Table of Contents for Neural Networks

Volume 10, Issue 7, 1 October 1997

Consciousness for Neural Networks

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1173, 1997

Printed in Great Britain
0893–6080/97 $17.00 + .00

PII: S0893-6080(97)00095-6

Introduction: 1997 Special Issue Neural Networks for Consciousness

Consciousness research is now coming to the forefront in cognitive neuroscience. The instruments of non-invasive imaging—PET, fMRI, EG and MEG are now being deployed to attempt to tease out how, when and where consciousness emerges in the brain and how it is then used to provide the efficiency of processing able to achieve acts of great creativity and reasoning. Yet, the wealth of data pouring in will not be able to be correlated together and used to build a clear picture of how the brain works, and in particular how consciousness itself is supported, without there being a corresponding underpinning at a theoretical level. This must be ultimately at the level of neural networks. That is why it is our pleasure to welcome you to this special issue on "Neural Networks for Consciousness". The issue contains contributions from across the range of workers in the field—from experimental neuroscientists to cognitive scientists to engineers to mathematicians. This breadth alone shows the vitality of the field and the great interest in it from across these disciplines. It also shows the strength future research can draw on since all of these disciplines can, in principle, contribute to solving the problems that consciousness raises. The problems themselves are very deep indeed, and we do not claim that we have assembled the last word on the subject. However, the acumen and experience in thinking about consciousness that the contributors undoubtedly possess will make this issue of both general interest and also of relevance to those concerned either in their own thinking processes or creating machines which can emulate them. We wish you enjoyment in reading the contributions.

John G. Taylor
London

Walter Freeman
Berkeley, CA

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1185–1194, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00055-5

1997 SPECIAL ISSUE

Non-Invasive Analysis of Awareness

J. G. TAYLOR AND H.-W. MUELLER-GAERTNER

Institute for Medicine, Research Centre-Juelich, D-52425

(*Received* 30 *September* 1996; *accepted* 2 *May* 1997)

Abstract—*A program is outlined which indicates in what manner awareness may be probed by the recently developed non-invasive instruments (PET, fMRI, MEG and EEG). The nature of the neural systems needed for the creation of awareness is reviewed, leading to a set of characteristics required for the crucial experiments able to uncover the processes involved. The assessment of experiments possible to probe the emergence of perception is then surveyed in terms of these criteria, and those which appear to be most effective are delineated. Brief conclusions are then drawn as to the viability of the program. © 1997 Elsevier Science Ltd.*

1. INTRODUCTION

There is considerable interest in the use of non-invasive instruments to detect the emergence of phenomenal awareness, say of a percept. The new machines give hope of being able to localise, both spatially and temporally, the neural processes which underpin such an event. However, there are numerous questions which must be answered before the search can be made well-enough defined to design paradigm experiments which can be expected to be of high import in this task. It is the purpose of this paper to explore these questions and come to conclusions as to the viability of this program.

Amongst these questions are:

(a) what is the nature of phenomenal awareness itself?
(b) how can the ''explanatory gap'' (which separates the internal experience from the external neural patterns of the behavioural response) be bridged?
(c) what is the expected spatial localisation of the process?
(d) what are the expected dynamical neural processes?

Some of these questions are very difficult, especially (a) and (b), but without at least a preliminary attempt to answer them it may be difficult to answer questions (c) and (d), and so give guidance to the proposed experimental program.

There is a considerable lack of agreement as to the answers to (a) and (b). However, it is becoming recognised that awareness involves use of previously encoded material to give content and meaning to new inputs (Baars, 1988; Baddeley, 1993; Edelman, 1989; Taylor, 1973, 1991). Initial processing before the awareness of percepts occurs is at the lowest level of coding, that of feature encoding and filtering in the early and associative cortices. At the next stage, higher level object representations are expected to be activated which begin to ''fill out'' and bind together the lower level activations into meaningful percepts. At various stages there would also be feedback helping to guide activation of the best representation able to fit the total input (Bullier, 1995). Awareness of a percept, say of a geometrical shape, of a colour or a moving region, will be expected to occur after the lowest level of coding has been achieved, so at the object (or equivalent) level, although exactly where that occurs is difficult to pin down, and is the object of the experimental program being discussed here. However, persistence of suitable encoded activity is expected to be essential. The continued activity arises in modules with the special property of being able to hold activity for a suitably long temporal duration. Such modules are different from those used in the first (non-conscious) stage of the processing.

This leads to a working definition of awareness at the lowest level as some form of encoded input which is available, over a suitably extended period of time, for further processing, and in particular for ''filling out'', by using closely related stored representations of experiences from the past, so as to be useable at higher levels (for action responses, reasoning, planning, for example; Taylor, 1998). This definition needs to be developed more fully in terms of detailed results from

Acknowledgements: J. G. Taylor would like to thank his co-author (H.-W. Mueller-Gaertner) for the excellent support given him whilst he was a guest scientist at the Institute of Medicine, Research Centre-Juelich where this work was performed.

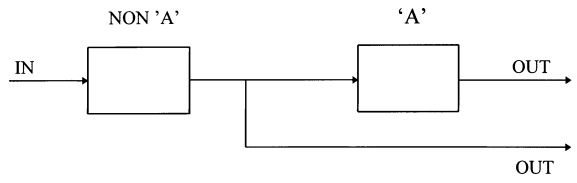


FIGURE 1. The hierarchical two-stage model of consciousness in which the early processing by the ''Non A'' modules is only at preconscious level whilst that by the ''A'' modules supports the emergence of consciousness.

the experimental program, but should help in initially clarifying what is to be looked for.

What results from the above brief sketch of the nature of awareness is a two-stage model for consciousness with a simplified flow chart as in Figure 1. In the figure, the ''Non A'' module performs the low-level coding of which a person is unaware. The output from this module is then passed on to the awareness module ''A'' or can bypass it so as to go directly to the motor response system. This two-stage model is to be distinguished from the continuous or one-stage model, in which consciousness arises in the same modules as were used in the non-conscious processing. The reason for such extra awareness of the activity would, for some reason or other in this one-stage model, be due to increased activity in the relevant modules.

There is much support for the two-stage model. Thus William James (1950) wrote ''the distribution of consciousness shows it to be exactly such as we might expect in an organ *added* for the sake of steering a nervous system grown too complex to regulate itself''. Also, from evidence associated with certain forms of prosopagnosia, Young (1994) wrote ''activation must cross some form of threshold before it can result in awareness''. Further evidence from neglect led Bisiach and Berti (1995) to note ''patients may be apparently unaware of what lies on the side opposite to the brain lesion despite its having been processed up to a high level, perhaps, at which meaning is captured''. Finally, there is evidence from study of the blindsighted subject GY that ''the visual input can reach V5 without passing first through V1 and that such input is sufficient for both discrimination and conscious awareness of the visual input'' (Weiskrantz et al., 1995). These quotations, especially the last two, indicate that there are cortical modules needed for the production of awareness beyond those involved in low-level processing.

The answer to (b) above, as to the bridging of the explanatory gap, is already indicated by the above answer to (a). It will be in terms of the manner in which the preprocessed neural activity representing an input is able to be made available for further transformations which lead to a higher cognitive level. Such availability will require suitable temporal duration of the representations so that they can be involved in activities in regions crucially concerned with such higher level processes, such as the frontal lobes.

It is natural to consider possible ''buffer'' posterior memory sites, holding activity encoded at object level, as the first candidates for sites of this availability. These sites sustain temporally extended activity, as is known from psychological tests of short-term memory functioning in the presence of distractors to prevent long-term memories from forming (Baddeley, 1986). Such short-term memory buffers have been observed directly by non-invasive techniques by a number of groups (Paulesu et al., 1993; Salmon et al., 1996; Smith and Jonides, 1995) after the initial proposal for their existence based on psychological data; the present evidence is summarised in Table 1. These sites give a tentative first answer to question (c). They also allow the model of Figure 1 to be extended to that of Figure 2.

In Figure 2, the highest level of coding at a non-aware level is denoted SM (for semantic memory in word processing) which feeds automatically to the working memory WM. It is in the latter that awareness arises, through lateral inhibitory and excitatory connections on the WM module. In Figure 3 are shown two of the coupled SM/WM pairs for different codes (denoted A and B). The manner in which awareness arises from joint activity of these two sets of modules has been proposed to involve a more global control system involving the thalamus and the nucleus reticularis thalami (a sheet of inhibitory neurons with good lateral connections; Taylor, 1992). It is through this sheet that a competition can be run between the two (or more) coupled SM/WM systems shown in Figure 2.

TABLE 1
PET results on posterior sites of working memory

Task	Brain area	Task/Code
Spatial discrimination	Left 40 (posterior parietal)[a]	Object shape
	Right 40/19 (occipito/parietal)[a]	
	Right 19 (occipital)[b]	Orientation discrimination
Object discrimination	Left 40 (posterior parietal)[a]	Object
	Left 37 (inferotemporal)[a]	
Phonemes/Words	Left 40 (posterior parietal)[c]	Word disambiguation

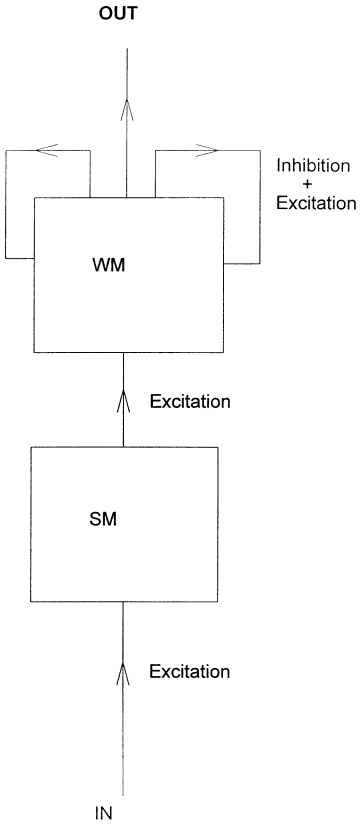


FIGURE 2. Competition for emergence into consciousness in the two-stage model where there is lateral inhibition on the second ''A'' or working memory (WM) stage modules producing consciousness from the earlier preprocessing on the ''Non A'' or semantic (SM) modules.

In a similar manner, the coupled SM/WM module pairs of Figure 2 provide a first guess at the temporal dynamics of the initial stages of awareness, question (d), as has been recently discussed by Taylor (1996b). This gave an analysis, for a lexical decision task (Marcel, 1980), of the manner in which subliminally processed data can retard or speed up processing of later data. It posited that the initial emergence of awareness of words arises from their persistence on the phonological store in Brodmann's area 40 (Paulesu et al., 1993), where they decay away over a fixed time of several seconds.

In this model, the initially unperceived neural activity at the highest level of preprocessing (in the lexical case up to the level of the semantic coding in Wernicke's area) has numerous representations active, corresponding to possibly contradictory interpretations. It is only at the next level up, that of the phonological store, that a competition is run between these various possibilities, and one is singled out by the use of previous activities there which had earlier reached awareness or had buffered subliminal activity. Thus, perception is regarded as the filter to achieve disambiguation between alternative interpretations based on earlier decision made there.

It is unclear at this stage exactly how to extend this model to other modalities, especially to vision, where numerous separate codes exist and the problem of the combination of the resulting activity may be most extreme. It is the purpose of the following discussion to consider several possibilities of doing that which will delineate features needing to be probed by suggested experiments, and hence give valuable criteria to apply to their choice.

2. NEURAL SUPPORTS AND CRITERIA

There appear to be two extremes of neural support which may be suggested as being at the basis of the emergence of perception:

1. in a single module (as suggested by the model of word perception in Taylor, 1996b)
2. as a distributed activity over a set of coupled modules.The second model is more likely to be valid in vision, given the number of codes (colour, motion, shape, texture, stereoscopy) and the level of interconnection between the various visual areas (Felleman and Van Essen, 1991). However, it is still to be expected that there will be certain of these visual areas which are playing a more decisive part in the emergence of a visual percept in some particular code, with characteristics as indicated at the end of the previous section.

Whilst the questions (c) and (d) above about the spatial and temporal characteristics of neural activity crucially involved in perception were posed separately, it is clear that the answers to them must be combined in some manner. This is achieved on positing (Taylor, 1996b) that the ''buffering'' memory sites be regarded as candidates for emergence of percepts in that particular code. Thus we are lead to the first criterion:

1. To search for those posterior cortical sites which are most effective in temporally sustaining activity as a percept becomes experienced.

There may well arise a set of sites satisfying the above criterion, in which there may be a spread of temporal duration over the various modules, but they all must have several seconds of duration for neural activity. How do we regard the emergence of the percept: from all of the modules conjointly or only, say, from that module which holds activity for the longest time? This is partly an experimental question, but there is great relevance to perform the related psychophysical experiments to ascertain the time of first experience by the subject of the percept.

An example of such a combination of psychophysical and non-invasive instrument experiments is that of the motion after-effect (Tootall et al., 1995), in which the experiential decay of adaptation was found to

match that of blood flow activation as observed by fMRI in suitable visual cortical regions. However, it is necessary to note if there are any regions in which activity persists even longer than that of the subjective experience. The model of lexical decision of (Taylor, 1996b) would indicate that activity on the phonological store persists longer than the subjective experience itself. This arose, according to the model, when the neural activity in the buffer site became stronger than some criterial value. The activity then decays away over a few seconds, and so should leave a signal of its activation trace lasting longer than the subjective experience. Thus we are led to the second criterion:

2. Attempt to order the modules observed to have extended temporal activity according to the duration of the observed activity (beyond that experienced by the subject); the best candidate for the emergence of perception will be that module with the longest temporal duration.

It is still not clear that the module holding neural activity for the longest time during the development of a percept will be the only one involved in the emergence of that percept in the subject's experience. There may indeed be others which hold activity for a very similar, but slightly shorter, duration. In that case it would be necessary to measure the manner in which there is a causal flow of activity between the modules to determine if one is dominating the others. Thus we must impose the third criterion:

3. The most suitable candidate(s) will be those modules which are causally the dominant one(s) in the temporal dynamics causing the emergence of the percept.

To apply this criterion, it will be necessary to measure accurately both the temporal dynamics of the neural activity as well as that of the subjective experience. This criterion puts a requirement on the appropriate machines to use as those with fast temporal sensitivity, i.e. EEG and MEG. It may also be possible to use suitable sequencing of the slicing in echo-planar imaging in fMRI to follow the dynamics of activity through the brain (Shah, 1997).

3. SURVEY OF RELEVANT EXPERIMENTS

The purpose of this section is to consider how effective various experiments will be in achieving the aim to capture the cortical sites of the emergence of perception under various visual (and other modality) stimulation conditions. The basic question that has to be addressed in this assessment is:

Does the experiment under consideration reveal areas which give ''added value'' to cortical processing beyond the level of non-conscious processing already known to occur, say at the retina or at the level of the LGN? More specifically do the areas satisfy the above three criteria?

The non-conscious processing areas have therefore to be delineated and separated from those involved in the emergence of a percept. That an experiment cannot do this will be taken as indication it will not be a crucial one for the program. However, the results could still be of great importance in determining the details of preprocessing necessary to build a more complete picture, both experimental and model-based, which can be used to probe the emergence of a percept in more detail.

3.1. Visual Modality Experiments

3.1.1. *Motion After-Effect (MAE).* This paradigm uses the relation between the perceptual and blood flow time courses of the decay of the adaptation to continuous motion, as in the case of looking at a waterfall for 30 seconds and then viewing the adjacent stationary rock face; the latter appears to move upwards. We have conjectured in the previous section that the MAE percept arises in that particular region which has the slowest such decay. It may be that further regions, with blood flow activity decaying as slow as or slower than the perceptual one, are crucially involved in percept formation; it is the observation of these further regions that must be done in more detail. That may also be observed by MEG analysis of the faster processing between the regions and the more detailed nature of that interaction.

In any case the experimental paradigm is satisfactory, as far as criteria 1 to 3 above are concerned (and has already been used to good effect in Tootall et al., 1995). It involves direct comparative analysis of timing effects across different areas.

3.1.2. *Orientation-Specific Adaptation.* Continued viewing of oriented stripes (for over 5 or so seconds) brings about adaptation to them, as observed by an increase of the contrast threshold for detection of stripes oriented in that direction as compared with those in a perpendicular direction (He et al., 1996). This effect has been shown, by means of ''crowding''—the input of further sets of oriented stripes in adjacent regions of visual space so as to cause a loss of awareness of the direction of orientation but not loss of non-conscious knowledge of that orientation—to be present in spite of the loss of the conscious experience by human observers of the particular orientation to which adaptation had occurred. The preservation of orientation adaptation knowledge, in spite of the loss of awareness due to crowding, indicates that ''visual awareness, and the crowding that blocks it, occur after orientation analysis in the visual information processing stream'' (He et al., 1996). Moreover, the asymmetry of crowding (being stronger in the upper half visual field) can only arise from areas after V1

(and even V2) since these areas are both symmetrical between the upper and lower half fields.

These effects—the differential placement of orientation adaptation (in V1 and V2) and crowding in the visual processing stream—lead to search for the placement of the computation of crowding itself. Let us suppose that it can be determined to be in a particular visual area, V*n* say. Does that imply that consciousness of crowding must also occur in V*n*? It is possible that the crowding effect is itself computed in an earlier module than would correspond to the working memory site of consciousness of the two-stage model of Figure 1. In other words, consciousness could arise at a later stage than the initial computation of the crowding effect, so that the first module performing the computation of the effect is not a ''tag'' for the emergence of awareness. It would be necessary to determine the time course of activity in order to pin down which module (or modules) were more precisely involved in the emergence of awareness of AM, following the criteria set out in Section 2.

In conclusion, the adaptation paradigm itself is not satisfactory to obtain the source of awareness. The crowding effect, probed as to its temporal course across various modules involved in its computation, does seem to be so. In that form it comes under subliminal processing, and as such will be considered in more detail in Section 3.3.

3.1.3. *Apparent Motion (AM).* This is a well-known paradigm in which a spot of light, say being shone at a particular position on a screen, is deleted and almost immediately (within about 300 msec) shone at a nearby point. This produces in the subject the experience of motion of the spot from the first to the second position.

The difficulty with the experiment as one for detecting the emergence of the awareness of motion is that it uses the motion pathway all the way up from V1 (and below, in LGN) due to the rapid onset and offset of the two stimuli, so that even when the experience of motion arises in the subject (when the spots have been moved close enough together for the motion percept to occur) there may be use of areas already activated by the sudden onset and offset which are involved in both non-conscious and conscious processing. The times of onset and offset might be used to separate out the motion effects they cause from the activation due to the further motion experience. However, this may need high spatial resolution.

There is also the difficulty that the initial lowest level computation of the track of motion may be achieved at a lower level than the computation of awareness; this is the same problem as that met above in orientation adaptation. It would be necessary to use the further criterion of temporal dynamics to ascertain the modules turned on by the AM signal which also held it for a suitably long temporal duration.

In conclusion, the AM paradigm is only satisfactory if its temporal dynamics is also probed with suitable precision.

3.1.4. *Blindsight.* The case of GY (or similar subjects), for whom there is a separation of ''contentless awareness'' as compared with knowledge without awareness, as stimulus speeds are altered (Weiskrantz et al., 1995), is an excellent one to observe possible regions which are activated by one paradigm compared with the other. This corresponds precisely to the case of subliminal versus conscious perception considered below. There is, however, the concern that there may be a considerable difference in the modules involved in such subjects as compared with those in the emergence of perception in normal humans. A blindsighted person may well have developed new connections and usage of modules that are different than those used in normally sighted people in the creation of awareness.

In conclusion, the paradigm is satisfactory (modulo the problem raised above).

3.1.5. *Normal Blindsight.* The phenomenon of blindsight, in which there is knowledge without awareness, is one which has been activated in normal subjects by means of the experimental paradigm of Kolb and Braun (1995). In this paradigm, a field of short oriented bars (or oppositely moving dots), all with the same orientation (or direction of motion), is embedded in a set of similar but oppositely oriented (or moving) objects. There is knowledge of the position of the small embedded subset of stimulus patterns, but no awareness occurs. It would be possible to modify the stimulus parameters, by increasing the size of the region of the embedded stimuli, so as to cause the emergence of perceptual experience in a controlled manner.

This experimental paradigm is seen to suffer from the difficulty that there is already considerable activation along the whole of the dorsal pathway caused by the surrounding pattern, of which the subject is clearly aware. The modification of this activity when there is ''normal blindsight'', as compared with when there is not, will therefore be submerged in a large level of activity involved with normal perception.

In conclusion, the paradigm is not satisfactory (due to the low signal for the signature of the emergence of awareness). It is also relevant to note that the paradigm is part of that of subliminal processing, so will be considered in more detail in Section 3.3.

3.1.6. *Rivalry/Binocular Rivalry.* Rivalry of concepts, such as arises in the Necker cube, has a long history of psychological investigation. The case of binocular rivalry has also been used in humans (Kolb and Braun, 1995) and in monkeys (Logothetis and Schall, 1989; see also Sheinberg and Logothetis, 1997) to attempt to ascertain where the computation occurs in cortex for the

production of the rivalrous percepts as they alternate. This has not yet proved possible. Yet there is a similar problem to that raised above for orientation adaptation and AM: it is necessary to probe the temporal dynamics on the modules involved very carefully.

In the case of switching between the two images in the Necker cube, the computation could occur at a level below that of awareness. Thus if the mechanism were of simple adaptation of one percept to give way to the other, for example through some form of competitive process, such adaptation may take place on a module at the NonAware level of Figure 1, with output being used to feed onto the further Awareness module A in that figure. This would not correspond precisely to the elaborated two-stage model of Figure 2, where disambiguation takes place only at the same level as awareness on the working memory site WM in that figure. However, it may be that there is some decoupling between disambiguation and temporal duration in the visual modality, especially due to the large number of codes involved as compared with the auditory one. The crucial tag of consciousness was claimed in Section 2 to be that of the extended temporality of neural activity; the disambiguation process is not necessarily associated with the onset of awareness (nor will it be so easy to identify).

In conclusion, it would be necessary to determine detailed temporal characteristics of neural activity during rivalrous phenomena; those modules in which there is temporal duration highly correlated with the oscillation between the concepts (but having the highest such duration, even extending beyond the switch-over times) would be the best candidates for the site of awareness. The observation of such sites may be difficult, however, due to highly similar coding which may occur for the rivalrous concepts. Thus, the experimental probing of such phenomena is of great interest but not at the top of the list of crucial experiments to probe awareness.

3.1.7. *Stabilised Images.* When images are stabilised on the retina they are know to fade (Pritchard, 1963). The manner in which they do so has very interesting characteristics, such as would seem to imply that both Gestalt and Hebbian ensemble actions are at work. This has been explained by Taylor (1996a) in terms of separate loss of activation of the ''what'' and ''where'' pathways. It would be possible to attempt to determine which regions lose activation as the percept fades from awareness.

It is unclear, when a stabilised image drops out of perception, that it does so only to the next lower level of non-conscious activation. It may be that there is no activity even from LGN, or from any other low-level module, where the most crucial levels of adaptation may occur. The discovery of the set of modifications (which area is on, which off, when the stabilised image blinks off) is of great interest, but not able to provide the desired answer.

In conclusion, it can only be said, for the above reason, that the paradigm is not satisfactory.

3.2. Auditory Modality

3.2.1. *The Double Click Experiment.* The double click experiment involves a subject attempting to detect those double clicks (with interstimulus delivery times of 0, 6, 12, 18 and 24 msec) presented to the ears which are detectable as double clicks and not single ones. At an interstimulus interval larger than about 12 msec there appears to be both double click detection and a phase shift in the observed 40 Hz signal observed by MEG (Joliot et al., 1994). Let us suppose it to be possible to detect the region involved in the computation of the double click interstimulus interval to produce such a phase shift. Is that necessarily the region where awareness of the double click arises?

