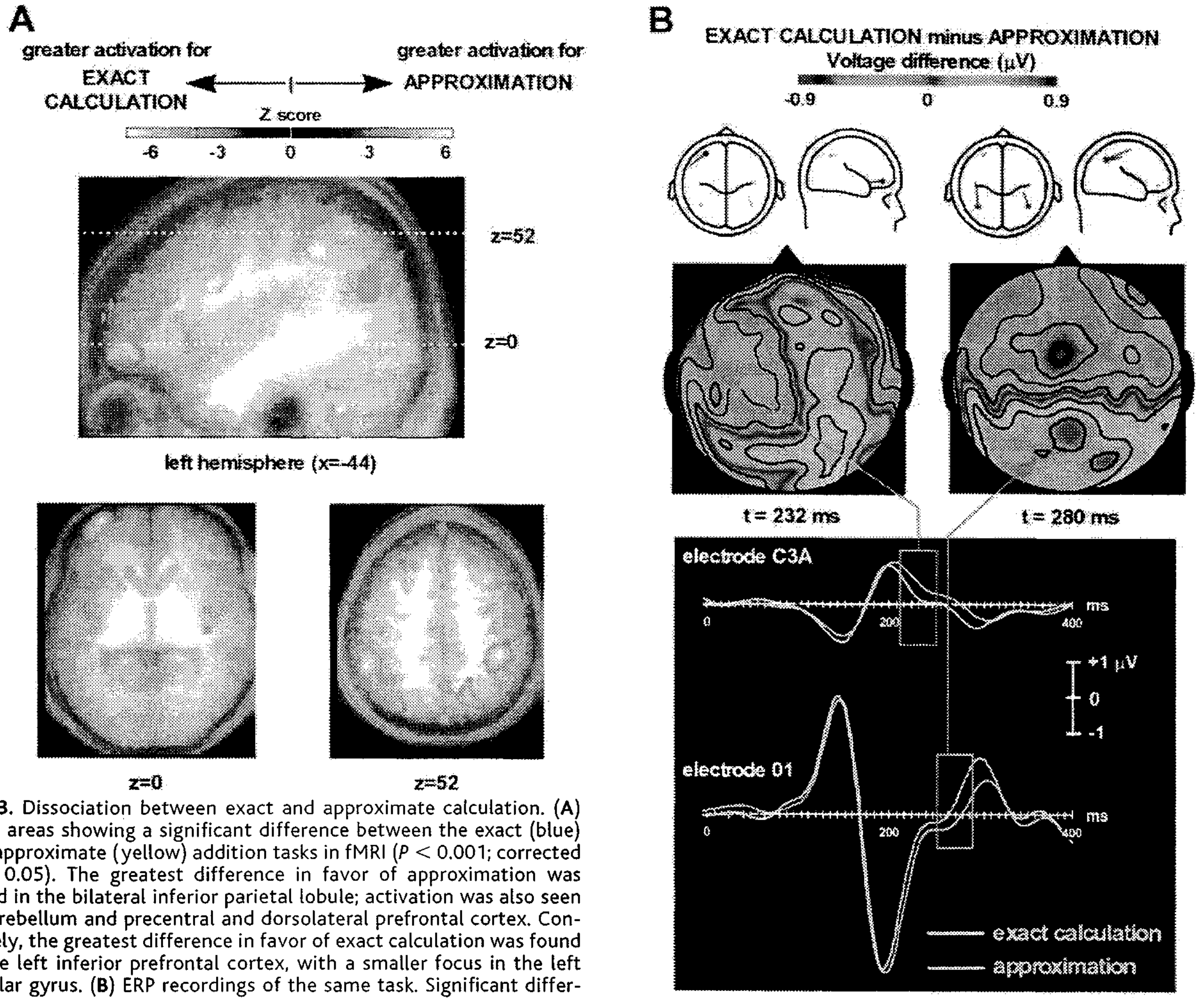

Fig. 3. Dissociation between exact and approximate calculation. (**A**) brain areas showing a significant difference between the exact (blue) and approximate (yellow) addition tasks in fMRI ($P < 0.001$; corrected $P < 0.05$). The greatest difference in favor of approximation was found in the bilateral inferior parietal lobule; activation was also seen in cerebellum and precentral and dorsolateral prefrontal cortex. Conversely, the greatest difference in favor of exact calculation was found in the left inferior prefrontal cortex, with a smaller focus in the left angular gyrus. (**B**) ERP recordings of the same task. Significant differences between exact and approximate calculation were found in two distinct time windows (red rectangles, $P < 0.05$), for which polar maps and dipole models of the corresponding interpolated voltage differences are shown. By 216 to 248 ms after the onset of the addition problem, ERPs were more negative during exact calculation over left inferior frontal sites (left). By 256 to 280 ms, ERPs were more negative during approximate calculation over bilateral parietal sites (right).