The answer is no, since it is possible to consider that the computation of the interstimulus intervals was performed at a lower non-aware level and sent on to the Awareness module of Figure 2. Such a possibility could be guarded against if there was a control experiment performed which required the subject to be distracted by a different stimulus load (say by counting down by threes from one hundred). The resulting activity would show whether or not the activity was still the same (the phase shift still having occurred at about 12 msec interstimulus interval). If the activity were still in the same cortical area then the computation of the ISI would clearly have been at a non-aware level. But if there were a difference in the sites of activity between the attended and distracted conditions it would still not be the case that the residual activity, after subtracting the aware from the distracted condition, would leave modules activated solely at an aware level. The distraction could have prevented processing in some parts of the non-aware circuitry but not throughout it all; the rest could still be producing the stimulus timing signal to send up to the awareness level.

In conclusion, again temporal duration must be used as a critical tag to separate non-aware from aware processing.

3.2.2. *Binaural Fusion.* This phenomenon involves the observation and comparison by means of MEG of those regions activated either by a tone or by a set of harmonics of the tone which lead to the same percept of the sound in a subject (Pantev et al., 1989). This experimental paradigm has only been performed under attentive processing conditions, so it is unclear that the crucial regions were at an aware level. If the experiment were also to be performed, as in the double click experiment, under a distracting load then one would still have the difficulty raised in the previous paradigm, that the computation, after subtracting the activity that was occurring under the distracting load, would not necessarily be at the

level of the awareness circuitry of Figure 1, but could still be down in the non-aware modules.

In both of these paradigms it would be necessary to determine timing effects with care. Thus in conclusion, the paradigm is interesting but does not single them out as leading automatically to the non-aware/aware boundary.

3.3. Subliminal Perception

There are different regions devoted to supporting subliminal as compared with aware activity, as Figure 1 denotes. Thus confounding the two forms of activity will make the flow of information in the brain very difficult to understand. This is particularly important as more data is gathered from the non-invasive instruments. It will become more and more difficult to comprehend without a proper distinction between low and top-level processing being introduced. Otherwise the data will just pile up to increase confusion over its interpretation.

Such seems to be the case in lexical processing (though not involving non-invasive studies yet), where recent results (Levelt et al., 1991) seem to contradict the simple connectionist approaches to lexical priming (in particular, the expected broad spread of activation at phonological level to semantic associates of an input word or target to be recognised). Yet these analyses and theoretical frameworks are not based on standard understanding of the phonological loop and its relevance to conscious experience. That needs to be developed on sure experimental grounds. The same situation will occur in vision and other modalities, although the theoretical structures are not as well developed to be able to appreciate the problem as effectively.

3.3.1. *Fully Subliminal Versus Partly Subliminal Paradigms.* One experimental paradigm to determine the extra machinery which divides perception without awareness from that with it would thus be to subtract results obtained from experiments on stimulus detection in which there is only a subliminally perceived stimulus from those obtained when awareness has occurred. The neuronal modules activated in giving the extra feature of aware beyond that of non-aware processing would then be detectable as having higher activation levels after subtraction, or more general comparison through a suitable correlational technique.

An alternative or additional paradigm is to consider an experiment in which there are different effects arising from stimuli perceived at the two levels (as in the Stroop effect). The results of such an experiment will have to be interpreted with care so as to be able to distinguish between the aware level of processing and that at a non-conscious level; certain paradigms should allow the separation of modules involved at the two different levels by means of use of different paths in the brain (as in foveal versus parafoveal experiments) or different parts of a module (as in the ''filling in'' phenomenon in audition).

The first of these paradigms will be termed ''fully subliminal'' experience, the second ''partly subliminal'' experience. These lead clearly to different experimental paradigms and to different criteria on the stimulus delivery systems. That will be explored below. The partly-subliminal is a paradigm which will be discussed further in Section 3.3.3 after the consideration of the full subliminal stimulus paradigm in the next section.

3.3.2. *Fully Subliminal Stimuli.* The basic idea behind such a visual stimulus delivery paradigm is that they be presented to subjects on a screen in front of them so that some of the stimuli are experienced at a subliminal level, whilst, in comparison and at different times, other stimuli are to be fully experienced. To ensure that a visual stimulus is subliminally experienced it is necessary to present the stimulus for a brief time, say 50 msec, and then flash another stimulus, called the ''masking stimulus'', on the screen at a controlled time afterwards so as to persist for, say 500 msec. The time between the offset of the first and onset of the masking stimulus, the so-called stimulus onset asynchrony (SOA), needs to be varied in steps of 5 msec to be effective in changing the level at which a given subject is able to perform at just above chance level in a lexical decision task (deciding whether the letter string composing the first stimulus is a word or not; Marcel, 1980).

Fully subliminal perception occurs in at least the modalities of touch, vision and audition, and there is a considerable psychological literature on the nature of these phenomena (Merikle, 1992).

The important problem of stimulus attention (so as to keep attention aroused even in the apparent absence of any stimulus and allow stimulus timing to be effective) may be approached by the following choice of paradigm: require the subject to respond, after the offset of the mask, to one of the questions:

1. was there a stimulus preceding the mask?
2. to which of two probe words (presented after offset of the mask) was that which preceded the mask most similar graphically?
3. to which word (as in the paradigm for (2) above) was the word preceding the mask most similar semantically?

In this manner, attention can be kept focused onto a given task (the final response) yet neither the subliminal stimulus nor the mask should be perturbed by the response of the subject.

This is valid for visual stimuli, but can be extended to somatosensory or auditory stimuli (in the former case with stimuli possibly delivered to one or other hand, or to two sites on the hand, and ask which hand or which site was stimulated, whilst for audition the stimulus could be delivered to either ear, with response to the question as to which ear).

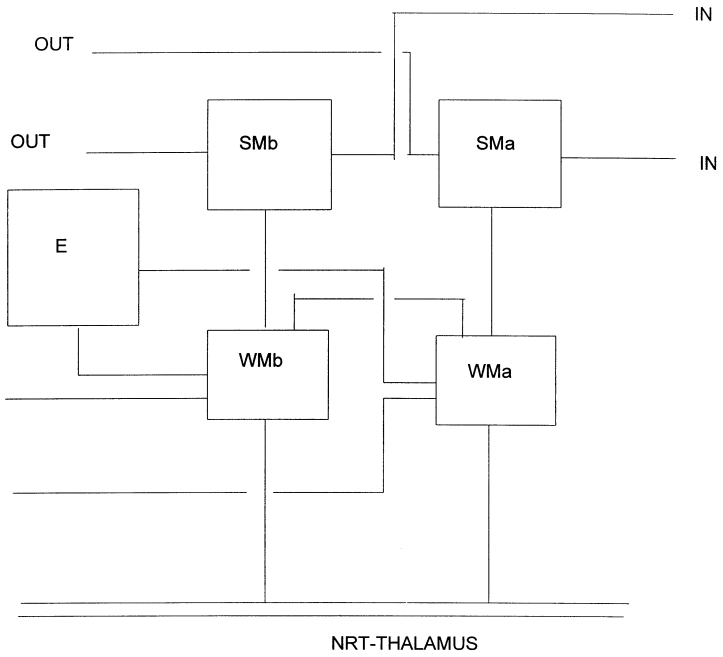


FIGURE 3. Coupling between two sets of two-stage modules of Figure 1 and Figure 2 to achieve competition between the activities on the second stages through the interaction of those stages with the NRT-thalamic system. There is also support from episodic memories encoded in the store denoted E in the figure.

3.3.3. *Partly-Subliminal or Unattended Stimuli.* A form of this paradigm particularly relevant to fMRI experimentation is that of two words presented to a subject, one foveally and the other parafoveally, for 120 msec, with attention being paid to the foveally presented word. At an inter-stimulus interval (ISI) of 200 or 2000 msec a target word is presented with the request to classify the stimulus as an animal or not by pressing one or other of two buttons, in the paradigm of DiPace et al. (1991). Reaction time measurements to the response at the earlier ISI should show priming from words semantically related to either the foveally or the parafoveally presented stimuli; at the later ISI only the foveally presented prime is effective. This was found to be so experimentally. The activation of the parafoveally presented word has died away before it has had time to reach awareness. This was also determined by asking the subjects if they had experienced the parafoveal word; in only 15% of cases did this happen.

There are many other stimuli which produce partly subliminal stimuli; in vision, the Stroop test has been mentioned. In audition, the use of a prime, which is attended to, and a probe which can then be used as a subliminal modifier of response to the attended prime, is common in lexical analysis, and again there is a very extensive literature. The accuracy of timing in these latter cases is only at the 20 msec level or above.

3.3.4. *Conclusions on Subliminal Stimuli.* There has been considerable analysis of the phenomenon of subliminal perception, in which there is knowledge acquired for inputs below a so-called ''objective threshold'' but only awareness when the input is above a higher ''subjective threshold'' (Cheesman and Merikle, 1986; Merikle, 1992). Thus the phenomenon would itself lead to the analysis of the emergence of perception by attempting to detect, under one of the numerous

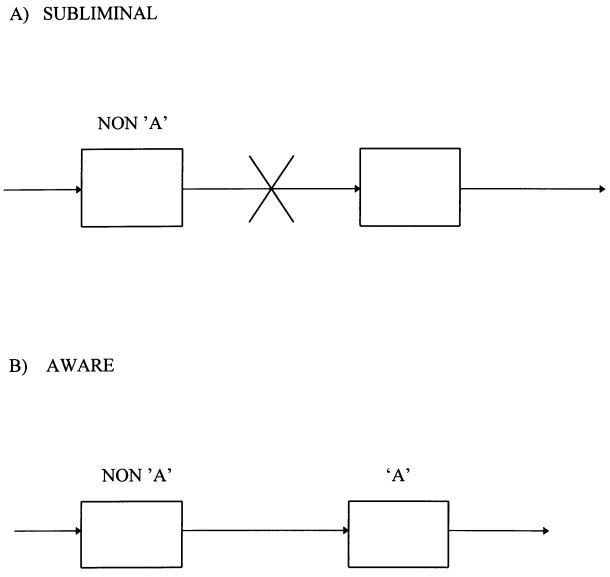


FIGURE 4. The difference between subliminal processing (a) and aware processing (b); in the former activity fails, for some reason (low level, degradation, backward masking), to reach the second, working memory stage of the construction of consciousness denoted in Figures 1–3.

TABLE 2

Paradigm	G/NG	Reason/Comments
1) Blindsight	G	But may have altered circuits
2) MAE	G	Area with longest duration of adaptation (& longer than psychophysical effects)?
3) Binocular rivalry	NG	Adaptation in module
Concept rivalry	NG	Below creation of ''A''?
4) Orientation-specific adaptation	NG	Adaptation not create ''A''
5) Double click	NG	Where is ''A''?
	G	If use non-attentive conditions in addition
6) Filling in	NG	As for double click
	G	If use non-attentive conditions in addition
7) Non-attentive	G	Must ensure full non-attention (but attentive ''A'')
8) Subliminal processing (many paradigms)	G	Must ensure only subliminal experience is occurring

G denotes ''good'', indicating that the paradigm is effective in discerning the location of the emergence of phenomenal awareness; NG denotes ''no good'').

paradigms in which it has been observed, the difference between those areas involved in the subliminal level of processing and those involved in the further creation of awareness in the subject.

The nature of these paradigms, in terms of the two-stage processing model of Figure 2, is as shown in Figure 4. The first figure shows purely subliminal processing, in which the awareness modules ''A'' of Figure 2 are not activated; the second figure indicates what happens when activity also arises at the aware level during conscious processing.

These seem to be the best of all possible paradigms, since they directly attack the separation of tasks into those involved with subliminal, just pre-conscious, activations from those in which there is awareness of the inputs.

In conclusion, these subliminal paradigms appear to be highly satisfactory. The optimal ones involve the fully subliminal stimuli of Section 3.3.2, since there is then less activity expected to arise at the aware level in the processing of the purely subliminal stimuli. In the partly subliminal stimulus case, there would be expected to be possible interference of the attended stimulus processing at the subliminal level. It would be relevant to develop paradigms in which such interference is expected to be minimised.

4. CONCLUSIONS

Of a range of experimental paradigms analysed to probe the emergence of perception, very few were found to satisfy the criteria presented above in Sections 2 and 3. A summary of the results of the above discussions is given in Table 2. In particular, the remaining successful experiments were:

1. MAE
2. subliminal processing
3. dissociated blindsight
4. comparison of activations caused by various inputs between attentive and non-attentive conditions.

Of these, (1) and (2) seem to be the most precise and clean, since they involve direct measurement of regions of long holding of the neural activity directly causing the emergence of the relevant percept. In (3), there is also the danger that blindsighted subjects could have rerouted their activity to other modules, so the neural sites discerned would not necessarily be those used by sighted subjects. In (4), there is the danger of only partially attended or unattended processing.

Given that there are two paradigms which would appear to be effective, we can conclude that there are some paradigms which satisfy all of the criteria of Sections 2 and 3. The next steps are (a) to perform and analyse the experiments in detail and (b) to model the detailed dynamics observed, so as to obtain a better understanding of the complex phenomena involved, and so as to make detailed predictions, from the models, in order to extend the experimental understanding.

REFERENCES

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

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

Baddeley, A. (1993). Working memory and conscious awareness. In S. E. Gathercole, A. F. Collins, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 11–28). Hillsdale, NJ: Erlbaum.

Bisiach, E., & Berti, A. (1995). Consciousness in dyschiria. In *The cognitive neurosciences* (chap. 88, pp. 1331–1340). Cambridge, MA: MIT Press.

Bullier, J. (1995). Parallel versus serial processing: New vistas on the distributed organization of the visual system. *Current Biology, 5,* 497–503.

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Can. J. of Psychology, 40,* 343–367.

DiPace, E., Longoni, A. M., & Zoccolotti, P. (1991). Semantic processing of unattended parafoveal words. *Acta Psychologica, 77,* 21–34.

Edelman, G. (1989). *The remembered present*. New York: Basic Books.

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

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

James, W. (1950). *The principles of psychology*. New York: Dover Books.

Joliot, M., Llinas, R., & Ribary, U. (1994). Human oscillatory activity near 40 Hz coexists with cognitive temporal binding. *Proc. Nat. Acad. Sci. (USA), 91,* 11748–11751.

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature, 377,* 336–338.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review, 98,* 122–142.

Logothetis, N.K., & Schall, J.D. (1989). Neuronal correlates of subjective visual perception. *Science, 245,* 761–763.

Marcel, A. (1980). Conscious and preconscious recognition on polysemous words: Locating selective effects of prior verbal contexts. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, NJ: Erlbaum.

Merikle, P. M. (1992). Perception without awareness. *American Psychologist, 47,* 792–795.

Pantev, C., Hoke, M., Lutkenhoner, & Lehnertz, K. (1989). Tonotopic organization of the auditory cortex: Pitch versus frequency representation. *Science, 246,* 486–488.

Paulesu, E., Frith, C., & Frakowiak, R. S. J. (1993). The neural correlates of the verbal components of working memory. *Nature, 362,* 342–345.

Pritchard, R.M. (1963). Stabilised images on the retina. *Scientific American, 204,* 72–78.

Salmon, E., Van der Linden, M., Collette, F., Delfiore, G., Maquet, P., Degueldre, C., Luxen, A., & Franck, G. (1996). Regional brain activity during working memory tasks. *Brain, 119,* 1617–1625.

Shah, J. (1997). Private communication.

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organisation. *Proc. Nat. Acad. Science (USA), 94,* 3408–3413.

Smith, E. E., & Jonides, J. (1995). Working memory in humans: Neurophysiological evidence. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (chap. 66, pp. 1090–1020). Cambridge, MA: MIT Press.

Taylor, J. G. (1973). Seminar, Tubingen University (unpublished).

Taylor, J. G. (1991). Can neural networks ever be made to think?. *Neural Network World, 1,* 4–11.

Taylor, J. G. (1992). Towards a neural network model of mind. *Neural Network World, 2,* 797–812.

Taylor, J. G. (1996a). Invited talk, WCNN96, San Diego, CA.

Taylor, J. G. (1996b). Breakthrough to awareness: A preliminary neural network model of unconscious perception in word processing. *Biol. Cybernetics, 75,* 59-72.

Taylor, J. G. (1998). *The race for consciousness.* Boston, MA: MIT Press.

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

Weiskrantz, L., Barbur, J. L., & Sahrie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proc. Nat. Acad. Sci. (USA), 92,* 6122–6126.

Young, A. W. (1994). Neuropsychology of awareness. In A. Revonsuo & M. Kamppinen (Eds.), *Consciousness in philosophy and cognitive neuroscience* (chap. 8, pp. 173–204). Hillsdale, NJ: Erlbaum.

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1175–1183, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00054-3

1997 SPECIAL ISSUE

Three Centuries of Category Errors in Studies of the Neural Basis of Consciousness and Intentionality

WALTER J. FREEMAN

University of California at Berkeley

(*Received* 2 *September* 1996; *accepted* 20 *January* 1997)

Abstract—*Recent interest in consciousness and the mind–brain problem has been fueled by technological advances in brain imaging and computer modeling in artificial intelligence: can machines be conscious? The machine metaphor originated in Cartesian ''reflections'' and culminated in 19th century reflexology modeled on Newtonian optics. It replaced the Aquinian view of mind, which was focused on the emergence of intentionality within the body, with control of output by input through brain dynamics. The state variables for neural activity were identified successively with animal spirits, élan vital, electricity, energy, information, and, most recently, Heisenbergian potentia. The source of dynamic structure in brains was conceived to lie outside brains in genetic and environmental determinism. An alternative view has grown in the 20th century from roots in American Pragmatists, particularly John Dewey, and European philosophers, particularly Heidegger and Piaget, by which brains are intrinsically unstable and continually create themselves. This view has new support from neurobiological studies in properties of self-organizing nonlinear dynamic systems. Intentional behavior can only be understood in relation to the chaotic patterns of neural activity that produce it. The machine metaphor remains, but the machine is seen as self-determining. © 1997 Elsevier Science Ltd.*

Keywords—Brain dynamics, Chaos, Consciousness, Existentialism, Information, Intentionality, Nerve energy.

1. INTRODUCTION

Studies of the neural basis of consciousness have recurred in the biomedical literature for 2500 years, beginning with Hippocrates:

> One ought to know that on the one hand pleasure, joy, laughter, and games, and on the other, grief, sorrow, discontent, and dissatisfaction arise only from [the brain]. It is especially by it that we think, comprehend, see, and hear, that we distinguish the ugly from the beautiful, the bad from the good, the agreeable from the disagreeable... Furthermore, it is by [the brain] that we are mad, that we rave, that fears and terrors assail us—be it by night or by day—dreams, untimely errors, groundless anxiety, blunders, awkwardness, want of experience. We are affected by all these things when the brain is not healthy, that is, when it is too hot or too cold, too moist or too dry, or when it has experienced some other unnatural injury to which it is not accustomed (Clarke and O'Malley, 1968, pp. 4–5).

The last strong outpouring from biologists came 40 years ago following the discovery of the midbrain and thalamic reticular activating systems and their roles in arousal and attention (Adrian et al., 1954). In the past two decades, contributions have come from researchers in brain imaging, psychology, psychiatry, neurology, philosophy, mathematics, physics, computer science, and artificial intelligence. As a result, a shift of immense magnitude is taking place in our understanding of ourselves, but none of us has the perspective yet to grasp its nature and significance. The limitation stems from the circumstance that the bulk of new data has been obtained within the confines of the machine metaphor for mind/brain function. This essay aims to explore the origin of that metaphor from what preceded it, and to indicate a new approach to mind/brain studies.

2. THE ORIGIN OF BRAIN DYNAMICS IN THE MACHINE METAPHOR

Behaviorists have a long history of using natural science

Acknowledgements: This work was supported by a grant MH06686 from the National Institute of Mental Health. The paper was read at the First Tucson Conference on ''Toward a Scientific Basis for Understanding Consciousness'' on 13 April 1994. It provided materials for Chapter 2 in ''Societies of Brains'' (1995), and is given here with the permission of the publisher, Lawrence Erlbaum Associates, Hillsdale, NJ.

E-mail:wfreeman@garnet.berkeley.edu.

to explain what they can observe among animals and their fellow humans. The Greeks invented a chemistry of the soul, in which the four elements (air, earth, fire and water) combined to make the four humors of the body (phlegm, blood, black bile and yellow bile), whence the Hippocratic temperaments: phlegmatic, sanguine, melancholic and choleric. Aristotle applied his physics to conceive the brain as a radiator to cool the blood. The foundation for a dynamics of behavior was laid by Descartes, who proposed that the brain operated as a pump for fluids termed ''animal spirits'', which flowed through the ventricles from the brain into the spinal cord and out into the muscles. The valves such as the pineal gland were controlled by the soul in humans but were automatic in animals, they being soulless machines.

The seminal importance of this formulation can be seen from the following standpoint. In a dynamical system some material thing moves with time. Since Newton and Leibniz the preferred description has been a differential equation, which relates time as it elapses independently (in either direction) to something in the system that is changing. That ''something'' is the state of the system, and its descriptor is a ''state variable''. It is essential to measure accurately both the time lapse and the state variable, because the equation gives the relation between the numbers that represent time and the state variables. In Cartesian studies of the brain, the state variable would have represented the flow of animal spirits, had the equations been written. No one did this. The difficulty was that there was no way to measure the flow of animal spirits through the nerves. Descartes postulated that the muscles were shortened by being pumped full of the fluid like a balloon. Physiologists tested this prediction by inventing the plethysmograph to measure the volume of muscle before and during contraction. There was no increase but, in fact, a slight decrease owing to expulsion of blood from the veins by muscle contraction, which showed that animal spirits had no volume to measure. Hence animal spirits could not be represented by a string of numbers in a model. This is an example of what Gilbert Ryle (1949) called a category error, in this case the assignment of the physical property of volume to an immaterial entity. There were more such errors to follow.

Late in the 18th century a Bohemian ophthalmologist named Giri Prochaska made a remarkable discovery. Until that time most scientists had assumed that the brain was a source of animal spirits. Prochaska observed the behavior of newborn anencephalics and found their behavior to be entirely normal. As we now know the human cerebrum is essentially nonfunctional at birth. His findings led him to propose that sensory receptors are the source of animal spirits, which are released by the action of stimuli from the environment. He then drew explicitly on Newtonian optics to formulate a theory of reflex action:

> The reflection of sensory impressions into motor...is not performed solely according to physical laws...but follows special laws inscribed, as it were, by nature on the medullary pulp... The general law...is that of our preservation: so that certain motor impressions may follow external impressions about to harm our body and produce movements aimed at warding off and removing the harm...or sensory impressions about to be favorable to us, and produce movement tending to preserve that pleasant condition longer (Prochaska, 1784, p. 116).

The brain was seen as complex but passive mirror. ''Reflections'' became reflexes. Three decades later Prochaska made another category error in identifying his ''sensory impressions'' with electricity newly discovered by Galvani and Volta. This hypothesis was disputed by Carlo Matteucci, who maintained that nerves carried spiritual force, that came to be identified with ''élan vital''.