ready active at about 230 and 280 ms poststimulus. This demonstrates that the calculation itself, not just the decision, is performed using distinct circuits depending on whether an exact or an approximate result is required.

This conclusion is also strengthened by previous neuropsychological observations of patients with calculation deficits, in whom the lesion localization fits with the present fMRI results. Several lesion sites can cause acalculia (*23*). However, on closer examination, at least two distinct patterns of deficit are found (*24*). Some patients with left parietal lesions exhibit a loss of the sense of numerical quantity (including an inability to decide which number falls between 2 and 4 or whether 9 is closer to 10 or to 5), with a relative preservation of rote language-based arithmetic such as multiplication tables (*24, 25*). Conversely, aphasia following left-hemispheric brain damage can be associated with a selective impairment of rote arithmetic and a preserved sense of quantity, including proximity and larger-smaller relations between numbers (*24*). Particularly relevant to the present work is the case of a severely aphasic and alexic patient with a large left-hemispheric lesion who could not decide whether 2 + 2 was 3 or 4, indicating a deficit for exact addition, but consistently preferred 3 over 9, indicating preserved approximation (*26*). Thus, lesion data confirm that distinct circuits underlie the sense of quantity and knowledge of rote arithmetic facts.

In conclusion, our results provides grounds for reconciling the divergent introspection of mathematicians by showing that even within the small domain of elementary arithmetic, multiple mental representations are used for different tasks. Exact arithmetic puts emphasis on language-specific representations and relies on a left inferior frontal circuit also used for generating associations between words. Symbolic arithmetic is a cultural invention specific to humans, and its development depended on the progressive improvement of number notation systems (*27*). Many other domains of mathematics, such as the calculus, also may depend critically on the invention of an appropriate mathematical language (*28*).

Approximate arithmetic, in contrast, shows no dependence on language and relies primarily on a quantity representation implemented in visuo-spatial networks of the left and right parietal lobes. An interesting, though clearly speculative, possibility, is that this language-independent representation of numerical quantity is related to the preverbal numerical abilities that have been independently established in various animals species (*29*) and in human infants (*30*). Together, these results may indicate that the nonverbal representation that underlies the human sense of numerical quantities has a long evolutionary history, a distinct developmental trajectory, and a dedicated cerebral substrate (*31*). In educated humans, it could provide the foundation for an integration with language-based representations of numbers. Much of advanced mathematics may build on this integration.