3. NERVE ENERGY REPLACES VIS NERVORUM

The hegemony of physics was re-established by the Young Turks. Du Bois-Reymond discovered the injury current and the ''negative variation'' (the nerve action potential). Helmholtz measured its conduction velocity. Sechenov developed an animal model for anencephaly by pithing a frog to study reflexes. The centerpiece of this antivitalist movement was the experimental demonstration of the First Law of Thermodynamics, the conservation of energy. In the grandest category error of them all, the animal spirits and élan vital were replaced with nerve forces and nerve energies, which flowed from the environment through the sensory receptors into the brain and back out again through the muscles, after being stored in nerve cells and then ''liberated''. The new doctrine was announced by Herbert Spencer (1863, p. 109):

> [It is]...an unquestionable truth that, at any moment, the existing quantity of liberated nerve-force, which in an inscrutable way produces in us the state we call feeling, must expend itself in some direction—must generate an equivalent manifestation of force somewhere... [A]n overflow of nerve-force, undirected by any motive, will manifestly take the most habitual routes; and, if these do not suffice, will next overflow into the less habitual ones.

Charles Darwin (1872, p. 70), continued:

> This involuntary transmission of nerve force may or may not be accompanied by consciousness. Why the irritation of nerve-cells should generate or liberate nerve force is not known; but that this is the case seems to be the conclusion arrived at by all the greatest physiologists such as Mueller, Virchow and Bernard, and so on.

The application of Newtonian dynamics was also explicit in the writings of J. Hughlings Jackson (1884,

pp. 42–44):

> ...we speak of the dynamics of the nervous system... A normal discharge starting in some elements of the highest centres overcomes the resistance of some of the middle, next the resistance of some of the lowest centers, and the muscles are moved... A fit of epilepsy is an excessive caricature of normal physiological processes during what is called a voluntary action... We have, in the case of ''discharging lesions,'' to consider not only the quantity of energy liberated, but the rate of its liberation... *Resistances* will be considered later.

A note in Jackson's handwriting was later found in the margin of this text: ''No more of this was published.'' It may have already become clear that while nerve tissue did have electrical resistance, the barrier to the flow of ''nerve energy'' was not so simple. The principle of the conservation of momentum was also used by Sigmund Freud as a foundation for his project of a scientific psychology, in which he confused his neuronic inertia with dendritic current:

> This line of approach is derived directly from pathological clinical observations, especially those concerned with excessively intense ideas... These occur in hysteria and obsessional neurosis, where, as we shall see, the quantitative characteristic emerges more plainly than in the normal... What I have in mind is the principle of neuronic inertia, which asserts that neurones tend to divest themselves of quantity (Q)... We arrive at the idea of a ''cathected'' neurone (N) filled with a certain quantity... The principle of inertia finds expression in the hypothesis of a current, passing from the cell-processes or dendrites to the axone... The secondary function [memory] is made possible by supposing that there are resistances which oppose discharge...in the contacts [between the neurones] which thus function as barriers. The hypothesis of ''contact-barriers'' is fruitful in many directions (Freud, 1895, pp. 356–359).

Two years later these barriers were named by Michael Foster and Sir Charles Sherrington:

> Such a special connection of one nerve-cell with another might be called a synapsis (Foster and Sherrington, 1897, p. 929).

Some four decades later the hypothesis of synaptic resistance was undermined by Otto Loewi's discovery of chemical neurotransmission, though it persists in more realistic treatments of electrical synapses, in which it refers to electrical current and not to nerve energy.

Another physical principle, the field of potential that was developed by Michael Faraday to explain electrical and magnetic forces, was coopted by Gestalt psychologists to explain their data from studies in perception:

> ...let us think of the physiological processes not as molecular, but as molar phenomena... Their molar properties will be the same as those of the conscious processes which they are supposed to underlie (Koffka, 1935, p. 57).

> The task of psychology...is the study of behavior in its causal connection to the psychophysical field (Koffka, 1935, p. 67).

The environment was conceived as a source of nerve energy, which flowed through the sensory receptors into the brain with striking motivational consequences:

> ...things in our environment tell us what to do with them... Their doing so indicates a field of force between these objects and our Egos...which...leads to action. ...A handle wants to be turned, ...chocolate wants to be eaten, ... (Koffka, 1935, p. 353).

One of the principle architects of Gestalt psychology, Wolfgang Köhler, presented meticulous studies of interactive phenomena in perception:

> Our present knowledge of human perception leaves no doubt as to the general form of any theory which is to do justice to such knowledge: a theory of perception must be a field theory. By this we mean that the neural functions and processes with which the perceptual facts are associated in each case are located in a continuous medium (Köhler, 1940, p. 55).

He pressed further into physics by identifying the perceptual fields with the electrical fields of the newly discovered electroencephalogram. This hypothesis was identical in form to the category error of Prochaska. It was quickly disproved by Roger Sperry (1958), who placed strips of mica and silver needles in the visual cortex of trained cats and monkeys and showed that the resulting distortions in electrical fields had negligible effects on behaviors involving visual perception. Unfortunately, for this and other reasons, the body of Gestalt theory was discredited among neurobiologists.

With continuing advances in the analysis of anatomical pathways in the cerebrum it became increasingly obvious that the concept of mass flow of energy made no sense. According to Lashley (1942, pp. 302–306):

> Generalization [stimulus equivalence] is one of the primitive basic functions of organized nervous tissue. ...Here is the dilemma. Nerve impulses are transmitted...from cell to cell through definite intercellular connections. Yet all behavior seems to be determined by masses of excitation... What sort of nervous organization might be capable of responding to a pattern of excitation without limited specialized paths of conduction? The problem is almost universal in the activities of the nervous system.

He had already noted the difficulty of finding useful concepts:

> ...expressions like mass action, stress patterns, dynamic effects, melodies of movement, vigilance, or nervous energy [are] all highly metaphorical and unproductive of experimental problems (Lashley, 1929, p. 254).

Yet he continued to borrow from the physical sciences

fuzzy concepts such as ''reverberatory circuits'', ''equivalent nervous connections'', ''systems of space coordinates'', ''wave interference patterns'', ''tuned resonating circuits'', etc. (Lashley, 1950).

4. INFORMATION REPLACES NERVE ENERGY

The repeated failure of the energy metaphor opened the way for a new approach that came from the communication sciences. Basing their work on Golgi analyses of the entorhinal cortex by Rafael Lorente de Nó (1934), Warren McCulloch and Walter Pitts introduced the concept of nerve cells operating as binary switches in neural networks to compute Boolean algebra. John von Neumann used this concept to develop programmable digital computers. Shannon and Weaver developed the theory of information by *divorcing it from meaning*. This led to the replacement of ''energy'' by ''information'' as a descriptor of neural activity. Information and energy are both conceived as flows from environmental ''sources''. They are transduced through sensory systems, transmitted by axonal tracts as channels, carried by action potentials (bits), transformed (processed) in brains by synapses working as binary switches, stored as fixed patterns (representations), recalled by read-out under constraints of finite channel capacities and entropic losses, like the content addressable memories in computers, and matched or cross-correlated with new input patterns.

Information metaphors are found at four levels. At the level of nerve cells the single neuron is seen as generating a pulse train to represent a meaning corresponding to the Logical Positivist element of a word or an object, such as a grandmother. The frequency of the train represents the probability that the object is present (Barlow, 1972). Members of a distributed collection of neurons that symbolize the same object are called ''cardinal cells'', deriving from a College of Cardinals running the brain, as distinct from the pontifical cell decried by Sherrington (1940, pp. 177–178):

> In the great head end which has been mostly darkness spring up myriads of twinkling stationary lights and myriads of trains of moving lights of many different directions... The brain is waking and with it the mind is returning. It is as if the Milky Way entered upon some cosmic dance. Swiftly the head-mass becomes an enchanted loom where millions of flashing shuttles weave a dissolving pattern though never an abiding one; a shifting harmony of subpatterns.

An alternative formulation holds a neuron to be a ''feature detector'' by virtue of its afferent synaptic connections, which are modified and adaptively shaped by prior learning. A collection of feature detectors defines an object when their pulse trains become synchronized, a proposed solution to ''the binding problem'' (Milner, 1974; von der Malsburg, 1983) of getting feature detectors to work together. Large arrays of such neurons form Hebbian nerve cell assemblies and neural networks, which provide the basis for neurocomputation or computational neural science. A well-known example is the tensor system for coordination of hand movements under visual guidance by cerebellar circuitry developed by Pellionisz and Llinás and reviewed by Churchland (1986).

At the behavioral level and among successors of the Gestalt school, most notable is J.J. Gibson, in whose work the ''affordances'' denote information that flows into the brain from outside the body through exteroceptors and from inside the body through the proprioceptors:

> ...the affordance, being invariant, is always there to be perceived. An affordance is not bestowed upon an object by a need of an observer and his act of perceiving it. The object offers what it does because it is what it is. ...But this does not in the least imply separate realms of consciousness and matter, a psychophysical dualism. It says only that the information to specify the utilities of the environment is accompanied by information to specify the observer himself... [E]xteroception is accompanied by proprioception...to perceive is to coperceive oneself (Gibson, 1979, p. 139).

Information is delivered into resonant circuits in the brain, and it flows out again as ''effectivities'' from muscles and glands in object-oriented actions. According to Shaw et al. (1990, pp. 586–587):

> Gibson, like Tolman, would disagree with [the view of] Skinner...that the organism is merely a ''through-put system''. For Tolman, cognition can embellish the stimulus, while for Gibson, stimulus must be informative about the environment in ways that a stimulus, as a physiological ''goad;'' or a reflexive ''force'', could never be. They both endow [the organism] with a complex interior—which Tolman cites as the residence of cognitive functions and Gibson as the seat of a tunable (not necessarily linear) information detection operator which resonates to qualitative environmental properties (i.e. affordances). For Gibson, the environment that surrounds an organism is real and objective for each given organism.

The metaphorical ''resonance'' is reminiscent of Karl Lashley's ''tuned resonating circuits''. It is metaphorical, because no physiological embodiment has yet been demonstrated. The difficulties of pursuing this line of theory have been well formulated by Shaw and his colleagues. For example, the infinite complexity of ''objects'' in unstructured environments is treated by converting objects to numbers:

> ...by following Cantor's fractalization rule, we have a way to rescale continuous geometric objects so that their dimensionality is reduced. Moreover, by following it with finite recursion, we find that there are objects without integer dimensions... These are Mandelbrot's fractals... (Shaw and Kinsella-Shaw, 1988, pp. 197–198).

They resolve the paradox that is inherent in combining past experiences and future goals by postulating a dual

Minkowski space with the two cones of past and future melded at the apex of the present, and they propose:

> ...an environment of information which dynamically links a socially invariant exterior with both a biologically invariant interior frame, on the one hand, and with a still more exterior physically invariant frame on the other. That psychological inverse dynamics must couple energy with information across a frame exterior (observable) to one and interior (controllable) to the other, and vice-versa, defines what is meant by an ecological map (Shaw et al., 1990, p. 587).

Their proposed exploration of the role of intentionality in purposive behavior by ecological psychology will depend heavily on the development of graphical computer displays for the three levels of maps and the derivation of equations to describe the operations required for transfer of information between them. At present, their efforts are remote from direct observations made on the dynamics of brains.

At the subneuronal level, the discovery of DNA as the basis for transmission of genetic information has also stimulated search for stores of experiential information in the form of ''memories'' in RNA molecules as well as synapses. The search for the ''molecular alphabet'' of learning is now among the hottest areas of neurobiological research (Alkon, 1992), although studies of synaptic modification have not yet progressed beyond habituation, facilitation, and simple go/no go reflex arcs. Holger Hyden demonstrated a change in RNA in the brains of rats trained to climb a wire and suggested that it indicated the storage of a procedural memory in the neurons of the vestibular nuclei. This line of thinking culminated in studies by worm runners to transfer the memory of working a T-maze from trained Planarian flatworms to their naive, cannibalistic siblings. After initial success this hypothesis failed in the hands of trained scientists, but it is still being ''replicated'' in high school science fairs annually across the country.

At the submolecular level is a variant of the information hypothesis, in which ''quanta of sentience'' emerge as Heisenbergian potentia from a ''sea of universal consciousness'' (Herbert, 1993, p. 26):

> Though materialists agree that mind (defined as ''inner experience'') is no more than a particular motion of matter, they differ concerning how complex matter's movement must be actually to produce a noticeable sensation, to generate what might be called a ''quantum of sentience''...analogous to physicist Max Planck's famous quantum of action.

The main hope that drives these investigations is for discovery of new laws of physics (Penrose, 1989), which will explain such paranormal phenomena as teleportation, precognition, distance viewing, and related forms of extrasensory perception (Herbert, 1993, p. 248):

> Most quantum models of consciousness are similar to Crookes's coherer proposal in that they consider the synapse to be a sensitive receiver of mental messages that originate outside the brain. The main difference between the coherer model of mind and quantum consciousness models is that...-mind is somehow resident in Heisenberg's quantum potentia rather than in electromagnetic ether.

Criticisms that brains, neurons, organelles and receptor molecules are neither small enough nor cold enough to afford quantum coherence have been met with the rejoinder that superconductivity is a macroscopic state of coherence that has already been achieved at temperatures approaching Siberian winter nights, and that it may soon be found also at normal brain temperatures. Less easily handled have been the criticisms that, for all its power in chemistry, quantum mechanics is a linear, first-order, discrete approximation for protons, that it is even inadequate to describe the collapse of the wave function, and that it is poorly suited for describing the nonlinear continuous time dynamics displayed by the nervous system at all levels. More to the point, this line of thought is new wine in an old bottle. The same properties are invoked as for energy and information: environmental sources of input, sinks for output, tuned receptors to resonate with selected inputs, and connectionist mechanisms for storage and retrieval. The elemental building blocks (reflexes, action potentials, bits, words, symbols, primitives, and quanta) change with the centuries, but the underlying concepts have been passed whole from one generation to the next.

5. THE UNIQUENESS OF BRAIN FUNCTION

Three insights are lacking from these input–output approaches to brain dynamics. The first insight is that the tissue formed by neurons in animal brains is unique. There is no other substance like neuropil in the known universe. Phylogenetically it has emerged by evolution repeatedly and independently in the brains of molluscs, crustaceans, and vertebrates, always as the basis for adaptive, goal-directed behavior. Being unlike anything else, it offers us the opportunity to discover its ''laws'', which might constitute the ''new laws of physics'' sought by Penrose (1989). Second, it follows that the machine metaphor cannot be serve to identify brain state variables with the state variables of any other machine. Third, brains organize their own goals, which machines cannot now do.

The second insight holds that, while neural activity is based in flows of transmitter molecules, inorganic ions, and electric currents fueled by metabolic energy and controlled by conformational structural changes in cell membranes, and while it carries both meaning and information, it cannot be defined by any of these physical or conceptual quantities. As noted by Karl Lashley the terms ''nerve force'', ''nerve energy'', ''information'', and ''representation'' are *metaphors* and not measurable descriptors of brain events. The spatiotemporal dynamic patterns of neuroactivity are observed by its

electrochemical manifestations (''signs'' according to Adrian, 1947), but neuroactivity cannot be observed directly, and models of it must include transforms that describe postulated relations of the signs to the activity (Freeman, 1975). The contents of the signs are inferred from behavioral correlates and introspection (Freeman, 1991, 1995), which are used to characterize and model the operations by which brains construct *meanings* from sensory stimulation. *This is not ''information processing''*. The problem is, how to define neuroactivity? Just as ''force'' in physics is defined as a relation between mass, time and distance, ''neural activity'' must be defined by relations between its electrochemical signs and overt, measured behaviors. Neuroactivity does not flow across the receptors, the muscles, or the blood–brain barrier as energy, matter and information do. Brains are closed systems with respect to meaning, though not to energy or information. This enclosure explains the inaccessibility of qualia between brains, because the feelings and the associations in experiences that come in each brain with expectations, actions, and stimuli are rooted in its undivided body of past learning and its present chemical state.

The third insight is that the patterns of neural activity are endogenous. Their structure emerges from within and is not imposed solely by flows of energy, information or quanta from the bodies. The theory of chaos in nonlinear dynamical systems offers a set of techniques for describing the conditions required for emergence of self-organized patterns (Freeman, 1992, Skarda and Freeman, 1987). The process of self-determination is inherent in the Aquinian concept of intentionality (Martin, 1988, Pegis, 1948) by which each soul (mind/brain) intends (''stretches forth'') outside itself and into the world. It acquires knowledge by shaping itself (learning) in accordance with the effects (sensory feedback) of its endogenous action. The soul creates itself and its virtue by its own actions. Descartes discarded this medieval doctrine and mechanized the body by relegating perceptual events to the status of ''representations'' of the world, so the soul understood through logic, not prelogical learning. Kant deliberately revolutionized the mechanization by postulating that the rules of knowing were embedded as absolute ideas in human nature. He had no framework in Newtonian science to cope with the emergent processes of intentionality. Brentano and Husserl reintroduced the word as denoting what the Kantian representations were ''about'', whether or not the objects or events so represented actually existed in the world (Freeman, 1995). This meaning now constitutes the mainstream interpretation among analytic philosophers (Putnam, 1990).

6. NONLINEAR BRAIN DYNAMICS AND NEUROEXISTENTIALISM

The neo-Aristotelian philosophy of the later Middle Ages brooked no doubts about the unique causal efficacy of each person. The machine metaphor undermined that certainty and induced conflicts between ideas of free will and universal determinism that persist into the modern era. Several philosophers in the present century constructed powerful theories that avoided the machine metaphor and did not lead to the cruel dichotomy. The earliest were John Dewey with American pragmatism, Henri Bergson with ''Creative Evolution'', Martin Heidegger with ''Being and Time'', and Jean Piaget with developmental psychology. Sir Frederic Bartlett (1932) described the problem from the viewpoint of his studies on remembering:

> ...some widely held views [of memory] have to be completely discarded, and none more completely than that which treats recall as the re-excitement in some way of fixed and changeless ''traces'' (p. vi).

> The picture is one of human beings confronted by a world in which they can live and be masters only as they learn to match its infinite diversity by increasing delicacy of response, and as they discover ways of escape from the complete sway of immediate circumstances (p. 301).

> There is one way in which an organism could learn how to do this. It may be the only way... An organism has somehow to acquire the capacity to turn round upon its own ''schemata'' and to construct them afresh. This is a crucial step in organic development. It is where and why consciousness comes in; it is what gives consciousness its most prominent function... I wish I knew exactly how it was done (p. 206).

Jason Brown (1977) described it from a philosophical viewpoint:

> The structural organization of cognition is no less dynamic than the psychological systems it supports... The incessant flow of cognition, the continual appearance and disappearance of new form at each moment of our waking and sleeping life, are manifestations of the activity of the structure as a whole as it achieves one or another level of realization (pp. 2–11).

> Affect is not an energy that invades and charges an idea... There is no need for the concept of psychic energy (instinct, motivation) as a motivating force in cognition. The orderly sequence and unfolding of cognitive levels repeats and extends the phylogenetic and ontogenetic pattern. The progression from depth to surface, the incessant repetition of developmental form, and the striving toward higher levels are all part of an evolutionary trend that leads in a forward direction simply because it is in the nature of the organization to unfold in this manner (pp. 127–133).

Ilya Prigogine (1980) has applied his theory of ''dissipative structures'', which feed on energy and evolve complex patterns in states that are far from equilibrium, to understand the nonlinear dynamics of brains. Hermann Haken (1983) has applied his theory of synergetics to

comprehend the physical principles, by which masses of neurons can interact to generate spatiotemporal patterns of activity.

Only one of several theories was effectively linked to neurobiology by its author. Maurice Merleau-Ponty (1942, 1945) drew heavily on work in clinical neurology from the First World War, particularly describing the phenomena now known as phantom limb and sensory neglect in brain-damaged individuals. With careful reasoning over voluminous biological details, he discarded the ''materialist'' view that minds were the results of linear causal chains of reflexes and chemical reactions. He likewise dismissed the ''idealist'' Kantian and Husserlian views of minds consisting of collections of representations, that were processed according to logical algorithms. He proposed instead the ''existentialist'' view, taken from his teacher, Heidegger, and classmate, Sartre: mind is ''the structure of behavior'', that creates itself by circular causality in its own ''action–perception cycle'' (Freeman, 1995).

7. THE NEUROBIOLOGY OF INTENTIONALITY

These several theories provide the warp and woof with which to weave the pattern of a strong neurobiological theory of self-organization of mind and brain. Biologists offer observations of the space–time patterns from brain imaging of human and animal brains during the performance of intentional behavior. Psychologists offer the measurements and analyses of behavior, in order to provide the essential behavioral structures that are to be correlated with the brain data. Physicists offer the dynamical systems theory by which to model the data and verify the capacity for brains to create and evolve their own unique space–time patterns of neural activity. Philosophers offer the conceptual framework required to bring the large picture into focus. The question remains: how do brains work? After three centuries of dynamics, answers are still elusive.

Nonlinear dynamics gives the technical tools needed to learn how it is done. The concept of the self-organizing brain, with its ever-shifting basins and attractors, its evolving trajectories (Tsuda, 1991), and its global cooperativity, enables us to model brain functions that transcend the present limitations of computational and representational schemata, and enter into those domains of nonrational and nonlogical construction from which consciousness emerges.

The complementary foundation in the mental sciences has been built in the past century by outstanding philosophers and psychologists, who can be grouped under the term ''existentialists'', and whose work has remained outside the main stream of modern neurobiology, owing to the unsolved problem of self-organization. The essential message of existentialism is that humans—and animals—create themselves by their actions. This insight has been arrived at independently in the 20th century by (among others):

Dewey (1914): Pragmatism—''Actions are not reactions to stimuli; they are actions into the stimuli.''
Heidegger (1927) and his students, Sartre and Merleau-Ponty (1942): Existentialism—''Mind is the structure of behavior.''
Koffka (1935): Gestalt psychology—''a field of force between objects and our Egos...leads to action.''
Piaget (1930): The cycle of ''action, assimilation, and adaptation'' in the sensorimotor stage of childhood development.
Gibson (1979): Ecopsychology—''An affordance...of an object offers what it does because it is what it is.''

In each of these systems sensation takes place as part of an ''action–perception cycle'' that Merleau-Ponty (1942) described as ''circular causality'' to contrast it with the ''linear causality'' of conditioned reflex chains and machine metaphors of brain function, such as clocks, telegraph nets, thermodynamic engines, chemical reaction systems, computers and holographs. Animals and humans receive and perceive stimuli as the end result of goal-oriented search for knowledge in the environment, and they learn about the world and shape themselves accordingly entirely in terms of the consequences of their own actions. The word ''intentionality'' has three widely accepted meanings. In analytic philosophy it means that a thought, belief, word, phrase or mental act is ''about'' something, whether an object or a person or a state of affairs, whether in the world or in the mind. In the psychological sciences it means that a thought, action or speech has a purpose, goal or intent, which is both outwardly directed toward manipulating objects in the world and inwardly directed toward satisfying biological drives, needs or instincts. In medicine it refers to the process of healing from injury, the re-establishment of wholeness of the body (Freeman, 1995). All the meanings stem from Medieval philosophy, which was synthesized in the 13th century by Aquinas. The mind is conceived as having unity that serves to distinguish itself from nonself; wholeness that expresses its direction of growth to maturity and the full realization of its potential; and intent (''stretching forth''), by which mind thrusts itself into the nonself by the actions of its body, and learns about the world by shaping itself in accordance with the outcomes of its actions, namely by learning from the sensory stimuli that were sought by its own actions (Freeman, 1995).