References and Notes

1. J. Hadamard, *An Essay on the Psychology of Invention in the Mathematical Field* (Princeton Univ. Press, Princeton, NJ, 1945).
2. A. Einstein, cited in (*1*), p. 142.
3. L. E. J. Brouwer, *Cambridge Lectures on Intuitionism* (Cambridge Univ. Press, Cambridge, 1981).
4. N. Bourbaki, *J. Symb. Logic* **14**, 1 (1948); D. Hilbert and W. Ackermann, *Principles of Mathematical Logic* (Chelsea, New York, 1950); A. N. Whitehead and B. Russell, *Principia Mathematica* (Cambridge Univ. Press, Cambridge, 1910).
5. H. Poincaré, *Science and Hypothesis* (Walter Scott Publishing Co., London, 1907).
6. J. I. D. Campbell and J. M. Clark, *J. Exp. Psychol. Gen.* **117**, 204 (1988); J. I. D. Campbell, *Cognition* **53**, 1 (1994); S. Dehaene, *ibid.* **44**, 1 (1992).
7. S. Dehaene and L. Cohen, *Math Cogn.* **1**, 83 (1995).
8. Participants were three female and five male bilingual Russian-English speakers, aged 18 to 32 years (mean age, 22.5 years), who began to learn English at a mean age of 15.3 years, had been in the United States for an average of 4.9 years and demonstrated fluent comprehension of both Russian and English on formal testing. All were undergraduate or graduate students, and all performed accurately on a variety of elementary arithmetic problems administered in Russian and in English during a pretest. Participants were trained on 12 sums of two two-digit numbers, totaling between 47 and 153. On each trial, an addition problem and two candidate answers were presented on a computer screen in word form, either in English or in Russian. Subjects selected one of the two answers, which appeared left and right of center, by pressing a corresponding key with the left or right hand. For exact addition, the candidate answers were the exact answer and a distractor in which the tens digit was off by one unit. For approximate addition, they were the multiple of ten closest to the correct sum, and another multiple off by 30 units. Each subject participated in 2 days of training with a fixed language and task (for example, exact addition in Russian) which was randomized across subjects (six repetitions of the 12 problems per day). Subjects also were trained for 2 days in a multiplication task (E. Spelke and S. Tsiukin, data not shown) with the same range of numbers in their other language, thus equalizing exposure to the two languages. On the fifth day, subjects were tested twice on each trained problem and twice on 12 similar untrained problems. Testing was done both in the original language of training and in the other language in different blocks.
9. The fact that subjects can estimate the solution of simple addition problems does not necessarily entail that the underlying mental representations are approximate. The present experiments only allow us to conclude that these representations are language-independent and encode numerical proximity.
10. Participants were four male and four female native Russian speakers, aged 18 to 24 years (mean age, 19.8 years), who were introduced to English at a mean age of 15.4 years, had lived in the United States for an average of 3.8 years, and fluently comprehended both Russian and English. All were undergraduate and graduate students who performed accurately in Russian and English pretests of elementary arithmetic skills. Each subject was trained in one language on 12 two-digit base-10 addition problems with addend of 54, 12 base-6 addition problems with two- to three-digit answers, and 12 cube-root estimation problems for numbers under 5000. The same subject was trained in the other language on 12 two-digit base-10 addition problems with addend of 63, 12 base-8 addition problems with two- to three-digit answers, and 12 base-2 logarithm estimation problems for numbers under 8500. Subjects received 2 days of training in each language, with six training trials per problem per training day. Order of training problems and pairings of problems and training languages were counterbalanced across subjects. Subjects received 1 day of testing in each of their languages on all the trained problems, plus an equal number of novel problems within the same six categories (two test trials per problem per testing day).
11. Anecdotal reports have suggested that when bilinguals calculate, they often revert to the original language in which they acquired arithmetic facts [P. A. Kolers, *Sci. Am.* **218**, 78 (March 1968); B. Shanon, *New Ideas Psychol.* **2**, 75 (1984)]. Furthermore, bilinguals solve arithmetic problems with greater speed and accuracy when the problems are presented in their first language [C. Frenck-Mestre and J. Vaid, *Mem. Cogn.* **21**, 809 (1993); L. G. Marsh and R. H. Maki, *ibid.* **4**, 216 (1976); L. McClain and J. Y. Shih Huang, *ibid.* **10**, 591 (1982)]. These observations, however, might simply reflect easier word comprehension and production processes in the first language, rather than a language-dependent encoding of arithmetic knowledge itself [M. McCloskey, P. Macaruso, T. Whetstone, in *The Nature and Origins of Mathematical Skills*, J. I. D. Campbell, Ed. (Elsevier, Amsterdam, 1992), pp. 493–537]. Our results, by contrast, showed a language-switching cost for exact calculation regardless of whether training was in the first or second language. Indeed, the cost of switching from the subjects' first language (Russian) to their second language was no greater than the cost of switching in the reverse direction (519 and 810 ms, respectively, for all exact tasks combined).
12. Participants were right-handed French students aged between 22 and 28 years (three men and four women in the fMRI version, five men and seven women in the ERP version). The project was approved by the regional ethical committee, and all subjects gave written informed consent. Stimuli were addition problems with addends ranging from 1 to 9 and sums ranging from 3 to 17. Ties such as 2 + 2 were excluded. For the exact task, the two candidate answers were the correct result and a result that was off by, at most, two units. In 90% of exact problems, the wrong result was of the same parity as the correct result, thus preventing the use of a short-cut based on parity checking [L. E. Krueger and E. W. Hallford, *Mem. Cogn.* **12**, 17 (1984)]. For the approximation task, the two alternatives were a number off by one unit, and a number off by at least four units. A third control task of letter matching was also introduced, in which digits were replaced by the corresponding uppercase letter in the alphabet, and subjects depressed the button on the side on which a letter was repeated from the initial pair. Tasks were presented in runs of alternating blocks of trials with a 4-s intertrial interval, separated by resting periods of 24 s. Four such runs were presented in semi-random order, two runs alternating exact calculation (three blocks of 18 trials each) with letter matching (three blocks of nine trials each), and two similar runs alternating approximation with letter matching. For fMRI, we used a gradient-echo echo-planar imaging sequence sensitive to brain oxygen-level dependent contrast (30 contiguous axial slices, 5 mm thickness, TR = 4 s, TE = 40 ms, angle = 90°, field of view 192 mm by 256 mm, matrix = 64 by 64) on a 3-T whole-body system (Bruker, Germany). High-resolution anatomical images (three-dimensional gradient-echo inversion-recovery sequence, TI = 700 ms, TR = 1600 ms, FOV = 192 mm by 256 mm, matrix = 256 × 128 × 256, slice thickness = 1 mm) were also acquired. Analysis was performed with SPM96 software (www.fil.ion.ucl.ac.uk/spm). Images were corrected for subject motion, normalized to Talairach coordinates using a linear transform calculated on the anatomical images, smoothed (full width at half maximum = 15 mm), and averaged across subjects to yield an "average run" in each condition (the results were replicated when the same analysis was applied to individual data with 5-mm smoothing). The generalized linear model was used to fit each voxel with a linear combination of two functions modeling early and late hemodynamic responses within each type of experimental block. Additional variables of noninterest modeled long-term signal variations with a high-pass filter set at 320 s. Because approximate and exact calculation blocks were acquired in different runs, the statistics we report used the interaction term (exact calculation – its letter control) – (approximate calculation – its letter control), with a voxelwise significance level of 0.001 corrected to