The neural mechanisms for intentionality in invertebrate animals and humans clearly reside in the limbic system. The evidence for this conclusion comes from diverse areas of study of animal and human behavior. Comparative neuroanatomists have shown that the forebrain of the most primitive surviving vertebrates representative of the ancestral line is composed of the essential sensory, motor and associational parts

of the limbic system, including the primordial hippocampus, septum and amygdala (Herrick, 1948). Selective lesions brains have shown that removal of the entire neocortex but sparing the ancient limbic structures impairs sensory guidance and elaboration of behavior, but the impoverished actions are clearly identified as intentional (Broca, 1973, Goltz, 1874). Intentional actions must take place within a space–time matrix for spatial orientation (the ''cognitive map'') and temporal integration (''short-term memory''). Electrophysiological investigations of the hippocampus (O'Keefe and Nadel, 1978) combined with studies of selective lesions of the hippocampal formation (Milner, 1966) have shown the importance of the limbic system for this matrix. An essential role in intentional action is played by ''corollary discharge'' (Sperry, 1950) and reafference (Kay, 1994) first identified by von Helmholtz (1879), which clearly is focused in the entorhinal cortex through its dense reciprocal connections both with the hippocampus and with all primary sensory cortices (Lorente de Nó, 1934) and the frontal lobes (Freeman, 1995).

8. CONCLUSIONS

Electrophysiological studies of visual, auditory, somatic and olfactory EEGs (Barrie et al., 1996) have shown that spatial patterns of neural activity emerge by construction with each act of perception, and that they depend on the context of the present and equally on the past experience of each subject, not merely on the stimuli. This is a reflection of the unity of function of the forebrain.

The implication is that each perceptual act has been organized in the context of the present state of the limbic system, expressing a desired state that is elaborated into a plan of action and an expectancy of the sensory consequences of that action (Freeman, 1995). All past experience is available and operating in the global dynamical state of the forebrain with each act in a continual succession of acts. The objective aspect that is observable by behaviorists is the flexible and adaptive sequence of acts from which intent is inferred. The subjective aspect is consciousness within the individual of the context that is brought to bear in each moment of choice with each act.

By this interpretation there is little further to be said about the biology of consciousness, because animals cannot describe their states of awareness using language, and the requisite electrophysiological studies in humans are not admissible. Problems abound in the biology of intentionality, including the dynamics by which structure emerges from the chaos of neural activity, by which limbic activity patterns are re-shaped into commands into the motor systems and corollary discharges into the sensory systems, and how the chain of nuclei in the brain stem that provide the neurohormones bathing the entire forebrain are integrated into the intentional structure of brain activity. These may not constitute ''the hard problem'' of Chalmers (1996), but as several authors have commented (Hameroff et al., 1996), they are hard enough for this generation of researchers.

REFERENCES

Adrian, E. D. (1947). *The physical background of perception*. Oxford, UK: Clarendon Press.

Adrian, E. D., Bremer, F., & Jasper, H. H. (1954). *Brain mechanisms and consciousness*. Oxford, UK: Blackwell.

Alkon, D. (1992). *Memory's voice: Deciphering the mind–brain code*. New York: HarperCollins.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1*, 371–394.

Barrie, J. M., Freeman, W. J., & Lenhart, M. (1996). Spatiotemporal analysis of prepyriform, visual, auditory and somesthetic surface EEGs in trained rabbits. *Journal of Neurophysiology, 76*, 1–20.

Bartlett, F. C. (1932). *Remembering*. Cambridge, UK: Cambridge University Press.

Broca, P. (1973). *Mémoires d'Anthropologie*. Paris: Reinwald.

Brown, J. (1977). *Mind, brain and consciousness*. New York: Academic Press.

Chalmers, D. J. (1996). Facing up to the problem of consciousness. In S. R. Hameroff, A. W. Kaszniak, A. C. Scott (Eds.), *Toward a science of consciousness* (chap. 5, pp. 5–28). Cambridge, MA: MIT Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind–brain*. Cambridge, MA: MIT Press.

Clarke, E., & O'Malley, C. D. (1968). *The human brain and spinal cord: A historical study illustrated by writings from antiquity to the twentieth century*. Los Angeles: University of California Press.

Darwin, C. (1872). *The expression of emotion in man and animals*. London: Murray.

Dewey, J. (1914). Psychological doctrine in philosophical teaching. *Journal of Philosophy, 11*, 505–512.

Foster, M., & Sherrington, C. S. (1897). *A textbook of physiology. III: The central nervous system* (7th ed., p. 929). London: MacMillan.

Freeman, W. J. (1975). *Mass action in the nervous system*. New York: Academic Press.

Freeman, W. J. (1991). The physiology of perception. *Scientific American, 264*, 78–85.

Freeman, W. J. (1992). Tutorial in neurobiology: From single neurons to brain chaos. *International Journal of Bifurcation and Chaos, 2*, 451–482.

Freeman, W. J. (1995). *Societies of brains: A study in the neuroscience of love and hate*. Hillsdale, NJ: Lawrence Erlbaum.

Freud, S. (1895/1954). The project of a scientific psychology. In M. Bonaparte, A. Freud, E. Kris (Eds.), E. Mosbacher, J. Strachey (Trans.), *The origins of psycho-analysis*. New York: Basic Books.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goltz, F. (1874). Über die funktionen des Lendenmarks des Hundes. *Pflügers Archiv gesamte Physiologie, 8*, 460–498.

Haken, H. (1983). *Synergetics: An introduction*. Berlin: Springer.

Hameroff, S. R., Kaszniak, A. W., & Scott, A. C. (Eds.), (1996). *Toward a science of consciousness*. Cambridge, MA: MIT Press.

Heidegger, M. (1927/1962). *Being and time* (J. Macquarrie, E. Robinson, Trans.). New York: Harper.

von Helmholtz, H. (1879/1925). *Treatise on physiological optics: Vol. 3. The perceptions of vision* (J.P.C. Southall, Trans.). Rochester, NY: Optical Society of America.

Herbert, N. (1993). *Elemental minds: Human consciousness and the new physics*. New York: Dutton, Penguin.

Herrick, C.J. (1948). *The brain of the tiger salamander*. Chicago, IL: University of Chicago Press.

Jackson, J. H. (1884/1958). Evolution and dissolution of the nervous system. Lecture III. In J. Taylor (Ed.), *Selected writings*. New York: Basic Books.

Kay, L. (1994). Distribution of gamma and beta oscillations in olfactory and limbic structures during olfactory perception in rats: Evidence for reafference. *Proceedings of the World Conference on Neural Networks, WCNN'94* (vol. 2, pp. 675–680).

Koffka, K. (1935). *Principles of gestalt psychology* (pp. 7, 353). New York: Harcourt Brace.

Köhler, W. (1940). *Dynamics in psychology*. New York: Grove Press.

Lashley, K. S. (1929). *Brain mechanisms of intelligence*. Chicago: University of Chicago Press.

Lashley, K. S. (1942). The problem of cerebral organization in vision. In J. Cattell (Ed.), *Biological symposia VII* (pp. 301–322). Lancaster, PA: Cattell Press.

Lashley, K. S. (1950). In search of the engram. *Symposia. Society of Experimental Biology*, *4*, 454–482.

Lorente de Nó, R. (1934). Studies in the structure of the cerebral cortex: I. The area entorhinalis. *Journal von Psychologie und Neurologie*, *45*, 381–438.

Martin, C. (Ed.) (1988). *The philosophy of Thomas Aquinas: Introductory readings*. New York: Routledge.

Merleau-Ponty, M. (1942/1963). *The structure of behavior* (A. L. Fischer, Trans.). Boston, MA: Beacon Press.

Merleau-Ponty, M. (1945/1962). *Phenomenology of perception* (C. Smith, Trans.). New York: Humanities Press.

Milner, B. (1966). Amnesia following operation on the temporal lobes. In C. W. M. Whitty, O. M. Zangwill (Eds.), *Amnesia* (pp. 109–133). London: Butterworths.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review*, *81*, 521–535.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, UK: Clarendon.

Pegis, A. C. (Ed.) (1948). *Introduction to Saint Thomas Aquinas*. New York: Modern Library.

Penrose, R. (1989). *The emperor's new mind*. Oxford, UK: Oxford University Press.

Piaget, J. (1930). *The child's conception of physical causality*. New York: Harcourt Brace.

Prigogine, I. (1980). *From being to becoming: Time and complexity in the physical sciences*. San Francisco: W.H. Freeman.

Prochaska, G. (1784). *Adnotationum academicarum*. Prague: W. Gerle. See also Unzer, J. A. (1851). *The principles of physiology by George Prochaska* (T. Laycock, Trans.). London: Sydenham Society.

Putnam, H. (1990). *Realism with a human face*. Cambridge, MA: Harvard University Press.

Ryle, G. (1949). *The concept of mind*. New York: Barnes and Noble.

Shaw, R. E., & Kinsella-Shaw, J. (1988). Ecological mechanics: A physical geometry for intentional constraints. *Human Movement Science*, *7*, 155–200.

Shaw, R. E., Kugler, P. N., & Kinsella-Shaw, J. M. (1990). Reciprocities of intentional systems. In R. Warren, A. Wertheim (Eds.), *Control of self-motion* (chap. 22, pp. 579–619). Hillsdale, NJ: Lawrence Erlbaum.

Sherrington, C.S. (1940/1951). *Man on his nature* (2nd ed.). Cambridge, UK: Cambridge University Press.

Skarda, C.A., & Freeman, W.J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, *10*, 161–195.

Spencer, H. (1863). *Essays: Moral, political, and aesthetic*. New York: Appleton.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response. *Journal of Comparative Physiology*, *43*, 482–489.

Sperry, R. W. (1958). Physiological plasticity and brain circuit theory. In H. F. Harlow, C. N. Woolsey (Eds.), *Biological and biochemical bases of behavior*. Madison, WI: University of Wisconsin Press.

Tsuda, I. (1991). Chaotic itinerancy as a dynamical basis of hermeneutics in brain and mind. *World Futures*, *32*, 167–184.

von der Malsburg, C. (1983). How are nervous structures organized? In E. Basar, H. Flohr, H. Haken, A. J. Mandell (Eds.), *Synergetics of the brain* (pp. 238–249). Berlin: Springer-Verlag.

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1195–1206, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00060-9

1997 SPECIAL ISSUE

A Neural Global Workspace Model for Conscious Attention

JAMES NEWMAN,[1] BERNARD J. BAARS,[2] AND SUNG-BAE CHO[3]

[1]Colorado Neurological Institute, Denver, [2]The Wright Institute, Berkeley and [3]Department of Computer Science, Yonsei University

(*Received* 9 *July* 1996; *accepted* 24 *April* 1997)

Abstract—*Considerable progress is being made in interdisciplinary efforts to develop a general theory of the neural correlates of consciousness. Developments of Baars' Global Workspace theory over the past decade are examples of this progress. Integrating experimental data and models from cognitive psychology, AI and neuroscience, we present a neurocognitive model in which consciousness is defined as a global integration and dissemination system — nested in a large-scale, distributed array of specialized bioprocessors — which controls the allocation of the processing resources of the central nervous system. It is posited that this global control is effected via cortical 'gating' of a strategic thalamic nucleus. The basic circuitry of this neural system is reasonably well understood, and can be modeled, to a first approximation, employing neural network principles. © 1997 Elsevier Science Ltd.*

Keywords—Attention, Binding, Consciousness, Global Workspace theory, Nucleus reticularis thalami, Programmable blackboard, Thalamocortical circuit, Thalamus.

1. INTRODUCTION

Consciousness has been widely portrayed as an intractable or irrelevant problem for cognitive science (e.g. Harnad, 1994; Penrose, 1994; O'Nuallain et al., 1997). Certainly its serious consideration is a fairly recent development (Jackendorf, 1987; Baars, 1988; Johnson-Laird, 1988; Edelman, 1989; Crick and Koch, 1990a), although a brief enthusiasm for the subject surfaced, and submerged, three decades ago (Eccles, 1966; Penfield, 1975). While it is not widely realized, the experimental neuroscience which served as the basis for that earlier enthusiasm is proving increasingly relevant to the present recrudescence of interest in conscious processes (see e.g. Stryker, 1989; Newman, 1995a, 1997). Beginning with historical developments in both AI and neuroscience, this paper reviews a growing body of evidence that some of the basic mechanisms underlying consciousness can be modeled, to a first approximation, employing variations upon current neural network architectures (see also Taylor, 1992; Taylor, 1996; Baars et al., in press; Newman et al., 1997).

Baars (1983, 1988, 1992, 1994) has developed a set of 'Global Workspace Models', based upon a unifying pattern, and addressing a substantial domain of evidence explicitly related to conscious experience. These models explicate an architecture in which many parallel, non-conscious experts interact via a serial, conscious and internally consistent Global Workspace (GW), or its functional equivalent. GW, or blackboard, architectures were first developed by cognitive scientists in the 1970s and this framework is closely related to the Unified Theories of Cognition of Simon, Newell and Anderson (see Newell, 1992).

The HEARSAY model of speech understanding (Reddy et al., 1973) was one of the earliest attempts to simulate a massively parallel/interactive computing architecture. The notion of a global workspace was initially inspired by this architecture, consisting of a large number of knowledge modules, or 'local experts', all connected to a single 'blackboard', or problem-solving space. Activated experts could compete to post 'messages' (or hypotheses) on the blackboard for all the other experts to read. Incompatible messages would tend to inhibit each other, while the output of cooperating experts would gain increasing access to the blackboard until a global solution emerged. Blackboard architectures are relatively slow, cumbersome and error-prone, but are capable of producing solutions to problems too novel or complex to be solved by any extant modular knowledge source. Once such 'global solutions' are attained, however, the original problems can be allocated to modular processors for 'non-conscious' solution.

Corresponding Author, Present address: 740 Clarkson Street, Denver, CO 80218, USA; Tel.: (303) 832-9062; e-mail: newmanjb@aol.com.

McClelland (1986) attested to the significance of this set of models to subsequent developments in cognitive science when he described HEARSAY, not only as ''a precursor of the interactive activation model'', but ''of the approach that underlies the whole field of parallel distributed processing'' (p. 122). We consider McClelland's own 'Programmable Blackboard Model of Reading' as a connectionist example of a global workspace architecture, and discuss its applicability to modeling conscious processes in a concluding section.

Another class of models that may turn out to be compatible with GW theory comes from 'distributed artificial intelligence' (DAI), which Durfee (1993) characterizes as the study of ''how intelligent agents coordinate their activities to collectively solve problems that are beyond their individual capabilities'' (p. 84). He cites examples of DAI applications, such as generic conflict resolution, unified negotiation protocols, and search-based models of coordination/cooperation. DAI applications appear to more closely approximate human interpersonal behaviour than purely logic-driven AI. They require that agents learn to be ''knowledgeable and skilled in interacting with others'' (p. 86). DAI models would appear to reflect an intelligent balance between competitive self-interest and cooperative problem-solving that is essential to optimizing overall outcomes in complex 'social' organizations. This, like GW theory, is consistent with other well-known metaphors in cognitive science, such as Minsky's 'Society Theory' (Minsky, 1979) and Gazzaniga's 'Social Brain' (Gazzanigga, 1985).

A similar, globally-integrative balancing of priorities appears to characterize the optimal processing of conscious information. Conscious percepts are characterized by unified gestalts of shape, texture, color, location and movement, despite the fact that these contributions to perception are initially processed in parallel areas of the cortex, in both hemispheres. Moreover, conscious intentions are generally single-minded and goal-directed. Of course, conflicts can and do arise, but a central purpose of consciousness seems to be resolving such conflicts (employing both integrative and inhibitory algorithms).

While such global states can be highly adaptive — indeed, are essential to explicit learning — GW theory maintains that the vast majority of cognitive tasks performed by the human brain are automatic, and largely non-conscious (Baars, 1988; Newman and Baars, 1993; Baars, 1997; Newman, 1997). Consciousness generally comes in play when stimuli are assessed to be novel, threatening, or momentarily relevant to active schemas or intentions.

> The defining properties of stimuli which engage conscious attention (i.e. the global allocation of processing resources) are that they: 1) vary in some significant degree from current expectations; or 2) are congruent with the current, predominant intent/goal of the organism. In contrast, the processing of stimuli which are predictable, routine or over-learned is automatically allocated to non-conscious, highly modularized cognitive systems (Newman, 1995b, p. 691).

Generally, we are conscious of what has the highest relevance to us at that moment. This may be a momentary threat, a sudden insight, a pleasant sensation, etc. (in relaxed moments, there may be no particular focus or intent, simply a stream of associations). Yet, while the range of our awareness is immense (limited only by our most developed cognitive capacities), we contend that the basic mechanism for the allocation of these capacities remains constant under virtually all contingencies; and the basic neural circuitry of that resource-allocation mechanism is reasonably well understood. Indeed, in subsequent sections, we suggest how it might be modeled based upon already existing neural network simulations (McClelland, 1985; Hampshire and Waibel, 1992; Taylor and Alavi, 1993; Llinas et al., 1994).

The relevance of Global Workspace theory extends beyond NN modeling, however. Indeed, it bears upon central philosophical problems in consciousness studies, such as the *homunculus* and Cartesian theater. The two are, of course, related. The image of a 'little man in our head' observing and manipulating the play of conscious images is beguiling, but absurd. For who is this strange being lodged in our brains? And who is watching **him**?

In Global Workspace theory the single homunculus is replaced by a large 'audience of experts'. The 'theater of consciousness' then becomes a workspace, with stage (Baars, 1997). Almost everyone in an audience has potential access to center stage (although most prefer to simply observe, or exert indirect influences). The focus of conscious activity, at any moment, corresponds to the 'work' produced by the most active coalition of experts, or modular processors: whoever has managed to win the competition for 'the spotlight'. There is no fixed, superordinate observer. Individual modules can pay as much or as little attention as suits them, based upon their particular expertise. At any one moment, some may be dozing in their seats, others busy on stage. In this sense, the global workspace resembles more a deliberative body than a theater audience. Each expert has a certain degree of 'influence', and by forming coalitions with other experts can contribute to deciding which issues receive immediate attention and which are 'sent back to committee'. Most of the work of this deliberative body is done 'off stage' (i.e. non-consciously). Only matters of greatest relevance in-the-moment gain access to consciousness.

While the GW is a teaming multiplicity, what is explicitly represented in consciousness is largely coherent and adaptive. The overall workspace serves as a 'global integration and dissemination system', in which all experts can participate, but only select coalitions dominate, momentarily, producing an orderly succession

of global representations. The stream of consciousness arises out of the operations of the GW system — and, over time, our sense of being a coherent 'I' (the memory and intention systems vital to this aspect are beyond the scope of this paper; see Baars et al., in press; Newman, 1997). It is this unitary awareness, not any agent or homunculus, that is globally superordinate. Of course, such a system is prone to inefficiencies and pathological perturbations, but this is consistent with the scientific literature concerning human consciousness (see Baars, 1988).

If we are to proceed beyond pleasing metaphors, however, it is necessary to operationalize the GW model in explicit neurocognitive terms. This process begins in the next section. To introduce it, we offer the following working definition:

> consciousness reflects the operations of a global integration and dissemination system, nested in a large-scale, distributed array of specialized bioprocessors; among the various functions of this system are the allocation of processing resources based, first, upon biological contingencies of novelty, need or potential threat and, secondly, cognitive schemas, purposes and plans.

2. MODELING GLOBAL, COMPETITIVE ATTENTION

We have introduced the theoretical background for the model. Newman and Baars (1993) and Newman (1997) present detailed accounts of its neural architecture. We would stress, however, that consciousness is a dynamic process, not a static structure. Also, it is not localized to some 'brain center', but arises out of the coordinated activities of widely distributed networks of neurons. Resource allocation is integral to these activities. The neural bases of resource allocation, or attention, have been extensively explored (see, e.g. Heilman et al., 1985; Mesulam, 1985; Posner and Rothbart, 1991; Posner, 1994; LaBerge, 1990, 1995). But, of course, not all forms of attention are conscious. As an example from AI, McClelland (1986) notes that in simulations of reading, activated modules must be 'sticky', that is ''interactive activation processes continue in older parts of the programmable blackboard while they are being set up in newer parts as the eye moves along...'' (pp. 150–151). This 'stickiness' would seem to entail a type of attention. It normally proceeds quite automatically, however, both in a reading machine and in a literate human being. Only when the process is disrupted by, say, a mis-spelled or unknown word, does that word becomes the focus of our conscious awareness. Normally, we are only conscious of the overall sense of the passage of text, and the images and thoughts it evokes, not particular semantic or syntactical operations. These linguistic processes became second nature to us long ago. Such 'particular operations' are hardly trivial aspects of language acquisition, but as Kihlstrom (1987) noted, in humans they tend to be ''automatized through experience and thus rendered unconscious'' (p. 285).

Conscious awareness clearly involves a higher order of resource allocation, which Newman and Baars (1993) call 'global attention'. The term ''refers to a level of cognitive processing at which a single, coherent stream of information emerges out of the diverse activities of the CNS'' (p. 258). The focus of that stream could (under atypical circumstances) be an individual word; but the conscious mind seldom confines itself to the processing of such rudimentary representations. Rather it seems to be decisively biased towards multifaceted, yet unified images. Thus, we are able to perceive a Necker Cube as projecting out of a two-dimensional page, alternately to the left, then to the right; but we are curiously incapable of perceiving these two perspectives simultaneously.

The processing load of global attention (like working memory), is both highly chunked and highly restricted (Baars, 1988). The non-conscious allocation of processing resources operates under no such constraints. For example, neuroscience has shown that specialized areas in the visual cortex process, in parallel, the contour, movement, color, spatial location, etc. of a stimulus (LaBerge, 1995). Yet our awareness is of a single, coherent object (and often includes tactile, auditory and associative aspects). Thus, neuroscience is faced with the 'binding problem' of how these multifarious representations, generated by widely separated areas, are integrated into real-time 'objects' of perception (see Crick and Koch, 1990a; Newman and Baars, 1993).

One would expect the neural mechanism for global attention to be complex, and widely distributed, which it is. But the basic circuitry can be described, to a first approximation, in terms of repeating, parallel loops of thalamo–cortico–thalamic axons, passing through a thin sheet of neurons known as the *nucleus reticularis thalami* (nRt). The loops are formed by long-axoned, excitatory neurons. The neurons of nRt are largely GABA-ergic, inhibitory neurons. Most, if not all, of the looping axons give off collaterals as they pass through nRt, while nRt neurons themselves project mainly to cells of the particular thalamic nucleus lying directly beneath them. There is an orderly topography to this array of axon collaterals and underlying thalamic nuclei (Scheibel and Scheibel, 1966; Mitrofanis and Guillery, 1993). It essentially mirrors, in miniature, the modular architecture of the cortex (see Newman and Baars, 1993; LaBerge, 1995; Newman, 1997, for reviews).