$P < 0.05$ for multiple comparisons. In a separate session, ERPs were sampled at 125 Hz with a 128-electrode geodesic sensor net reference to the vertex [D. Tucker, *Electroencephalogr. Clin. Neurophysiol.* **87**, 154 (1993)]. We rejected trials with incorrect responses, voltages exceeding $\pm$100 μV, transients exceeding $\pm$50 μV, electro-oculogram activity exceeding $\pm$70 μV, or response times outside a 200- to 2500-ms interval. The remaining trials were averaged in synchrony with stimulus onset, digitally transformed to an average reference, band-pass filtered (0.5 to 20 Hz), and corrected for baseline over a 200-ms window before stimulus onset. Experimental conditions were compared within the first 400 ms by sample-by-sample t tests, with a criterion of $P < 0.05$ for five consecutive samples on at least eight electrodes simultaneously. Two-dimensional maps of scalp voltage were constructed by spherical spline interpolation [F. Perrin, J. Pernier, D. Bertrand, J. F. Echallier, *Electroencephalogr. Clin. Neurophysiol.* **72**, 184 (1989)]. Dipole models were generated with BESA [M. Scherg and P. Berg, *BESA—Brain Electric Source Analysis Handbook* (Max-Planck Institute for Psychiatry, Munich, 1990)]. Three fixed dipoles were placed at locations suggested by fMRI (left inferior frontal and bilateral parietal), and the program selected the dipole orientation and strength to match the exact-approximate ERP difference on a 216- to 280-ms time window, during which significant differences were found.