Evidence for the central role of this 'thalamocortical circuit' (LaBerge, 1995) in attention and consciousness has been accumulating for decades (Jasper, 1960; Scheibel, 1980; Jones, 1985; Steriade and Llinas, 1988; Llinas and Pare, 1991). Skinner and Yingling (1977) first proposed a neural model for its role in selective attention. Our 'wagon wheel' model (next section) represents a

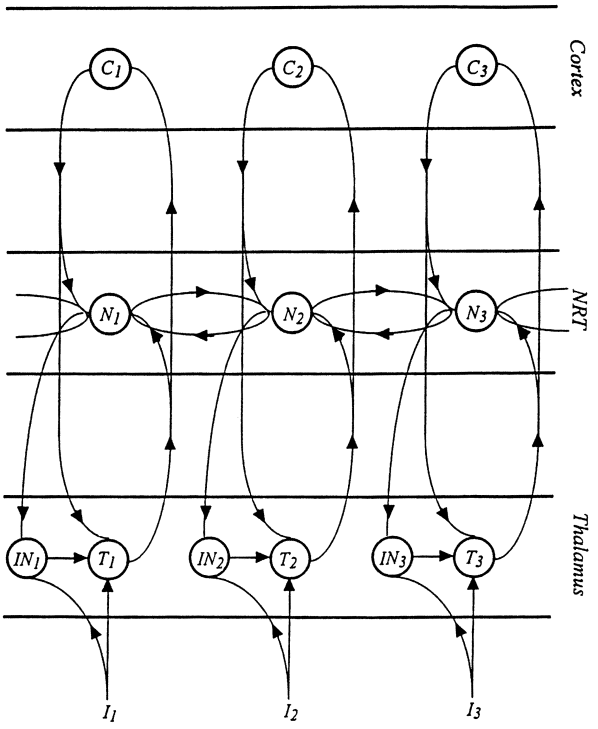


FIGURE 1. The wiring diagram of the main model of the thalamus–NRT–cortex complex. Input I_j is sent both to the thalamic relay cell T_j and the inhibitory interneuron IN_j, which latter cell also feeds to T_j. Output from T_j goes up to the corresponding cortical cell C_j, which returns its output to T_j. Both the axons T_jC_j and C_jT_j send axon collaterals to the corresponding NRT cell N_j. There is axonal output from N_j to IN_j, as well as collaterals to neighbouring NRT cells. There are also dendro–dendritic synapses between the NRT cells (from Taylor and Alavi, 1993).

synthesis of both the accumulated evidence, and related models (Skinner and Yingling, 1977; Scheibel, 1980; Crick, 1984; Taylor and Alavi, 1993; Llinas et al., 1994; LaBerge, 1995). These related models vary in their details, as the precise connectivities and physiology of the thalamocortical circuit are not fully worked out.

Most attentional models are based upon conventional simulations of mechanisms such as center-surround inhibition, or winner-take-all (WTA) competitions, among local circuits. Various researchers have described the network of nRt neurons as a mosaic, or array, of neural 'gatelets' acting to selectively filter the flow of sensory inputs to the cortex (Skinner and Yingling, 1977; Scheibel, 1980; Crick, 1984). The WTA dynamic may seem analogous to the 'competition' posited by GW theory. The problem with such conventional networks is that they are poorly suited to global forms of competition, because prohibitively long-range and geometrically increasing numbers of connections would be required. Moreover, most long-range, reciprocal connections in the CNS are excitatory. Inhibitory effects tend to be local.

Taylor and Alavi (1993), however, have modeled a competitive network for global attention based upon a highly simplified version of the 'thalamus–NRT–cortex complex'. Their model is unique, in that it takes into account the effects of dendro–dendritic interactions throughout nRt. The dendrites of nRt cells project out tangentially within the reticular sheet, bidirectionally. The physiology of information processing in dendritic trees is highly complex, and not well understood (Mel, 1994); but Koch and Poggio (1992) review evidence for the dendritic trees playing a role in several types of second-order, multiplicative computations. We will have more to say about this subsequently.

Figure 1 [taken from Taylor and Alavi (1993)] illustrates three thalamocortical circuits, as well as the nonlinear, dendro–dendritic connections between N1, N2, N3,... within NRT. We would refer the reader to the original paper for a detailed description of the simulations carried out, employing a network of 100 thalamocortical loops. To briefly summarize the results, the addition of dendro–dendritic connections to the looping circuits provided ''the basis for a simple version of the global gating model... that instantiates a form of competition in the spatial wavelength parameters of incoming inputs...'' (p. 352). In this version of the model, the entire nRt network oscillates with a wavelength,

> with the net strength given by the component of the input with the same wavelength.

The way in which global control arises now becomes clear. Only those inputs which have special spatial wavelength oscillations are allowed through to the cortex, or are allowed to persist in those regions of the cortex strongly connected to the NRT: the thalamus–NRT system acts as a spatial Fourier filter (p. 353).

Simulation runs demonstrated the global, wave-like properties of the competitive model. The overall pattern of activation in cortical units was shown to be exclusively dependent upon the wave pattern spanning across all of the NRT units (Figure 2). As LaBerge (1995) notes, the actual physiology of nRt gating in alert states remains unclear, but it is firmly established that nRt is the source of global oscillatory activity (at 8–13 Hz) initiating the descent into sleep.

> The RN cells are known to inhibit each other, and when inhibition hyperpolarizes an RN cell sufficiently, it produces a rebound burst. In this way a network of connected RN inhibitory cells can spread activity to every cell within the network, apparently without decrement in the intensity of the activity (p. 184).

Here then, is a plausible circuitry for a global, winner-take-all competition among the large array of specialized cortical processing areas.

Llinas et al. (1994) offer an interesting variation upon this circuitry, in which thalamocortical loops of the 'non-specific' intralaminar nuclei operate in parallel with the specific (input) loops described above. The synchronous

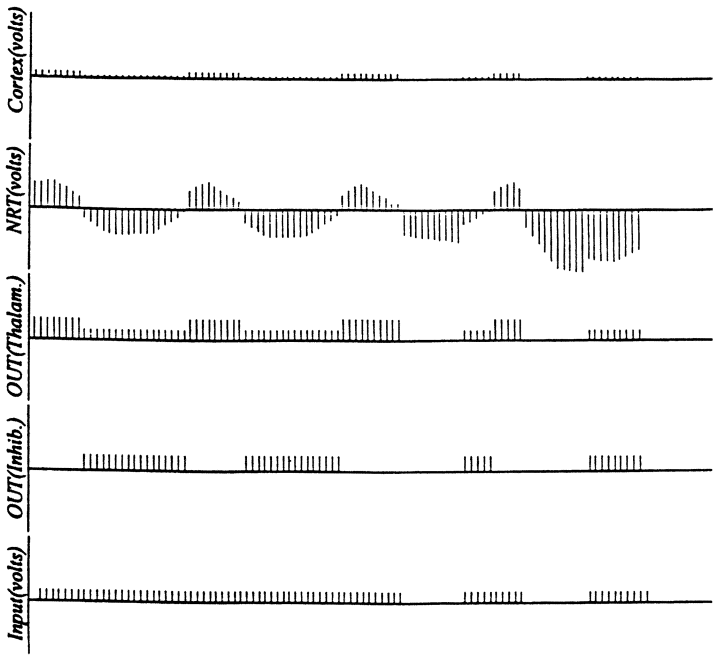


FIGURE 2. One of 15 simulation runs for the thalamus–NRT–cortex model showing full global control with semi-constant spatial input. Note that cortex activity is influenced by the NRT alone (from Taylor and Alavi, 1993).

activation of specific and non-specific loops is postulated to provide a basis for ''perceptual unity... by which different sensory components are gathered into one global image'' (p. 251). Their modeling is concerned with high-frequency EEG oscillations (and omits dendro–dendritic connections), yet appears to parallel much of what we discuss above.

> When the interconnectivity of these nuclei is combined with the intrinsic properties of the individual neurons, a network for resonant neuronal oscillations emerges in which specific corticothalamic circuits would tend to resonate at 40 Hz. According to this hypothesis, neurons at the different levels, and particularly those in the reticular nucleus, would be responsible for the synchronization of 40-Hz oscillations in distant thalamic and cortical sites...these oscillations may be organized globally over the CNS, especially as it has been shown that neighboring reticular cells are linked by dendrito–dendritic and intranuclear axon collaterals (Deschenes et al., 1985; Yen et al., 1985, pp. 253–254).

3. A NEURAL MODEL FOR GLOBAL RESOURCE ALLOCATION

We have introduced a set of convergent models for the basic circuitry of a Global Workspace system involved in the integration and dissemination of the processing resources of the nervous system. This 'bare bones' version accounts for how a global, winner-take-all competition might be mediated between various external inputs and cortical modules, to produce ''a single, coherent stream of information out of the diverse activities of the CNS'' (Newman and Baars, 1993). There remains to be explained how the thalamocortical circuit fits in with the second half of our working definition for the conscious system: the allocation of processing resources

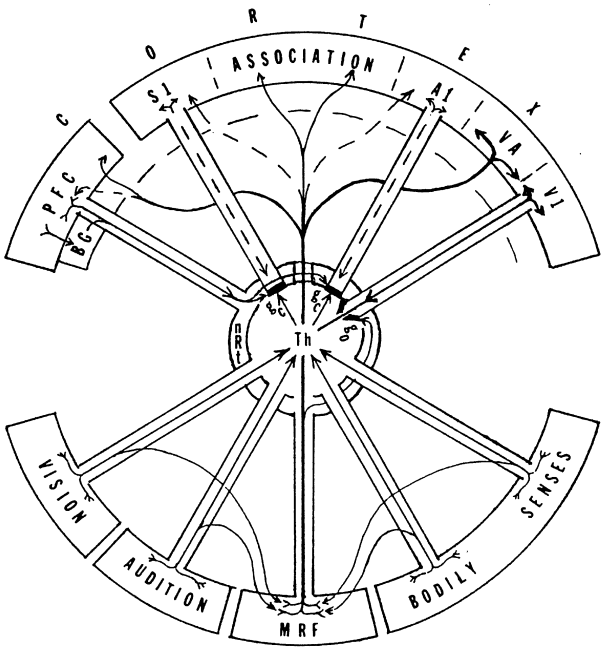


FIGURE 3. 'Wagon wheel' model of CNS systems contributing to global attention and conscious perception. A1, primary auditory area; BG, basal ganglia; g_c, 'closed' nRt gate; g_o, 'open' nRt gate; MRF, midbrain reticular formation; nRt, nucleus reticularis thalami; PFC, prefrontal cortex; S1, primary somatosensory area; Th, ventral thalamus; V1, primary visual cortex (from Newman et al., 1997).

based, first, upon biological contingencies of novelty, need or potential threat and, secondly, cognitive schemas, purposes and plans. In keeping with our definition, we will first add a sub-cortical component for orienting to ''novelty, need, or potential threat'', and then discuss the much more complex aspects of cortically-mediated effects upon the system.

This extended version of the model is schematically illustrated in Figure 3 as a 'wagon wheel', with the thalamus (Th) as its 'hub'. The reticular nucleus (nRt) corresponds to the metal sleeve fitted around the hub. The upper rim of the wheel represents the cerebral cortex (PFC/S1...V1), and closely associated basal ganglia (BG). The lower half shows the major sensory systems and subcortical nuclei whose projections converge upon the thalamus. The outer 'spokes' represent the sensory pathways for vision, audition and the bodily senses. These project, in an orderly topography, to modality-specific nuclei in the thalamic 'hub'. As they ascend towards the thalamus, these pathways give off collaterals to the midbrain reticular formation (MRF) (see also Newman et al., 1997). Scheibel (1980) reviewed three decades of experimental evidence indicating that these midbrain collaterals serve as the basis for an initial 'spatial envelope', or global map, of the environment surrounding the animal.

> Most reticular [MRF] neurons...appear multimodal, responding to particular visual, somatic and auditory stimuli, with combinations of the last two stimuli most numerous. The common receptive fields of typical bimodal cells in this array show a significant degree of congruence. For instance a unit responding to stimulation of the hind limb will usually prove maximally sensitive to auditory stimuli originating to the rear of the organism. These twin somatic and auditory maps retain approximate register and overlap the visuotopic map laid down in the...superior colliculus...These data might be interpreted to mean that each locus maps a point in the three-dimensional spatial envelope surrounding the organism. Further studies suggest the presence of a deep motor map closely matching and in apparent register with the sensory map. (p. 63)

More recent research has supported Scheibel's portrayal of the superior colliculus as the visual component of what Crick and Koch (1990b) termed a 'saliency map' for eye movements, involved in orienting the animal to biologically relevant stimuli. Subsequent findings have both confirmed Scheibel's analysis, and revealed a number of 'top-down' projections that modulate activities in MRF. LaBerge (1995) writes:

> the superficial area [of the superior colliculus] receives strong cortical inputs from V1, V2 and V3 [primary and secondary visual cortex], the deep layers in the monkey SC receive their main cortical inputs from the posterior parietal area (Lynch *et al*., 1985), from the prefrontal areas (Goldman and Nauta, 1976), and the frontal eye fields (Leichnetz *et al*., 1981). The deep layers contain a map of visual space that is stacked adjacent to maps for auditory and somatosensory spaces in a manner that cells corresponding to points in space lie along the same vertical axis (Merideth and Stein, 1990). Stimulation of these cells by microelectrodes produces movements of eyes and head... (LaBerge, 1995, p. 145)

LaBerge goes on to describe basal ganglia inputs that ''are of particular importance because they tonically inhibit activity in the SC cells''. It has long been known that the frontal eye fields, and posterior parietal area ''exert strong influences on eye movements and must be considered together with the superior colliculus in accounting for...orienting of attention'' (p. 142). These facts emphasize two key aspects of the 'conscious system' we are modeling: 1) it is polymodal, integrating not just visual, auditory and somatosensory inputs, but motor and 'higher-order' cortical effects; and 2) it is extended, with input/output relations reaching from the brain stem core to association cortices. Indeed, the general term we have used to describe it elsewhere is the 'extended reticular–thalamic activation system', or 'ERTAS' (Baars, 1988; Newman and Baars, 1993; Newman, 1995a, 1995b, 1997).

The third key aspect of the system (as exemplified by the 'wagon wheel' model) is that it converges on the thalamus. We have already discussed this in terms of the thalamocortical circuit, which connects to ''virtually every area of the cerebral cortex'' (LaBerge, 1995, p. 221). Scheibel (1980) described the MRF portion of the system as:

> sweep[ing] forward on a broad front, investing the [intralaminar complex of the] thalamus and nucleus reticularis thalami. The investiture is precise in the sense that the sites representing specific zones of the spatial envelope (receptive field) project to portions of the nucleus reticularis concerned with similar peripheral fields via projections from both sensory thalamus and sensory association cortices (p. 62).

The fact that Scheibel's (1980) 'spatial envelope' projects with some topographic precision upon nRt, would appear to enable it to disinhibit particular arrays of nRt gatelets, selectively enhancing the flow of sensory information to the cortex. The 'intralaminar complex' (Newman and Baars, 1993) is also integral to the ERTAS system, as the non-specific portion of the thalamocortical circuit. It is intralaminar projections which relay MRF activation to the cortex (illustrated by the vertical MRF-Th projection, above which it branches out to all areas of CORTEX). As noted above, Llinas et al. (1994) hypothesize the perceptual unity of consciousness (binding) to be brought about by the global synchronization of specific and non-specific circuits via nRt. Scheibel (1980) earlier concluded as much concerning the role of this extended activation system in 'selective awareness':

> From these data, the concept emerges of a reticularis

> complex [nRt] selectively gating interaction between specific thalamic nuclei and the cerebral cortex under the opposed but complementary control of the brain stem reticular core [MRF] and the frontal granular cortex [PFC]. In addition, the gate is highly selective; thus, depending on the nature of the alerting stimulus or central excitation, only that portion of the nucleus reticularis will open which controls the appropriate subjacent thalamic sensory field. The reticularis gate [thus] becomes a mosaic of gatelets, each tied to some specific receptive zone or species of input. Each is under the delicate yet opposed control of: (a) the specifically signatured sensory input and its integrated feedback from cortex [S1...V1]; (b) the reticular core [MRF] with its concern more for novelty (danger?) than for specific details of experience; and (c) the frontal granular cortex-medial thalamic system [PFC/BG] more attuned to upper level strategies of the organism, whether based on drive mechanisms (food, sex) or on still more complex derivative phenomenon (curiosity, altruism). Perhaps here resides the structuro- functional substrate for selective awareness, and in the delicacy and complexity of its connections, our source of knowing, and of knowing that we know (p. 63).

Here, as well, is a summary description of a neural substrate for the global allocation of the processing resources of the CNS. All that it lacks is the mechanisms for a global competition (Taylor and Alavi, 1993) and binding (Llinas et al., 1994) introduced in the previous section. But we must tie the operations of this thalamus-centered system more closely to those of the cortex and basal ganglia, or most of the functions routinely studied by cognitive science have no place in the model. This introduces an exponentially higher level of complexity (one of the hazards of dealing with global systems).

One of the values of GW theory, however, is that it provides a framework for understanding this complexity. First, it holds that the vast majority of cognitive functions are carried out, non-consciously, via changing arrays of specialized, modular processors. This is reflected, anatomically, in the immense number of cortico–cortical connections in the human brain, outnumbering those with subcortical nuclei by nearly ten to one. Thalamocortical projections are comparatively sparse, but serve at least two essential functions: 1) transmitting sensory inputs to the primary cortical areas (S1, A1, V1, Figure 3); and 2) providing a means to selectively amplify/synchronize cortex-wide activation (Section 2).

GW theory also reminds us that conscious functions operate upon an information load about the size of working memory. Thus, we are talking of a highly coarse-grained level of processing. In this context, global attention is (at least) a second-order operation, acting upon a highly selective stream of information. All this is to say that a relatively low density of widely distributed, yet highly convergent, circuits could be all that are required to create a conscious system; and these are the very characteristics of the neural model we have described.

However, most neural network modelers take a cortically-centered view of cognition, from which the brain stem functions so far described probably seem rather primitive or trivial (i.e. orienting, controlling eye movements) when compared to cortically-mediated processes such as language acquisition, pattern recognition, motor planning, etc. What evidence is there that cortical (and other forebrain systems) depend upon projections to the thalamus for effecting high-level cognitive processes?

Early support for such effects, mediated by prefrontal projections, was provided by animal experiments undertaken by Skinner and Yingling (1977). They found that selective activation of one portion of a fronto-thalamic tract could shut down sensory processing in visual, but not auditory, cortex. Activation of another 'spoke' of the prefrontal-thalamic tract shut down auditory processing, but allowed visual inputs to reach posterior cortex. Skinner and Yingling wrote ''This result implies that selective attention emerges via selective inhibition in certain sensory channels that the animal must **know in advance** are irrelevant to its situation'' (p. 54). To inhibit orienting based upon advanced knowledge is clearly a sophisticated use of cognition. Several lines of research have converged in recent years to support this concept. Summarizing the current state of knowledge of prefrontal regulation of subcortical systems, Newman (1997) wrote:

> It is now generally accepted that the prefrontal lobes (with the cingulate cortex) constitute an 'executive' over the limbic system mediating such functions as working memory, inhibition of conditioned responses, and goal-directed attention (see Fuster, 1980; Goldman-Rakic, 1988b; Damasio, 1994; Posner, 1994). More recent research on the basal ganglia (see reviews by Groenewegen and Berendse, 1994; Parent and Hazrati, 1995) have suggested that they constitute a 'motor programming extension' of the frontal lobes as well — routed through the thalamus (p. 112–113).

Newman (1997) goes on to cite evidence (Parent and Hazrati, 1995) that the BG 'extension' (like the thalamocortical loops) sends rich, collateral projections to nRt that effect not only its 'gating' of motor programs, but hippocampal-mediated episodic memory functions (see also Newman, 1995b).

Finally, we would note that cortico-thalamic projections to nRt and associated specific nuclei are both more topographically precise (Mitrofanis and Guillery, 1993) and more pervasive than had once been thought (Jones, 1985). Llinas and Pare (1991) estimate that, for every axon the thalamus sends to the cortex, the cortical area it projects to reciprocates with ten. Given the modular architecture of the neocortex, one might reasonably predict that these cortico-thalamic projections exert highly differentiated influences upon the flow of information through the thalamus. Efforts by experimental neuroscience throughout the 1980s to elucidate the precise effects of cortico-thalamic projections were

frustratingly inconclusive. But a recent review by Buser and Rougeul-Buser (1995) notes:

> The situation has however recently swung back, due to some new and perhaps consistent findings, indicating that the visual cortex appears to have a major action down onto the lateral geniculate nucleus, which may generate thalamic oscillations (Funke and Eyse, 1992; McCormick and Krosigk, 1992; Krosigk et al., 1993; Sillito et al., 1994) (p. 252).

While additional research is clearly needed, these recent findings suggest that Scheibel's (1980) early model of the converging influences of projections upon a thalamic hub — with the addition of basal ganglia inputs to nRt and the intralaminar complex — remains a viable model for ''global attention'', including the influences of cortically generated ''schemas, purposes and plans''. Newman (1997) discusses the contributions of the 'cortico-basal ganglia–thalamo–cortical loop' to memory and volitional processes in greater detail. The complexities of this system are beyond the scope of the models presented here, although Monchi and Taylor (1995) and Taylor and Michalis (1995), among others, have developed neural models simulating functions of the BG and hippocampal systems.

What we propose to do instead is present a much simpler, but highly relevant, connectionist model that simulates the sorts of second-order operations one would predict in a GW system employing a gating network to selectively filter and integrate inputs as a function of central knowledge stores. The basic heuristic for this type of model is described in Newman and Baars (1993). It posits that

> prefrontal cortex acts as an executive attentional system by actively influencing information processing in the posterior cortex through its effects upon the nucleus reticularis. In this manner, the highly parallel [processing] functions of the posterior cortex are brought into accord with increasingly complex and intentional cognitive schemes generated within the prefrontal regions of the brain (p. 281).

A defining property of an executive system is that it acts upon other sub-systems, modifying their inputs for its particular purposes. Posterior cortical areas act more like arrays of quasi-autonomous processing modules (or local experts) — the bread and butter of NN simulations. Note that an executive system is not an *essential* requirement for consciousness. That this is the case is illustrated by the literature on extensive damage to the frontal lobes of the brain. PFC damage results in significant deficits in such purposeful activities as: the inhibition of inappropriate responding; switching of response set, planning and monitoring of actions, etc.; but produces little or no alteration in basic mental status. Indeed, many patients with frontal lobe pathology perform at pre-morbid levels on intelligence tests (Walshe, 1978; Damasio, 1994). In terms of the GW model we have presented, it is not executive attentional processes, but the selective binding of coalitions of active cortical modules via a thalamocortical competition which is the *sine qua non* for the generation of a coherent stream of conscious representations. Examples of these aspects of the GW model have already been offered.

4. SECOND-ORDER MODELS FOR GLOBAL GATING

Let us return to the 'wagon wheel' model illustrated in Figure 3, and transform its components into a connectionist GW, with an executive system. To simplify things, the network will have only two sensory modules, one for processing auditory (A1) inputs, and one for visual (V1). In order to provide second-order control over processing in both modules, we will add a gating module (nRt) with the same number of units as connections in each sensory module. Each gating unit sends its output to a corresponding connection in A1 and V1. The connections between the gating units and sensory units are multiplicative. As Rummelhart et al. (1986) write about such connections:

> if one unit of a multiplicative pair is zero, the other member of the pair can have no effect, no matter how strong its output. On the other hand, if one unit of a pair has value 1, the output of the other passe[s] unchanged to the receiving unit....In addition to their use as gates [such] units can be used to convert the output level of a unit into a signal that acts like a **weight** connecting two units (p. 73).

In this manner, a fully connected gating module can actually program the connection strengths of one or more input modules to process a particular type of input, for example phonemes, or letters, into words. For maximum flexibility, it is preferable that the gating module not have fixed connections either, but simply relay gating (connection strength) information from a central module to which its units are connected. The central module contains (in this case) word-level knowledge needed to program the sensory modules to process words. Another central module might be specialized for knowledge for processing visual scenes or tactile shapes. To complete the system, each programmable input unit sends a corresponding connection to a central module unit.

The highly simplified network just described is really a variation on a 'Programmable Blackboard Model for Reading' developed by McClelland (1985, 1986). Its four modules correspond to those labeled in Figure 4: a Central Module (PFC); Connection Activation System (PFC-nRt); and two Programmable Modules (A1, V1). The connections described above are shown in Figure 5 (note: McClelland's modules are identical, and used only for reading (not hearing) words, but theoretically they could be programmed to process *any* type of input).