13. A. R. Damasio and N. Geschwind, *Annu. Rev. Neurosci.* **7**, 127 (1984); C. Price, *Trends Cogn. Sci.* **2**, 281 (1998).
14. R. A. Andersen, *Philos. Trans. R. Soc. London Ser. B* **352**, 1421 (1997); A. Berthoz, *ibid.*, p. 1437; M. Jeannerod, *The Cognitive Neuroscience of Action* (Blackwell, New York, 1997); L. H. Snyder, K. L. Grieve, P. Brotchie, R. A. Andersen, *Nature* **394**, 887 (1998).
15. R. Kawashima *et al.*, *Neuroreport* **7**, 1253 (1996); H. Sakata and M. Taira, *Curr. Biol.* **4**, 847 (1994).
16. H. Kawamichi, Y. Kikuchi, H. Endo, T. Takeda, S. Yoshizawa, *Neuroreport* **9**, 1127 (1998); S. M. Kosslyn, G. J. DiGirolamo, W. L. Thompson, N. M. Alpert, *Psychophysiology* **35**, 151 (1998); W. Richter, K. Ugurbil, A. Georgopoulos, S. G. Kim, *Neuroreport* **8**, 3697 (1997).
17. M. Corbetta, F. M. Miezin, G. L. Schulman, S. E. Petersen, *J. Neurosci.* **13**, 1202 (1993); M. Corbetta, G. L. Shulman, F. M. Miezin, S. E. Petersen, *Science* **270**, 802 (1995); A. C. Nobre *et al.*, *Brain* **120**, 515 (1997); M. I. Posner and S. Dehaene, *Trends Neurosci.* **17**, 75 (1994).
18. S. Dehaene, *J. Cogn. Neurosci.* **8**, 47 (1996); S. Dehaene *et al.*, *Neuropsychologia* **34**, 1097 (1996); P. E. Roland and L. Friberg, *J. Neurophysiol.* **53**, 1219 (1985); L. Rueckert *et al.*, *Neuroimage* **3**, 97 (1996).
19. The exact and approximate tasks did not differ in mean response time (fMRI: 772 and 783 ms; ERPs: 913 and 946 ms, respectively) nor in error rate (fMRI: 4.4 and 5.3%; ERPs: 2.7 and 2.3%, respectively) (all Fs < 1).
20. In individual analyses ($P < 10^{-3}$, corrected), the finding of greater intraparietal activation during approximation than during exact addition was replicated in five of seven subjects, whereas significantly greater left inferior frontal activation during exact addition was observed in four of seven subjects.
21. S. E. Petersen, P. T. Fox, M. I. Posner, M. Mintun, M. E. Raichle, *Nature* **331**, 585 (1988); M. E. Raichle *et al.*, *Cereb. Cortex* **4**, 8 (1994); R. Vandenberghe, C. Price, R. Wise, O. Josephs, R. S. J. Frackowiak, *Nature* **383**, 254 (1996); A. D. Wagner *et al.*, *Science* **281**, 1188 (1998).
22. Y. G. Abdullaev and N. P. Bechtereva, *Int. J. Psychophysiol.* **14**, 167 (1993); Y. G. Abdullaev and M. I. Posner, *Neuroimage* **7**, 1 (1998); A. Z. Snyder, Y. G. Abdullaev, M. I. Posner, M. E. Raichle, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1689 (1995).
23. H. Hécaen, R. Angelergues, S. Houillier, *Rev. Neurol.* **105**, 85 (1961); M. Rosselli and A. Ardila, *Neuropsychologia* **27**, 607 (1989).
24. S. Dehaene and L. Cohen, *Cortex* **33**, 219 (1997).
25. A. L. Benton, *Arch. Neurol.* **49**, 445 (1992); Y. Takayama, M. Sugishita, I. Akiguchi, J. Kimura, *ibid.* **51**, 286 (1994); M. Delazer and T. Benke, *Cortex* **33**, 697 (1997). Note that in such left parietal cases, the preservation of exact arithmetic facts is clearest for very small multiplication and addition problems (*23*). As the numbers involved in an exact arithmetic problem get larger, subjects are more and more likely to rely on quantity-based strategies to supplement rote verbal retrieval [(*7*); see also J. A. LeFevre *et al.*, *J. Exp. Psychol. Gen.* **125**, 284 (1996); J. LeFevre, G. S. Sadesky, J. Bisanz, *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 216 (1996)]. Indeed, supplementary fMRI analyses (available from S. Dehaene) showed that, within the present exact addition task, increasing problem sizes caused greater activation of the bilateral intraparietal circuit in regions identical to those active during approximation. This finding suggests that verbal and quantity representations of numbers are functionally integrated in the adult brain. Although only the verbal circuit is used for well-rehearsed exact arithmetic facts, both circuits are used when attempting to retrieve lesser-known facts.
26. S. Dehaene and L. Cohen, *Neuropsychologia* **29**, 1045 (1991).
27. T. Dantzig, *Number: The Language of Science* (Free Press, New York, 1967); G. Ifrah, *Histoire Universelle des Chiffres* (Robert Laffont, Paris, 1994).
28. M. Kline, *Mathematical Thought from Ancient to Modern Times* (Oxford Univ. Press, New York, 1972).
29. S. T. Boysen and E. J. Capaldi, Eds., *The Development of Numerical Competence: Animal and Human Models* (Erlbaum, Hillsdale, NJ, 1993); E. M. Brannon and H. S. Terrace, *Science* **282**, 746 (1998); S. Dehaene, G. Dehaene-Lambertz, L. Cohen, *Trends Neurosci.* **21**, 355 (1998); C. R. Gallistel, *Annu. Rev. Psychol.* **40**, 155 (1989).
30. K. Wynn, *Trends Cogn. Sci.* **2**, 296 (1998).
31. S. Dehaene, *The Number Sense* (Oxford Univ. Press, New York, 1997).
32. Supported by NIH grant HD23103 (E.S.) and by the Fondation pour la Recherche Médicale (S.D.). We gratefully acknowledge discussions with L. Cohen, D. Le Bihan, J.-B. Poline, and N. Kanwisher.