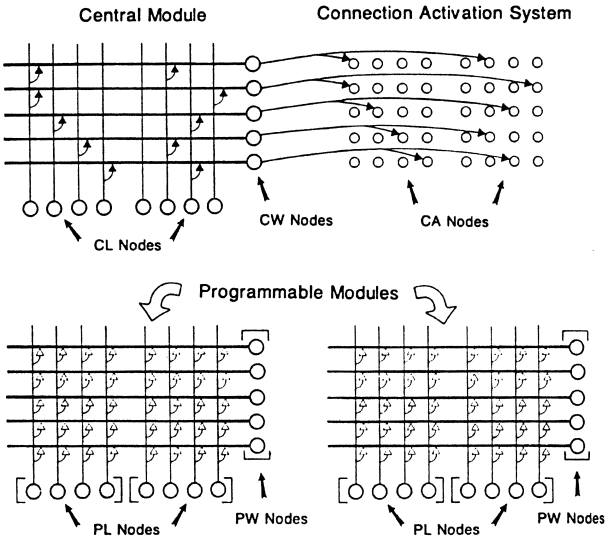


FIGURE 4. A simplified example of a Connection Information Distributor (CID), sufficient for simultaneous bottom-up processing of two two-letter words. The programmable modules consist of the programmable letter (PL) and programmable word (PW) nodes, and programmable connections between them (open triangles). The central module consists of a set of central letter (CL) nodes and a set of central word (CW) nodes, and hard-wired connections between them (filled triangles). The connection activation system includes the central word nodes, a set of connection activator (CA) nodes, and hard-wired connections between them. Connections between the central knowledge system (central module plus connection activation system) and the programmable blackboard are shown in Figure 6 (from McClelland, 1985).

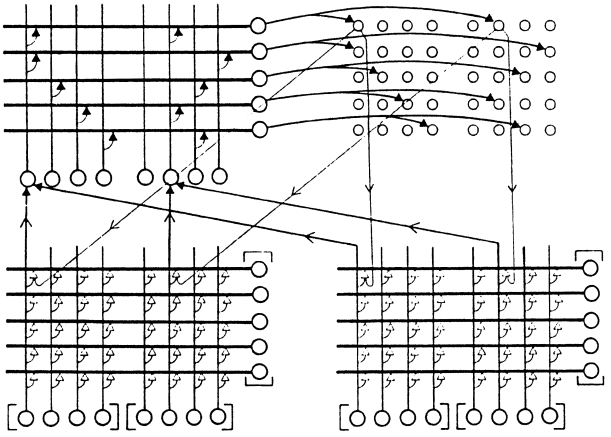


FIGURE 5. Each CA node projects to the corresponding connection in both programmable modules, and each central letter node receives projections from the corresponding programmable letter node in both programmable modules. The inputs to two central letter nodes, and the outputs from two CA nodes are shown (from McClelland, 1985).

In the brain, of course, the primary areas (A1, V1, S1) send no direct projections to PFC; but they do send convergent projections (as in Figure 5) to secondary association areas, which send projections directly to PFC (as well as posterior association areas). Although these feed-forward projections to PFC are less topographically precise (e.g. the receptive fields of visual neurons in the secondary areas are much larger), they maintain a fair degree of parallel distribution, indicating that much of the prefrontal cortex is as modular in its organization as the posterior 'association' cortex. Moreover, PFC 'modules' reciprocate these parallel, feed-forward projections, although in a more divergent pattern (Goldman-Rakic, 1988a; LaBerge, 1995). Interestingly, this convergence/divergence pattern is paralleled by the connections in Figure 5 for the central module.

In the actual prefrontal cortex there are hundreds (if not thousands) of 'central modules'. Feed-forward inputs allow them to use and store highly processed information from the posterior (sensory) cortex. Of course, feedback (or re-entrant) connections enable PFC to influence processing in the posterior areas as well. But such divergent and indirect feedback pathways are poorly suited to exercising momentary, direct effects upon processing at the input level. Nor could such centrally-stored knowledge be employed to guide, or anticipate, how inputs are processed (re the 'knowing in advance' Skinner and Yingling (1977) attributed to PFC-Th circuits). This is where direct projections to the primary processing areas (actually the thalamocortical circuit) could prove quite valuable. Instead of the sensory input units (A1-Th; V1-Th) responding based upon fixed connection strengths, a central module could program input modules to process (i.e. pay attention to) particular categories of inputs. McClelland (1986) calls this form of activation 'connection information distribution' (CID) and compares its benefits to those of:

> the invention of the stored program....The use of centrally stored connection information to program local processing structures is analogous. It allows the very same processing structures to be programmed to perform a very wide range of different tasks.... [CID] also carries out a form of what is known in production systems as 'resolution', binding the right tokens in the blackboard together into higher-order structural patterns (p. 165).

Finally, he notes analogous aspects in the CID's operations to 'working memory', a process which has been tied by neuroscientists to a prefrontal/thalamic/hippocampal system (e.g. Fuster, 1980; Goldman-Rakic, 1988b). These comparisons between the Wagon Wheel and Programmable Blackboard models, of course, have purely heuristic value (although McClelland's (1986) PABLO simulation of his model contained a sufficient programmable blackboard to read lines of text up to 20 characters long). But the use of gating networks to generate useful 'higher-order structural patterns' is fairly widespread.

For engineering problems such as object recognition and robot motion control, the concept of combining modular networks using gating connections has been actively exploited to develop highly reliable systems (Jacobs et al., 1991; Hampshire and Waibel, 1992;

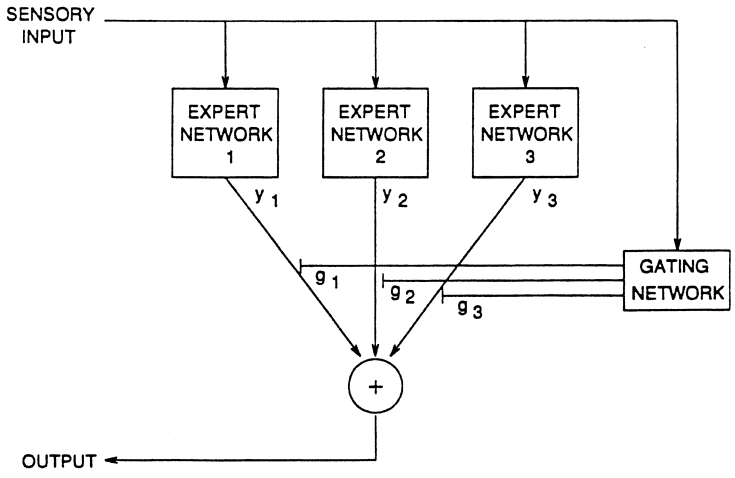


FIGURE 6. Schematic diagram of modular neural networks with three expert networks and a gating network. The output of the entire architecture, denoted by Y, is $Y = g_1y_1 + g_2y_2 + g_3y_3$, where y_1 denotes the output of the ith expert network.

Jacobs and Jordan, 1993; Cho and Kim, 1995). The key issue in this approach is how to combine the results of the individual networks to give the best estimate of the optimal overall result. Architectures used in this approach consist of two types of networks: an expert and a gating network. Basically, the expert networks compete to learn the training instances, and the gating network facilitates cooperation by the overall mediation of this competition. The expert networks may be trained separately using their own preassigned sub-tasks and differing modalities (e.g. vision and touch), or the same modality at different times (e.g. the consecutive 2-D views of a rotating 3-D object). The gating network need only have as many output units as there are expert networks.

To train such a gating network, Hampshire and Waibel (1992) developed a new form of multiplicative connection, which they call the 'Meta-Pi' connection. Its function is closely aligned with predecessors described in McClelland (1986). The final output of the overall system is a linear combination of the outputs of the expert networks, with the gating network determining the proportion of each local output in the linear combination. Figure 6 illustrates this architecture with three expert networks.

The final output of the overall system is a linear combination of the outputs of the expert networks, with the gating network determining the proportion of each local output in the linear combination. The Meta-Pi gating network allocates appropriate combinations of the expert networks when stimuli are assessed to be novel, while an automatic ('non-conscious') decision process operates in instances where a single expert can execute the task. This coupling of modular, expert networks and gating controls produces new levels of cooperative behavior. The expert networks are local in the sense that the weights in one network are decoupled from the weights in other expert networks. However, there is still some indirect coupling because if some other network changes its weights, it may cause the gating network to alter the responsibilities that get assigned to the expert network.

These examples from engineering applications of multiplicative, gating networks are not based upon the Wagon Wheel model or, for that matter, any specific neural circuitry. Yet Koch (1997) notes that

> Multiplication is one of the most common operations carried out in the nervous system (for example, for estimating motion or the time-to-contact with an approaching stimulus) (p. 208).

We are not aware of any studies of either the axon collateral or dendro–dendritic projections in nRt demonstrating multiplicative properties, but Mel (1994) has modeled such connections in the NMDA-rich dendritic trees of cortical pyramidal cells. He postulates that they perform nonlinear pattern discrimination and correlative operations. Given the role of the bidirectional dendritic trees of nRt cells in globally synchronizing the thalamocortical circuit (Taylor and Alavi, 1993; LaBerge, 1995), it seems likely that they will eventually be found to have important computational functions as well.

Even if it transpires that synchronous oscillations, not multiplicative connections, are the basis for the 'gating' functions of nRt upon the thalamocortical circuit, NN models based upon Meta-Pi connections may still be useful for simulating global workspace systems. The use of Meta-Pi connections has already been extended to synchronous oscillators in modular cortical neural networks. Indeed, computational simulations of phase-locked oscillations characteristic of neurons involved in the 'binding' of visual (Grossberg and Somers, 1991; Sompolinsky et al., 1991) and auditory (Vibert et al., 1994) features of an attended object have already been extended to synchronous oscillators using Meta-Pi connections. Such oscillatory circuits have also been

employed in modeling storage and retrieval in pattern recognition tasks (Yao and Freeman, 1990).

In this paper, we have introduced a collection of neuroscience and NN models for attention and binding, resource allocation, and second-order gating, which share important features and parallels with a Neural Global Workspace System for conscious attention (Newman and Baars, 1993). While the NN models we have presented only implement partial aspects of the GW system, and even our Wagon Wheel model largely neglects the influences of memory and affective systems upon the stream of consciousness, the outlines of a general framework for understanding conscious processes should be discernable (see Newman, 1997 for a fuller account). This is certainly great progress, given the virtual *terra incognita* consciousness has been for most of the history of science.

REFERENCES

Baars, B. J. (1983). How does a serial, integrated and very limited stream of consciousness emerge out of a nervous system that is mostly unconscious, distributed, and of enormous capacity? In G. R. Brock & J. Marsh (Eds), *CIBA Symposium on Experimental and Theoretical Studies of Consciousness* (pp. 282–290). London: John Wiley and Sons.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. (1992). *Experimental slips and human error: Exploring the architecture of volition*. New York: Plemun Press.

Baars, B. J. (1994). A global workspace theory of conscious experience. Baars, B. J. and Newman, J. (1994). A neuro-biological interpretation of a Global Workspace theory of consciousness. In A. Revonsuo and M. Kamppinen (Eds), *Consciousness in philosophy and cognitive neuroscience*. Hillsdale, NJ: Erlbaum.

Baars, B. J. (1997). *In the theatre of consciousness: The workspace of the mind.* Oxford: Oxford University Press.

Baars, B. J., Newman, J. & Taylor, J. G. (in press). Neuronal mechanisms of consciousness: A Relational Global Workspace framework. In S. Hameroff, A. Kaszniak, J. Laukes, & A. Scott (Eds), *Towards a science of consciousness: The second Tucson discussion and debates*. Cambridge, MA: MIT Press.

Buser, P., & Rougeul-Buser, A. (1995). Do cortical and thalamic bioelectric oscillations have a functional role? A brief survey. *Journal of Physiology (Paris), 89,* 249–254.

Cho, S.-B., & Kim, J. H. (1995). Multiple network fusion using fuzzy logic. *IEEE Trans. Neural Networks*, *6*, 497–501.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences, USA, 81,* 4586–4590.

Crick, F., & Koch, C. (1990a). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2,* 263–275.

Crick, F., & Koch, C. (1990b). Some reflections on visual awareness. *Cold Spring Harbor Symposium on Quantitative Biology, 15,* 953–962.

Damasio, A. R. (1994). *Descartes' Error*. New York: G.P. Putnam Sons.

Deschenes, M., Madariage-Domich, A., & Steriade, M. (1985). Dendrodendritic synapses in the cat reticularis thalami nucleus: A structural basis for thalamic spindle synchronization. *Brain Research, 334,* 165–168.

Durfee, E.H. (1993). Cooperative distributed problem solving between (and within) intelligent agents. In P. Rudomin *et al.* (Eds), *Neuroscience: From neural networks to artificial intelligence* (pp. 84–98). Heidelberg: Springer-Verlag.

Eccles, J. C. (1966). *Brain and conscious experience*. Heidelberg: Springer-Verlag.

Edelman, G. M. (1989). *The remembered present, a biological theory of consciousness.* New York: Basic Books.

Funke, K., & Eyse, U. T. (1992). EEG-dependent modulation of response dynamics of cat dLGN relay cells and the contribution of corticogeniculate feedback. *Brain Research, 573,* 217–227.

Fuster, J. M. (1980). *The prefrontal cortex*. New York: Raven Press.

Gazzanigga, M. S. (1985). *The social brain, discovering the networks of the mind*. New York: Basic Books.

Goldman-Rakic, P. S. (1988a). Changing concepts of cortical connectivity: parallel distributed cortical networks. In P. Rakic and W. Singer (Eds), *Neurobiology of the cortex* (pp. 177–202). Berlin: John Wiley and Sons Ltd.

Goldman-Rakic, P. S. (1988b). The prefrontal contribution to working memory and conscious experience. In O. Creutzfeld and J. Eccles (Eds), *The brain and conscious experience*. Rome: Pontifical Academy.

Groenewegen, H. J., & Berendse, H. W. (1994). The specificity of the 'nonspecific' midline and intralaminar thalamic nuclei. *Trends in Neuroscience, 4*(2), 52–58.

Grossberg, S., & Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks, 4,* 452–466.

Hampshire II, J.B., & Waibel, A. (1992). The Meta-Pi network: Building distributed knowledge representations for robust multisource pattern recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence, 14,* 751–769.

Harnad, S. (1994). Guest editorial — Why and how we are not zombies. *Journal of Consciousness Studies, 1*(2), 164–168.

Heilman, K. M., Watson, R. T. & Valenstein, E. V. (1985). Neglect and related disorders. In K.M. Heilman & E.V. Valenstein (Eds), *Clinical neuropsychology*. New York: Oxford University Press.

Jackendorf, R. (1987). *Consciousness and the computational mind.* Cambridge, MA: MIT Press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

Jacobs, R. A., & Jordan, M. I. (1993). Learning piecewise control strategies in a modular neural network architecture. *IEEE Trans. Systems, Man, and Cybernetics, 23,* 337–345.

Jasper, H. H. (1960). Unspecific thalamocortical relations. In J. Field, H. W. Magoun & V. E. Hall (Eds), *Handbook of neurophysiology, Vol. 1* (pp. 1307–1322). Washington, DC: American Physiological Society.

Johnson-Laird, P. N. (1988). *The computer and the mind.* Cambridge, MA: Harvard University Press.

Jones, E. G. (1985). *The thalamus.* New York: Plenum Press.

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science, 237,* 285–292.

Koch, C. (1997). Computation and the single neuron. *Nature, 385,* 207–210.

Koch, C., & Poggio, T. (1992). Multiplying with synapses and neurons. In T. McKenna, J. Davis & S. F. Zornetzer (Eds), *Single neuron computation* (pp. 315–345). Boston, MA: Academic Press.

Krosigk, von M., Bal, T., & McCormack, D. (1993). Cellular mechanisms of a synchronized oscillation in the thalamus. *Science, 261,* 361–364.

LaBerge, D. L. (1990). William James symposium: Attention. *Psychological Science, 1*(3), 156–162.

LaBerge, D. L. (1995). *Attentional processing: The brain's art of mindfulness.* Cambridge, MA: Harvard University Press.

Llinas, R. R., & Pare, D. (1991). Commentary: of dreaming and wakefulness. *Neuroscience, 44*(3), 521–535.

Llinas, R., Ribary, U. Joliot, M. & Wang, X-J. (1994). Content and context in temporal thalamocortical binding. In G. Busaki *et al.* (Eds), *Temporal coding in the brain* (pp. 251–272). Heidelberg: Springer-Verlag.

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for

programming parallel processing structures on the fly. *Cognitive Science*, *9*, 113–146.

McClelland, J. L. (1986). The programmable blackboard model of reading. In J. L. McClelland & D. E. Rummelhart (Eds), *Parallel distributed processing*, Vol. 2 (pp. 122–169). Cambridge, MA: MIT Press.

McCormick, D. A., & Krosigk, M. (1992). Corticothalamic activation modulates thalamic firing through glutamate metabotropic receptors. *Proceedings of the National Academy of Science USA*, *89*, 2774–2778.

Mel, B. W. (1994). Information processing in dendritic trees. *Neural Computation*, *6*, 1031–1085.

Mesulam, M. (1985). *Principles of behavioral neurology*. Philadelphia: F.A. Davis.

Minsky, M. (1979). The society theory. In P. H. Winston & R. H. Brown (Eds), *Artificial intelligence, an MIT perspective, Vol. 1* (pp. 423–450). Cambridge, MA: MIT Press.

Mitrofanis, J., & Guillery, R. W. (1993). New views of the thalamic reticular nucleus in the adult and developing brain. *Trends in Neuroscience*, *16*, 240–245.

Monchi, O. & Taylor, J. G. (1995). A model of the prefrontal loop that includes the basal ganglia in solving a recency task, *Proceedings of the International Neural Network Society Annual Meeting*, July 1995. Washington, DC: International Neural Network Society Press.

Newell, A. (1992). SOAR as a unified theory of cognition: Issues and explanations. *Behavioral and Brain Sciences*, *15*(3), 464–492.

Newman, J. (1995a). Review: Thalamic contributions to attention and consciousness. *Consciousness and Cognition*, *4*(2), 172–193.

Newman, J. (1995b). Reticular-thalamic activation of the cortex generates conscious contents. *Behavioral and Brain Sciences*, *18*(4), 691–692.

Newman, J. (1997). Putting the puzzle together: Towards a general theory of the neural correlates of consciousness. *Journal of Consciousness Studies*, *4* (1 and 2), 47–66 and 99–120.

Newman, J., & Baars, B. J. (1993). A neural attentional model for access to consciousness: A Global Workspace perspective. *Concepts in Neuroscience*, *4*(2), 255–290.

Newman, J., Baars, B. J. & Cho, S-B. (1997). A neurocognitive model for attention and consciousness. In S. O'Nuallain, P. McKevitt & E. MacAogdin (Eds), *Two sciences of mind: Readings in cognitive science and consciousness*. Philadelphia, PA: John Benjamins of North America.

O'Nuallain, S., McKevit, P. & MacAogdin, E. (1997). *Two sciences of mind: Readings in cognitive science and consciousness*. Philadelphia, PA: John Benjamins of North America.

Parent, A., & Hazrati, L.-N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia–thalamo–cortical loop. *Brain Research Reviews*, *20*, 91–127.

Penfield, W. (1975). *The mystery of the mind: A critical study of consciousness and the human brain*. Princeton, NJ: Princeton University Press.

Penrose, R. (1994). *Shadows of the mind—In search of the missing science of consciousness*. Oxford: Oxford University Press.

Posner, M. I. & Rothbart, M. K. (1991). Attentional mechanisms and conscious experience. In A. D. Milner & M. D. Rugg (Eds), *The neuropsychology of consciousness* (pp. 11–34). London: Academic Press.

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

Reddy, D. R., Erman, L. D., Fennell, R. D., & Neely, R. B. (1973). The Hearsay speech understanding system: An example of the recognition process. *Proceedings of the International Conference on Artificial Intelligence*, 185–194.

Rummelhart, D. E., Hinton, G. E. & McClelland, J. L. (1986). A general framework for parallel distributed processing. In J. L. McClelland & D. E. Rummelhart (Eds), *Parallel distributed processing*, Vol. 1 (pp. 43–76). Cambridge, MA: MIT Press.

Scheibel, M. E., & Scheibel, A. B. (1966). The organization of the nucleus reticularis: A Golgi study. *Brain Research*, *1*, 43–62.

Scheibel, A. B. (1980). Anatomical and physiological substrates of arousal: A view from the bridge. In J. A. Hobson & M. A. B. Brazier (Eds), *The reticular formation revisited* (pp. 55–66). New York: Raven Press.

Sillito, A., Jones, H., Gerstein, G., & West, D. (1994). Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, *369*, 479–482.

Skinner, J. E. & Yingling, C. D. (1977). Central gating mechanisms that regulate event-related potentials and behavior. In J. E. Desmedt (Ed.), *Progress in clinical neurophysiology: Attention, voluntary contraction and event-related cerebral potentials, Vol. 1* (pp. 30–69). Basel: Karger.

Sompolinsky, H., Golomb, D., & Kleinfeld, D. (1991). Cooperative dynamics in visual processing. *Physical Review A*, *43*(12), 6990–7011.

Steriade, M., & Llinas, R. R. (1988). The functional states of the thalamus and the associated neuronal interplay. *Physiological Reviews*, *68*(3), 649–742.

Stryker, M. P. (1989). Is grandmother an oscillation? *Nature*, *338*, 297–337.

Taylor, J. G. (1992). Towards a neural network model of the mind. *Neural Network World*, *2*, 797–812.

Taylor, J. G. (1996). A competition for consciousness? *Neurocomputing*, *11*, 271–296.

Taylor, J. G. & Alavi, F. N. (1993). Mathematical analysis of a competitive network for attention. In J. G. Taylor (Ed.), *Mathematical approaches to neural networks* (pp. 341–382). Amsterdam: Elsevier Science Publishers B.V.

Taylor, J. G. & Michalis, L. (1995). The functional role of the hippocampus in the organization of memory. *Proceedings of the International Neural Network Society Annual Meeting*, July 1995. Washington, DC: International Neural Network Society Press.

Vibert, J., Pakdaman, K., & Azmy, N. (1994). Interneural delay modification synchronizes biologically plausible neural networks. *Neural Networks*, *7*, 589–607.

Walshe, K. W. (1978). *Neuropsychology, A clinical approach*. Edinburgh: Churchill Livingstone.

Yao, Y., & Freeman, W.J. (1990). Model of biological pattern recognition with spatially chaotic dynamics. *Neural Networks*, *3*, 153–170.

Yen, C. T., Conely, M., Hendry, S. H. C., & Jones, E. G. (1985). The morphology of physiologically indentified GABAergic neurons in the somatic sensory part of the thalamic reticular nucleus in the cat. *Journal of Neuroscience*, *5*, 2254–2268

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1207–1225, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00064-6

1997 SPECIAL ISSUE

Neural Networks for Consciousness

J. G. TAYLOR

Department of Mathematics, King's College and Institute for Medicine, Research Centre Juelich

(*Received* 9 *March* 1997; *accepted* 28 *June* 1997)

Abstract—*The paper outlines a three-stage neural network model for (i) the emergence of consciousness at its lowest level of phenomenal experience, (ii) the development of actions on the emerged conscious activity so as to generate higher-order consciousness. In the model, the lower first stage involves modules transforming inputs into various codes. It is only at the second level that further modules are activated in a manner which brings about the emergence of phenomenal awareness. The evidence in support of the initial two-stage model is briefly summarised. The two-stage model is then developed by means of a detailed neural framework which allows analysis of the extended duration of activity in posterior working memory sites.The third stage is then described in terms of known neural structures in the frontal lobes, and the manner they might support the creation of higher consciousness is briefly presented. © 1997 Elsevier Science Ltd.*

1. INTRODUCTION

The race for consciousness has started (Taylor, 1998). There is a large amount of interest in the neuroscience and neural network community on the topic. Various groups are seriously trying to track it down by the use of non-invasive instruments (PET, fMRI, EEG and MEG) by observing the networks of connected modules which function so as to enable various high level cognitive tasks to be solved by the brain.

The current body of knowledge on consciousness is rapidly increasing. There is already a large reservoir of material on it discovered by psychologists over the last century, and this is now being combined with the results coming from non-invasive instruments, and at single cell level in monkeys, to attempt to indicate where and how consciousness arises, and is supported in the brain.

Various models of consciousness have been suggested in the past. These give answers to the two questions indicated in Figure 1 associated with the question marks inserted there. The figure shows schematically the preprocessing of input by the modules labeled 1, and its further analysis by the modules 2. It is the latter which are supposed to support the initial appearance of consciousness of the incoming stimuli. The two question marks attached to the flow of activity in Figure 1 are related to the following basic questions about consciousness:

1. Q1: how is content supplied to the conscious experience once it has been created?

   The second question mark is attached to the ''hard'' question of consciousness (Levine, 1983):

2. Q2: how does the transformation of activity from the modules 1 to those in 2 cause phenomenal awareness to arise in the first place?

These two questions are closely related to each other, but are logically distinct in that content may have many possibilities, but the underlying phenomenal experience itself appears to have a relatively constant set of attributes (which will be considered in due course). Question 1 attempts to explain how these variations of awareness occur, whilst question 2 explores the principles behind that emergence in the first place. It might be suggested that one should answer question 2 before question 1, but indeed the former is very hard to give a convincing explanation. That is why a number of reasonably acceptable proposals have been made to answer question 1, but with no acceptable solution to question 2 in sight.

The most promising of the answers to question 1 is in terms of some form of feedback from earlier memories, both of semantic and episodic form. Thus, there is the 'Remembered Present' (Edelman, 1989), the 'Relational

Acknowledgements: The author would like to thank Prof. H.-W. Mueller-Gaertner, Director of the Institute for Medicine, for hospitality at the Institute where this paper was written, and to his colleagues A. Ionnides, L. Jaencke, S. Posse and J. Shah for stimulating conversations. I would also like to thank S. Petersen for the simulation of Figure 10.

Requests for reprints should be sent to Dept of Maths, King's College, Strand, UK; E-mail: j.taylor@mth.kcl.ac.uk.

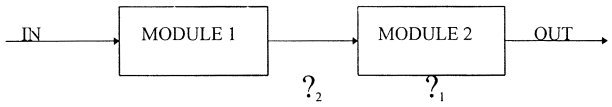


FIGURE 1. The basic two-stage model of awareness. Module 1 performs preprocessing in various modalities and codes; activity in module 2 is that on which phenomenal awareness is based. The two question marks denote the questions Q1 and Q2 stated in the text.

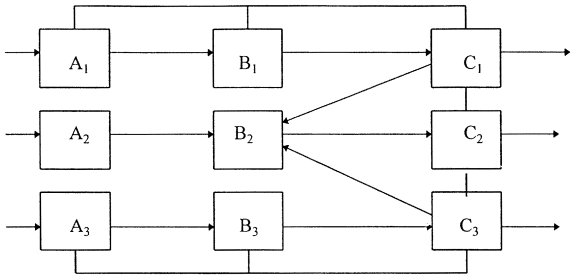


FIGURE 2. The three-stage model for the creation and control of awareness. The modules at the lowest or first stage are denoted A_1, A_2, ... and are involved only in low level feature analysis in the appropriate modality. A number of modules are involved in such processing. At the next stage are sites of 'buffer' memory in posterior cortex, denoted B_1, B_2, ... They are fed by preprocessed activity from first stage modules, and are coupled to each other so as to run a competition by lateral inhibition so as to select the most appropriate percept, given earlier context still present in activity traces on these modules, to enter phenomenal awareness. The final third stage is composed of modules which are frontal and possess attentional feedback control both to the second stage modules, so as to refresh activity there, and to the first stage modules to reduce or increase stimulus selectivity. Feedback is shown explicitly from the third to lower stages although it will also be present from the second to first stage (not all feedback is shown explicitly).

Mind' (Taylor, 1973, 1991) and the 'Predicted Present' (Gray, 1995) as candidate versions using, in one way or another, relational structures involving past memories in one form or another. However, these approaches do not seem to help solve question 2. It is the purpose of the first part of this paper to propose the bare bones of such a solution.

There is also the unresolved question as to how further higher level cognitive experience arises and is related to that of phenomenal experience. This leads to our third question:

Q3: Is there a dissociation between higher level cognitive processes (thinking, planning, self-awareness) and the lower level phenomenal experience, and if so what are the additional neural structures and mechanisms which support the higher level processes?

There has already been acceptance of some level of dissociation between passive awareness and controlled cognition in presenting the two-stage model of Figure 1, since otherwise it would not be possible to leave out the higher levels. The three stage model of Figure 2 does assume some dissociation but may simply be regarded as a summary of known anatomy, with the modules at stage 2 being in posterior cortex and stage-three modules in frontal lobe. However, the process of modeling the actions of the various parts of the more complete structures of Figure 2 would need to be performed with some knowledge of the level of any such dissociation. That will be considered at a later stage of the paper; for the moment we will assume that the methodology we are adopting, of modeling the lower two stages before attacking the third stage, will have some chance of being successful.

The detailed contents of the paper are as follows. It commences in Section 2 with an outline of the two-stage neural model for the emergence of consciousness. In this model the lower first stage involves modules transforming inputs into codes in the visual or other modalities. It is only at the second higher level that further modules are activated in a manner which brings about the emergence of phenomenal awareness. The evidence in support of such a two-stage model will be briefly summarised in Section 3.

The two-stage model is then developed in Section 4 by means of a detailed neural framework which allows analysis of the extended duration of activity in posterior working memory sites (which are posited as at the second stage in the model) using the notion of 'bubbles' of activity which are formed in neural models with enough neurons and suitably strong recurrent excitatory and inhibitory connections (Amari, 1977; Taylor, 1997a). Such models have recently been suggested as explaining the development of observed orientation sensitivity in primary visual cortex (Douglas et al., 1995; Somers et al., 1995; Ben-Yishai et al., 1995). The purpose of Section 4 is to discuss the implications of the resulting dynamical activity for the two-stage model of the emergence of phenomenal consciousness.

Detailed neural models are then briefly summarised for

1. the initial emergence of consciousness in terms of semi-autonomous 'bubbles' of activity in a neural system of the form of a continuum neural field theory (CNFT), as developed in a two-dimensional context (Taylor, 1997a);
2. application of CNFT to give brief initial explanations of various visual phenomena: apparent motion (due to moving bubbles) and stabilised image effects (destruction of bubbles);
3. how working memory modules, with extended temporal activity, can arise as from CNFT for suitably high density neuronal content (together with suitable ionically-driven adaptation);

The initial two-stage model is then extended to a third stage, as shown in Figure 2, in Section 5. There is much investigation presently as to the extent of possible dissociations between posterior sites of working memory, the

so-called 'buffer sites' of short term memory, and the anterior 'active memory' sites (Fuster, 1993). The nature of this evidence will be summarised and a third-stage model suggested in Section 6 which appears to be consistent with presently know data.

The paper finishes with a discussion, particularly on how the anterior activations can create the experienced features of higher consciousness.

2. THE TWO-STAGE MODEL OF AWARENESS

The two-stage model of the emergence of awareness in the brain is shown in Figure 1. There is supposedly no awareness of activity in stage 1 modules, whilst it is in modules at stage 2 that activity can enter awareness. What sort of neural networks would be suitable for such a separation? Some features of them have been summarised in Table 1, derived from analysis of introspection by psychologists and philosophers (and summarised in Metzinger, 1995). As seen from the table, there are spatial and temporal features of the networks which are of great importance. Thus, there are spatial features which require localised representations in sets of well-connected modules which are also well coupled to those modules involved in higher cognitive processes (the modules C of the third stage of processing in Figure 2).

The temporal features are also very specific, requiring suitably long temporal duration of activity (of the order of 300−500 ms) for phenomenal experience (PE) to arise but with little time between one experience and the next. Finally, there are emergent features which are also non-trivial, involving one-way and rapid emergence of PE at the highest level of a processing hierarchy with no ability to gain awareness of the activity of lower modules of stage 1.

Coding at the highest level of activity emerging into PE is most simply obtained by assuming that it arises from those first stage modules which are also coded at the second-stage level. The other most important feature in Table 1, that of the temporal aspect of duration, is a crucial part of the general features of the buffer stores in the psychologically-based working memory model (Baddeley and Hitch, 1974; Baddeley, 1986). These buffers have input from semantic-level coding, where all possible interpretations (in the case of words) have already been accessed in the semantic memory, so at the end of the stage 1 processing, before the buffer store is attained (Marcel, 1980; Levelt et al., 1991).

TABLE 1
Criterial features on neural networks to support PE

Spatial features:
(a) localised representations in localised modules
(b) modules well-coupled together
(c) modules well-coupled to higher-level modules
Temporal features:
(a) temporal continuation of activity
(b) time required to achieve PE activation
(c) no gap between different PE activations
Emergent features:
(a) one-way creation of PE activation
(b) PE created at highest level of a hierarchy
(c) rapid emergence once begun
(d) no ability to probe lower levels supporting PE but below PE creation level

The two-stage model of Figure 1 has been developed in a neural network implementation elsewhere (Taylor, 1996a). It uses semantic-level coding of activity at stage 1 and then buffered activity at stage 2, with lateral inhibition at this stage to achieve singling out of the contextually appropriate representation at the higher stage. It is this activity which then enters awareness by being broadcasting around the 'global workspace' of (Baars, 1988) or, in terms of the three stage model of Figure 3 gaining access to the anterior sites of working memory. Once at that third and highest level the material can be rehearsed or manipulated as desired. A neural network model of this higher level processing will also be described later in the paper.

Besides the construction of a model of the two-stage process, and due to the plethora of buffer stores now observed in posterior cortex in various modalities and codes (Salmon et al., 1996; Paulesu et al., 1993; Smith and Jonides, 1995) there must be some way of combining their various activities so as to achieve a unified experience of consciousness. Control structures in the brain are thereby needed which produce global correlations between the various component working memories. In particular, these control structures should support competition between various inputs, when suitably encoded, and only allow certain, most relevant, memories to be activated and related to the corresponding winning input. Such structures may also be involved with assessing the level of discrepancy of new incoming input with that predicted from later parts of activated stored pattern sequences. These aspects were considered (Taylor, 1992, 1993, 1994; Taylor and Villa, 1997), in terms of possible networks which could perform pattern matching and sustain competition. One of these involved the nucleus reticularis thalami (NRT), a sheet of mutually inhibitory neurons interposed between thalamus and cortex (see Figure 3).

It was suggested that the thalamus-nucleus reticularis thalami(NRT)-cortex complex may support such activities. This is due to the fact that the NRT is composed almost entirely of inhibitory neurons which are fed by activity coming from thalamus up to cortex and also by reciprocal cortico-thalamic connections. Since NRT also sends inhibitory inputs down onto thalamus, it is clear that such inhibition (which could also function as a release from inhibition if the NRT targets inhibitory interneurons in thalamus more effectively than the

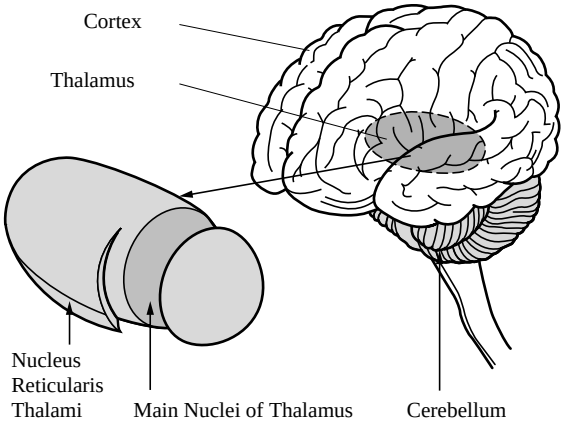


FIGURE 3. The Nucleus Reticularis Thalami (NRT) is a sheet of inhibitory neurons surrounding the upper and lateral parts of the thalamus. It only gives output to the thalamus as well as by lateral connections to other parts of itself, whilst receiving collaterals from both thalamo-cortical and cortico-thalamic axons.

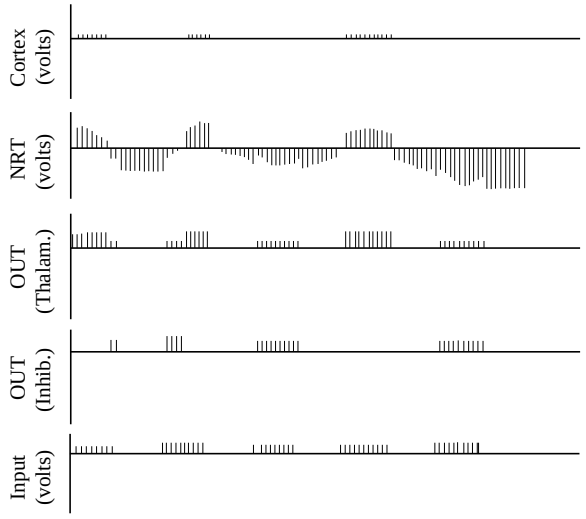


FIGURE 5. Results of a simulation of the circuit of Figure 4 for a line of 100 cells, for an input given on the bottom line. There is a global wave of activity set up over the NRT, as seen in the activation curve iin the second line of the figure. This causes a reduced level of input to cortical cells, as their activity indicate, shown on the top line.

excitatory relay cells to cortex) could exert a powerful control influence on cortical activity. This has been shown experimentally to be the case, with global effects of the NRT sheet especially being observed in the manner in which NRT controls the nature of cortical patterns of activity in sleep. There is also some evidence for a similar global form of control by NRT on allowed cortical activity in the non-sleep states (Skinner and Yingling, 1977; Villa, 1988).

It is possible, in a general manner, to understand this global control achieved by NRT of cortical activity in the following manner. Since any localised activity on

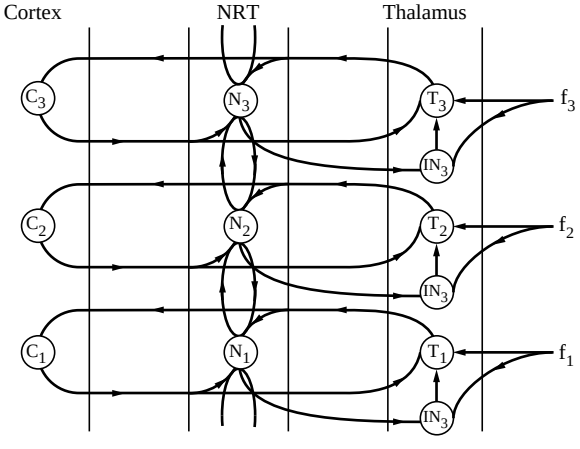


FIGURE 4. Circuit used to simulate the global competitive control action of the NRT over thalamic inputs to cortex, and hence to cortical activity itself. Input I enters both the main relay cells, denoted T, and the associated inhibitory interneurons IN in thalamus, and then feeds topographically to cortical cells C. There are collaterals from the cortico-thalamic axons and the thalamo-cortical ones feeding the NRT cells N, whcih are also laterally connected. These latter also feed back to both the inhibitory and relay cells in the thalamus, having greatest inhibitory effect on the former (from Alavi and Taylor, 1993).

the NRT will try to damp down activity on it elsewhere by lateral inhibition, the NRT can sustain 'bunched' spatially inhomogeneous activity, in which competition between neighbouring thalamic or cortical inputs onto it is occurring. This spatially structured activity may occur globally over the whole NRT sheet if it is well enough connected laterally, as is seen to be the case in spindle generation in sleep (Steriade et al., 1990). In this manner the NRT may function as a global controller of cortical activity. As such it appears of great relevance to include in models of the control circuitry for consciousness (Taylor, 1992, 1993, 1996b; Alavi and Taylor, 1992, 1993, 1995; Baars and Newman, 1993, 1994; Harth, 1995; Kilmer, 1996). Known local circuitry (Steriade et al., 1990) was used to construct a simulation of the coupled thalamo-NRT-cortical system following the circuit of Figure 4; the resulting activity arising from a simulation of 100 cortical neurons is shown in Figure 5. The resulting extended two-stage model is shown in Figure 6.

In summary the proposed two-stage model has the following features:

1. coding occurs up to semantic level (denoted SM) in the first (preprocessing) stage in Figure 6,
2. there is feedforward transfer of activity from the first to the second stage at approximately the same level of coding, with the feedforward map being
3. SM → WM
4. in a given code (as shown in Figure 6),
5. duration of activity is longest of all posterior modules in the WM of the second stage,
6. there is a process of global competition between

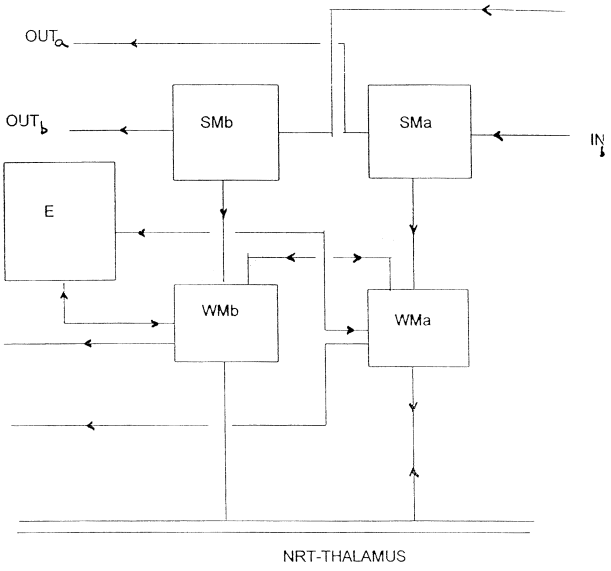


FIGURE 6. The resulting more detailed first two stages of the networks of the 3-stage model. The various features are discussed in the text.

activities on different WMs, supported by the TH/NRT/C system,

7. there is a strong reciprocal connectivity between the posterior WM sites and frontal areas at the highest level of coding (to allow rapid and effective access to the frontal 'active' working memory sites). Such connections are observed neuroanatomically.

The above model of the two-stages of phenomenal awareness leaves unanswered many detailed points. In particular:

1. what is the dynamical origin of the temporally extended activity traces on the working memory sites?
2. how are the codes in the WMs built up?
3. how do features become bound across codes and modalities?
4. what is the temporal dynamics of the emergence of awareness?

Answers to some of these questions (especially the 1st, 2nd and 4th) will be developed in Section 4. Before that we turn in the next section to describe some of the experimental basis for the separation of the emergence of awareness in two parts, as the 2-stage model claims.

3. EXPERIMENTAL SUPPORT FOR THE TWO-STAGE MODEL.

The two-stage model supposes that there are two sorts of cortical processing: firstly at the level of the stage-1 modules in Figure 1 and supporting neural activity of which there is no direct awareness, and secondly that at the level of stage-2 modules of Figure 1, of which there is direct awareness. This is not a new idea although it has been severely challenged in the past.

The reason for such a challenge is not too difficult to discern. If the two-stage model is true then there must be some properties possessed by the stage-2 modules which are not by the stage-1 modules. Such properties should be open to direct analysis by experiment and modeling, and perhaps allow the manner in which awareness is created to be determined. On the other hand, if there were no special modules in the brain supporting awareness then consciousness could preserve its monolithic mystery; it would still be some unknown feature possessed by a complex enough set of nerve cells but not one we can yet (if ever) see how to probe. Of course the two-stage model shades into the one-stage, monolithic one as the preprocessing stages of Figure 1 reduce in size and importance. However the existence of any modules at the non-aware level would still destroy the monolithic approach and make the task of probing consciousness more hopeful.

There are two main sources of support for the two-stage model: altered experience brought about by deficits, and that caused by suitably subtle paradigms for normal people. The first class contains modification of many forms of experience due to brain damage brought about by accident or disease. All of these are germane to the general program of explaining the mind in terms of the brain but some are more crucial to the task at hand, that of supporting the two-stage model, so will be described briefly in this section.

The two most important deficits to be discussed here are those of neglect and of blindsight. Neglect involves the inability to notice inputs to one side of the body. Thus, a patient will be unable to notice one side of their field of view; if such a patient is shown a picture containing two houses, placed one above the other, with one of the houses having flames emerging from its left side, then the patient may well prefer the house without the flames if asked to choose between them. This and other tests (Bisiach, 1988) indicate that the patient has knowledge up to a high level about the neglected side of their field of view but it is not part of their phenomenal experience.

Blindsight is a similar phenomenon in which there is knowledge without awareness. It is a phenomenon which has been well documented (Weiskrantz, 1986), and occurs to those who have lost a part of the primary visual cortex and so cannot see things in that part of the visual field. Yet they have knowledge of moving spots of light in their blindfield, since they can guess above chance as to where a spot of light has moved across a screen in front of them. It has even been found, in some cases, that the patient possesses some form of awareness of the movement-termed 'contentless awareness' (Weiskrantz et al., 1995), if it is fast enough. Thus, it would appear as if there is still the ability of modules after V1 (the primary visual area) to create awareness provided there is a large enough input by some alternate route.

Besides the cases brought about by damage and disease there is also the phenomenon of 'normal blindsight' (Kolb and Braun, 1995). A subject is required to pick out from an array of short bars oriented in a particular direction a small subset oriented in the perpendicular one. It is found that, provided the field of oppositely oriented bars is small enough, the subject is not aware of them but can be correct as to where they are more often than by chance alone. Here again there is knowledge without awareness, so that the orientation analysis, performed very likely in V1, is producing knowledge which is below the level of awareness.

Further support for V1 being below the level of awareness has also been given by an experiment which uses the adaptation of subjects to oriented bars, which takes place over about 5 s of viewing the bars (He et al., 1996). Once such adaptation has taken place, and before recovery, the contrast threshold for detecting oriented bars of the same orientation is found to have increased considerably. If the bars were 'crowded' by having bars of opposite orientation surrounding them in the field of view, loss of awareness of the particular bars now occurs but there is still an increase of the threshold for detecting bars presented at a variable contrast thresholds. The phenomenon of crowding is also found to be asymmetric between the upper and lower halves of the field of view, an asymmetry not possessed by V1. This means that V1 is involved in the adaptation to the orientation analysis but not in the final production of awareness. As the authors concluded 'activation of neurons in V1 is insufficient for conscious perception'.

This is further supported by the so-called waterfall effect or 'motion after-effect' (MAE). This occurs as an adaptation after about 30 seconds of viewing a moving scene; on cessation of the motion there is an apparent motion of the now static scene in the opposite direction. The time course of this has been investigated under fMRI (Tootell et al., 1996a) where they detected in subjects continued activity in the motion-sensitive area V5 for about the same length of time that the MAE persists in humans (about 9 s). even more interestingly they observed (Tootell, 1996b) a longer time course of the effect (by about 3 s) in higher visual areas beyond V5. This supports the notion that awareness arises in sites with about 3 seconds of duration of neural activations.