6 January 1999; accepted 16 March 1999

Discovery of a Small Molecule Insulin Mimetic with Antidiabetic Activity in Mice

Bei Zhang,[1*] Gino Salituro,[2] Deborah Szalkowski,[1] Zhihua Li,[1] Yan Zhang,[2] Inmaculada Royo,[4] Dolores Vilella,[4] Maria Teresa Díez,[4] Fernando Pelaez,[4] Caroline Ruby,[2] Richard L. Kendall,[5] Xianzhi Mao,[5] Patrick Griffin,[3] Jimmy Calaycay,[3] Juleen R. Zierath,[6] James V. Heck,[2] Roy G. Smith,[1†] David E. Moller[1]

Insulin elicits a spectrum of biological responses by binding to its cell surface receptor. In a screen for small molecules that activate the human insulin receptor tyrosine kinase, a nonpeptidyl fungal metabolite (L-783,281) was identified that acted as an insulin mimetic in several biochemical and cellular assays. The compound was selective for insulin receptor versus insulin-like growth factor I (IGFI) receptor and other receptor tyrosine kinases. Oral administration of L-783,281 to two mouse models of diabetes resulted in significant lowering in blood glucose levels. These results demonstrate the feasibility of discovering novel insulin receptor activators that may lead to new therapies for diabetes.

The actions of insulin are initiated by its binding to the insulin receptor (IR), a disulfide-bonded heterotetrameric membrane protein (*1–3*). Insulin binds to two asymmetric sites on the extracellular α subunits and causes conformational changes that lead to autophosphorylation of the membrane-spanning β subunits and activation of the receptor's intrinsic tyrosine kinase (*4*, *5*). Insulin receptors transphosphorylate several immediate substrates (on Tyr residues) including insulin receptor substrate (IRS) proteins (*6*). These events lead to the activation of downstream signaling molecules. The function of the receptor tyrosine kinase is essential for the biological effects of insulin (*1–6*).

The pathogenesis of type 2, non–insulin-dependent diabetes mellitus (NIDDM) is complex, involving progressive development of insulin resistance and a defect in insulin secretion, which leads to overt hyperglycemia. The molecular basis for insulin resistance in NIDDM remains poorly understood. However, several studies have shown modest ($\approx$30 to 40%) decreases in IR number with tissues or cells from NIDDM patients (*7*).

[1]Department of Molecular Endocrinology, [2]Department of Natural Product Drug Discovery, [3]Department of Molecular Diversity and Design, Merck Research Laboratories, R80W250, Post Office Box 2000, Rahway, NJ 07065, USA. [4]Centro de Investigación Básica, Merck, Sharp & Dohme de España, S. A. Josefa Valcárcel 38, Madrid 28027, Spain. [5]Department of Cancer Research, Merck Research Laboratories, Post Office Box 4, West Point, PA 19486, USA. [6]Department of Clinical Physiology, Karolinska Hospital, Karolinska Institute, S-171 76 Stockholm, Sweden.

*To whom correspondence should be addressed. E-mail: bei_zhang@merck.com

†Present address: Huffington Center on Aging and Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, M-320, Houston, TX 77030, USA.