Single cell data from monkeys are also relevant to the search for the modules involved in the creation of consciousness. Binocular rivalry (in which different images are presented to the two eyes and the percept will switch every few seconds between the two possibilities) has been used in monkeys to determine in which area there is a predominance of neurons signaling the percept when it is experienced (which they can be trained to do faithfully). The use of this technique, over several years, has led Logethitis and colleagues (Logothetis and Schall, 1989; Leopold and Logothetis, 1996) to be able to show that, in vision, certain earlier areas (V1, V2, V4)

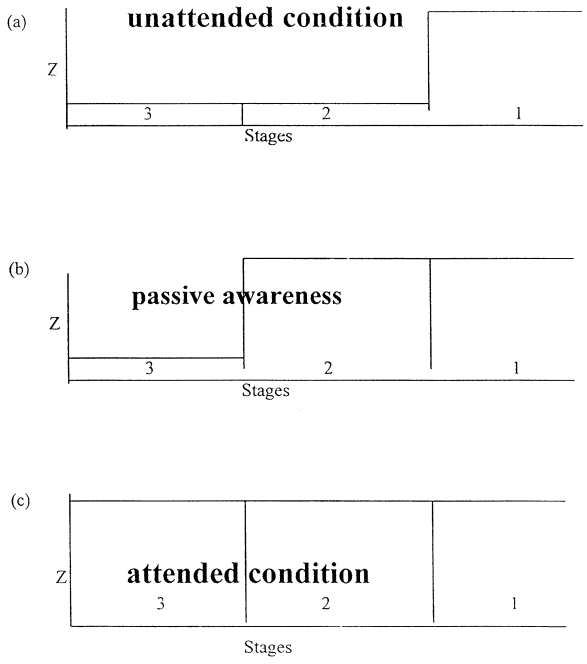


FIGURE 7. The levels of activation of the various modules of the 3-stage model for a given input under the three conditions of a) ignore b) passive awareness c) directed attention. Under condition (b) both the first two stage modules are activated, whilst under (c) all are on. There will be differences expected between the total levels of activation across the conditions due to feedback modulation, so the levels in the different conditions are normalized.

have about equal numbers of neurons firing both for the percept they support (in terms of direction of motion, for the paradigm in which the rivalry is between two sets of moving bars going vertically upwards versus moving downwards; a neuron supports a movement direction if it fires preponderantly when that direction of movement is presented) and the opposite one. As they wrote, 'The majority of cells in all areas continue to respond even when it is perceptually suppressed'. However their most recent analysis in the inferotemporal area (Sheinberg and Logothetis, 1997) has shown that in areas TE and TEO there is now a majority of active neurons in favor of the percept being experienced. Where the change-over area is sited is unclear and very important to discover.

There is also support from psychophysical experiments on word processing, in which the processing of a word presented visually is probed by a further auditorily-presented probe word (Levelt et al., 1991). The conclusion of the study was that a two-stage process occurs in which at the first stage there is automatic processing of all meanings of the word, which at a later stage are reduced to a suitably unique phonological representation. This is similar to the earlier two-stage 'logogon' model of (Morton, 1969, 1979) both supporting the proposal of the two-stage model of Section 2 that awareness arises at the second stage.

There is also increasing evidence in support of the

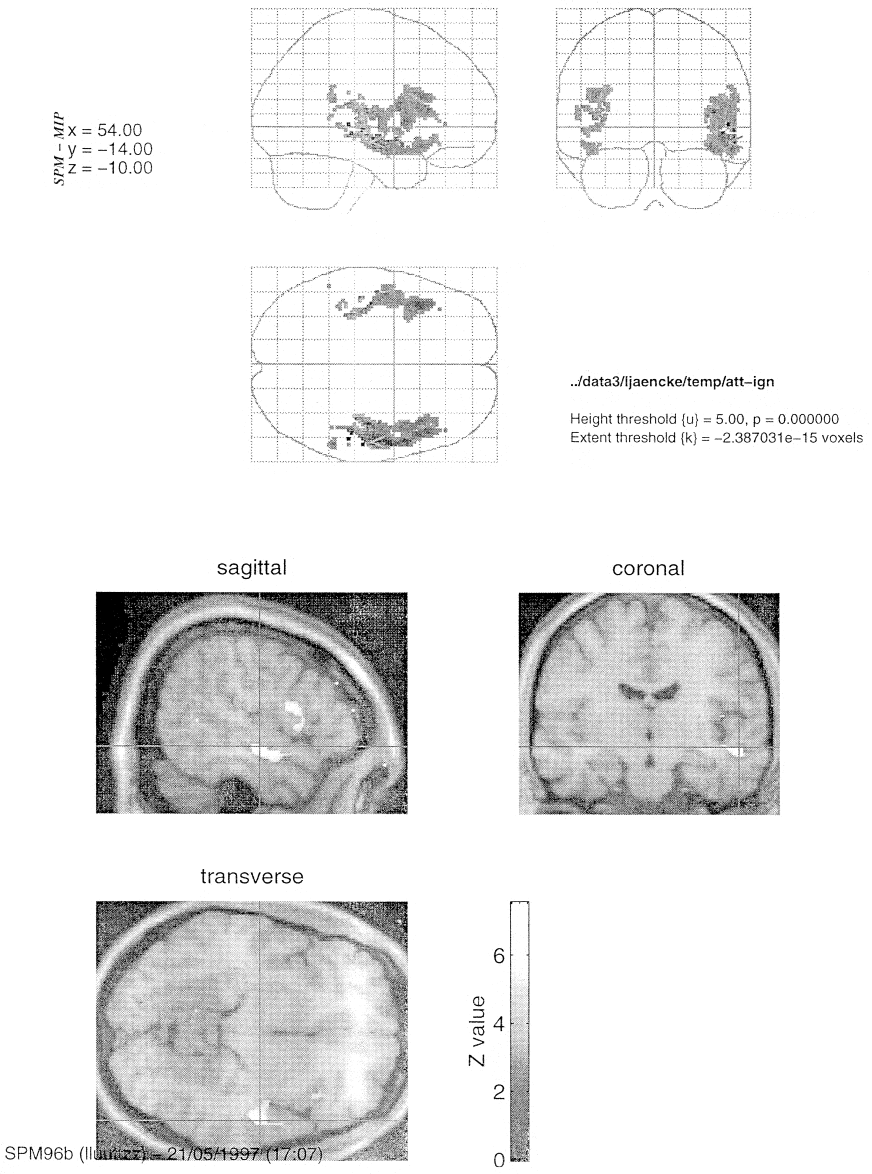


FIGURE 8. Differences in activation at various levels of the brain from an fMRI experiment (Taylor, Jaenke et al, 1997) in which the passive levels had subtracted from them levels due to ignore conditions for listening to a sequence of syllables. Note the non-zero levels of activity in various sites (threshold set at 5 leads to significance at $p < 0.01$).

2-stage model from non-invasive measurements. Thus, if a stimulus is ignored (by means of a suitably distracting task in another modality) the activity in cortical areas can be compared with that arising from passive awareness of the stimulus itself without any distractor. The 2-stage model would lead to the expectation of differences in activation along the lines of that shown in Figure 7. These differences in activation level have been seen in an auditory listening task with an added attentional load, as shown in Figure 8 (Taylor et al., 1997). This is taken from an fMRI study of five subjects in which there is comparison between subjects ignoring a sequence of syllables and listening passively to them. There is clear activation of regions in BA44/45 (Broca's area) and auditory areas when one passes from the ignore condition to the passive listening condition; these are candidates for the emergence of phenomenal awareness of syllables. The activation in Broca's area may be part of attetional processing and not due to phenomenal awareness.

Similar support arises from a recent MEG experiment (Vanni et al., 1996) in which subjects were presented with line drawings of objects or of non-objects, with a visual mask then delivered 30, 45 or 106 ms later. The level of the peak amplitude in only one visual area, the right lateral occipital (LO) area, was found to be significantly correlated with the level of awareness of the object drawings. This result indicates that the right cortical visual area LO is importantly involved in the creation of awareness of the objects described by the line drawings.

There is also a set of recent fMRI studies determining areas activated by various working memory tasks, such as the *n*-back task (Cohen et al., 1996). In this task a subject observes a sequence of letters or numbers presented one per second and has to respond when the same symbol occurs after n presentations (where *n* has been told to the subject by the experimenter before the particular run, with values of $n = 1$ to 4). Various areas were found to be significantly active compared to a background task, in particular posterior sites in BA 40 and BA 18/19, as well as prefrontal sites when *n* became 2 or greater. This was especially so for the dorsolateral prefrontal cortex (DLPFC), a fact important for our analysis of the third stage modules of Figure 2. However this area was not active for $n = 1$, so corresponding to posteriorly sited buffer memory (holding an object in memory over 10 s), and supporting the dissociation between frontal (stage 3) and posterior (stage 2) cortical areas under low-load conditions.

The conclusion of this section is that there is strong experimental support for

1. the existence of two separate stages at the lowest level in Figure 1,
2. the localisation of activity associated with the second stage of Figure 1 to specialised sites in posterior cortex, usually termed sites of 'buffer memory',
3. the determination of sites also involved in the stage-3 modules of Figure 3, involved in the creation of higher -order consciousness.
4. dissociation of the third from the lower stages under low attentional load.

4. A NEURAL NETWORK FOR THE TWO-STAGE MODEL

4.1. The Basics of Bubbles

We now turn to the development of a neural model to help support the two-stage model of Figure 1 and use it to give tentative answers to the questions raised at the end of Section 2. In particular we will consider the questions:

1. what is the dynamical origin of the temporally extended memory traces residing in the short-term working memory sites considered in this section?
2. how are the codes in these sites constructed?
3. what are the neural principles behind the creation of phenomenal awareness in these sites?

As stated in the Introduction we will employ continuum neural field theory (CNFT) whcih has been used over many years as a first approximation to cortical structure (Amari, 1977; Beurle, 1956; Ermentrout and Cowan, 1978). One of the important results of that period is the discovery of cortical 'bubbles' of activity which are initially created by input but which persist after stimulus offset (Amari, 1977). A brief introduction to bubbles and some of their properties in CNFT is given in this subsection, and relevant applications developed in the following one.

CNFT is based on a model of cortex as a continuous two-dimensional sheet of neurons with a lateral connection with the shape of a 'Mexican hat'. This allows the creation of localised activity, the bubbles, which can persist, and remain localised in spite of the absence of input due to the recurrent activity produced initially causing repeated firing of the neurons initially activated. The basic CNFT equation is constructed in terms of the membrane potential of a neuron and denoted by $u(x,t)$, at the point x and time t. It will be assumed that there is lateral connectivity other neural sheet defined by the lateral connection weight function $w(x - x')$ between the two neurons at the relevant points x. The connection weight will be usually taken to be of Mexican hat form as a function of the Euclidean distance $|x - x'|$. There is also an afferent connection weight function $s(x, y)$ from the thalamic position y to the cortical point x. The response function of a neuron will be taken to be determined by its mean firing rate, which is given as some function f of the membrane potential u of the relevant cell.

The membrane potential $u(x, t)$ will satisfy the CNFT equation (Amari, 1977)

$$\tau \partial u(x,t)/\partial t = -u(x,t) + \int dx' w(x-x') f[\, u(x',t)\,] + \int dy\, s(x,y)\, I(y,t) + h \qquad (1)$$

where $I(y, t)$ is the input to the thalamic position y at time t, h is the neuron threshold and the integration over the lateral connection weight is over the manifold M of neurons.

There are well-known autonomous solutions to (1) in the case when M is one-dimensional (Amari, 1977). In that case equation (1), for a static solution and with no input, becomes:

$$u(x) = \int w(x-x')\, 1[u(x')]\, dx' + h \qquad (2)$$

where the sharp threshold response function $f = 1$ (the step function) has been assumed in (2). A 'bubble' is defined to have a positive membrane potential over an interval, independent of input. This is formalised as

4.1.1. *Definition.* A 'bubble' of neural activity is a localised persistent solution to the CNFT equations. Its size is in general dependent on input, but its continued existence is not (to within effects of adaptation).

Let us consider the bubble extending from $x = 0$ to $x = a$:

$$u(x) > 0\,,\ 0 < x < a;\ u(0) = u(a) = 0 \qquad (3)$$

and otherwise $u < 0$. Then from (2) and (3), u is obtained

explicitly as

$$u(x) = \int_0^a w(x-x')\,dx' + h = W(x) - W(x-a) \quad (4)$$

where the function W is defined by

$$w(x) = \int_0^x w(x')\,dx' \quad (5)$$

Necessary conditions for the bubble to exist are that the membrane potential vanishes at the ends of the interval $[0, a]$, so

$$u(0) = u(a) = 0 = W(a)' + h \quad (6)$$

It is then possible to show that $u(x) > 0$ for $0 < x < a$ if $h < 0$; $u(x) < 0$ otherwise.

Stability of the resulting solution then requires

$$dW(a)/da < 0, \ \vee \ w(a) < 0 \quad (7)$$

Thus, the one-dimensional bubble exists under the conditions (6) and (7).

There are a number of further important results derived in Amari (1977) concerning the nature of bubble solutions and their extension to input dependence which will be briefly summarised here:

1. the parameter ranges for h and for the parameters in W can be determined so as to allow for autonomous solutions of various types ($\varnothing$ or the trivial one, ∞ or the constant non-zero one, an a-solution as the bubble of finite length a described above, and a spatially periodic solution),
2. complete determination of those patterns which are stable and those which are unstable, from amongst the stationary solutions described above,
3. response to input stimulus patterns: a bubble of finite length moves to a position of maximum of the input,
4. two bubbles interact, if close, with attraction (from the Mexican hat connection weight function), if more distant with repulsion, and if very distant with no effect on each other,
5. there can occur spatially homogeneous temporal oscillations (between a layer of excitatory and one of inhibitory cells)
6. traveling waves can persist.

Returning to the full two-dimensional bubble in the region D (which will be called an $R[2]$-solution here), we define

$$W_\infty = \lim_{R\to\infty} W(R) \quad (8)$$

where

$$W(R) = \int_D w(x-x')dx' \quad (9)$$

where x is only allowed to be on the boundary of D, which has radius $|x| = R$.

It is now possible to extend the methods of Amari, 1977 to deduce the same results as in the one dimensional case for the questions (a) to (f) raised above. Using the same enumeration:

1. Theorem 1. In the absence of input:

   There exists a $\varnothing$ solution iff $h < 0$.
   There exists an ∞-solution iff $W_\infty > -h$.
   There exists an $R[2]$-solution iff $h < 0$ and $R > 0$ satisfies

   $$W(R) + h = 0 \quad (10)$$

   It is possible to extend the classification of the solutions for varying levels of the stimulus h. Let $W_m = \max_x W(x)$.

2. Theorem 2. The nature of the various solutions for different parameter ranges is as in Figure 9.

   To determine which of these solutions is stable it is necessary to extend the one-dimensional discussion of Amari (1977) to two (or higher) dimensions. From the two-dimensional extension of equation (6) the boundary of D, defined by the radius $R(t)$ at time t, satisfies the constraint

   $$u(R(t), t) = 0 \quad (11)$$

   On differentiation of equation (11) with respect to t and use of equation (1) there results

   $$dR/dt = -[G(R(t)) + h]/\tau v \quad (12)$$

   where v is gradient of u normal to ∂D and is negative. The equilibrium case results on setting the right

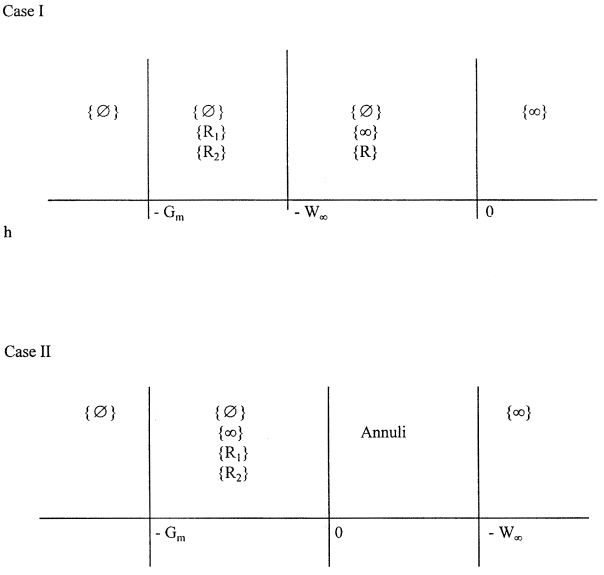


FIGURE 9. Parameter ranges for the existence of stable bubbles in two-dimensional CNFT (from Taylor, 1997a). The meaning of the paramteres is given in the text.

hand side of equation (12) to zero. The stability of this solution is determined by the sign of dG/dR: $dG/dR < 0 \Leftrightarrow$ stability. This leads to the stability classification of the solutions as given in theorem 2.

3. The response to stationary inputs of an $R[2]$-solution can also be treated as in Amari (1977). Consider a small stationary input $\varepsilon\ I(x)$, which is not assumed to be circularly symmetric so that the asymptotic bubble will not be circularly symmetric either. The equation of constraint is, following equation (12), that

$$dx/dt.\ \nabla u + \partial u/\partial t = 0\ on\ \partial D(t) \qquad (13)$$

Replacing the time derivative of u on the left hand side of equation (13) by equation (1) it is now possible to derive the condition for $x(t)$ on the boundary of $D(t)$,

$$dx/dt.\ \nabla u = (1\tau)[h + \varepsilon I + G(|x(t)|)] \qquad (14)$$

On expanding in a perturbation series in the small quantity ε, with

$$||x(t)|| = R_0 + \varepsilon(s,t) \qquad (15)$$

(where s denotes a parameter specifying which point on the boundary of the unperturbed $R[2]$ is being considered in the perturbation (15)) there results the constraint

$$d\varepsilon/dt.\ \nabla u = (1/\tau)[\varepsilon + \varepsilon.\ \nabla G(|x(t)|)\ ] \qquad (16)$$

where $\varepsilon = \varepsilon(s, t)$ is a vector describing the direction as well as the magnitude $\varepsilon(s, t)$ of the perturbation of the boundary point x, and the derivatives in (16) are evaluated at $\varepsilon = 0$, so at $x(t) = R_0$.
The result from the constraint (16) is that the net radial movement of the boundary of $D(t)$ is towards the region of largest input. There will be a movement of regions of $\partial D(t)$ towards lower values of the input, if these are positive, but there will be a larger velocity of movement towards those regions of the boundary nearer the maxima of I.

4. The one-dimensional result (d) above can also be extended in the same way, where the effect of one region (say $D1$) on another (say $D2$) is given, in terms of the lateral interaction term in equation (2), as the effective input to a neuron in $D2$ at the point x of amount

$$s(x) = \int_{D1} w(|x - x'|)dx' \qquad (17)$$

This will have the same effect as in the one-dimensional case, with attraction between the bubbles at $D1$ and $D2$ if they are close enough (as determined by $s(x)$), repulsion if the two regions are further separated, and ultimately no interaction between the bubbles at all if they are beyond the range of the lateral interaction term w (if that is finite).
5. The case of spatially homogeneous oscillations extends immediately to the two-dimensional case, since only ∞-solutions are being considered.
6. This case involves temporal structure and is considered more fully in Taylor (1997a).

4.2. Applications of Bubbles

We will apply the bubble solutions only to the cases of apparent motion, the fading of stabilised images and the lifetime of bubbles in cortex. Bubbles have also been applied to a variety of other cortical processes: the development of topographic maps, both in the one dimensional case (Takeuchi and Amari, 1979) and for two dimensions (Taylor, 1997a), control of saccades by the superior colliculus, on which the bubbles are supposed to form (Kopecz and Schoner, 1995), the modifications of the somatosensory constant topographic map by rapid unmasking due to removal of portion of an input (Petersen and Taylor, 1996a) or by relearning (Petersen and Taylor, 1996b), the guidance of head-directed cells (Zhang, 1996) and in explaining pre-attentive auditory memory (May et al., 1995). The two cases to be considered are closely related to features associated with the possible manner in which bubbles could enter directly into perception, and so are most relevant for discussion here.

4.2.1. *Apparent Motion.* This is the very well-known phenomenon in which a spot of light appearing on a screen in one place within 200 or so ms after the disappearance of one within 5 degrees of the second leads to the experience to a viewer of the motion of the first spot to the second. There has been considerable research on this phenomenon (see, for example, Cavanagh and Mather, 1990 and Ullman, 1979) and interesting attempts to correlate and explain it (Dawson, 1991). Here CNFT will be used to give an explanation at the level of principle, without attempting to explain all of the details of the phenomenon by the model (which is discussed more fully elsewhere).

The basic idea is the visual input from the initial spot creates a bubble in a cortical area (very likely MT) which is then dragged across the area by the appearance of the new spot in the manner arising from the result (d) above. The disappearance of the first spot and the appearance of the second causes a change in the background in which the bubble created by the first spot resides. It is thus attracted to the central region of activity brought about

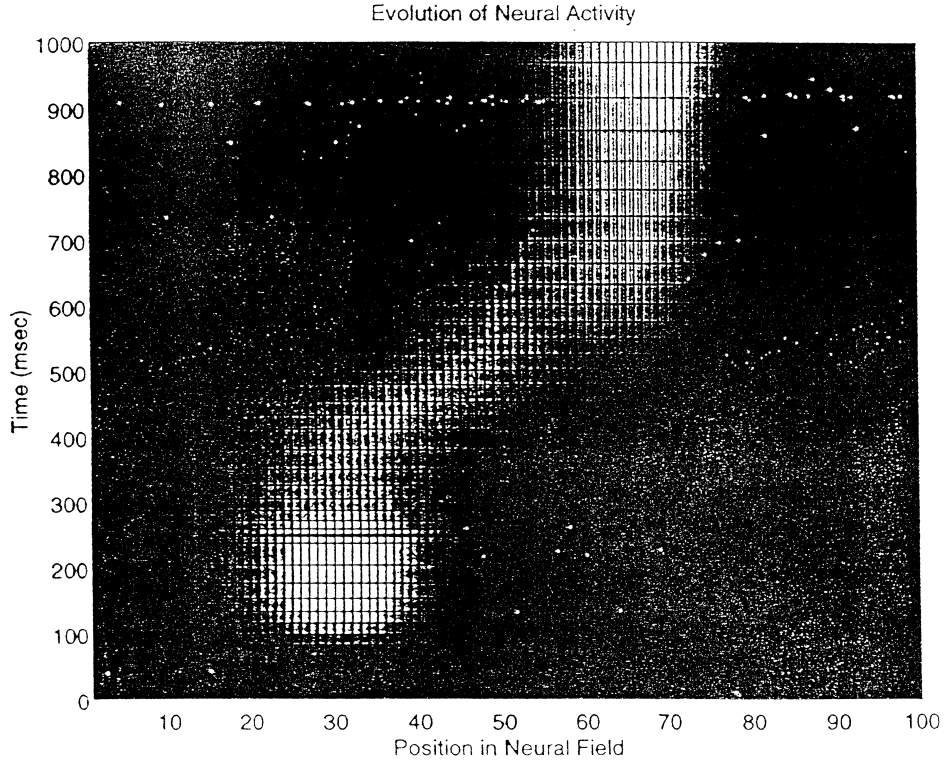


FIGURE 10. The dragging of a bubble, due to a localised input at the point 30 at time 0-300, formed in a line of neurons when the new input at the position 70 is switched on at the time 250. This is a proposed explanation of AM (see text).

by the second spot on its appearance, following the above argument. This process is difficult, although not impossible, to analyse in detail using the time-dependent equation (1), but a one-dimensional simulation is shown in Figure 10. In this, a bubble is created by a spot at time $t = 0$ at the position 30, using equation (1) with suitable parameters for the input to cause the creation of a bubble. This initial input is then removed at time 250 and a new input used at the position 70 at that time. As seen from the figure the bubble created by the first spot reduces in size when its input is removed but persists to move over to the position of the second input. It leaves a trail as it goes, causing the impression of a moving input.

4.2.2. *Stabilised Images.* These are experienced when the retinal image is kept fixed on the eye and the image ultimately fades from view (Pritchard, 1963). The initial loss of image may be due to adaptation in the retinal pathway, so reducing the image able to reach later stages. The process leads to very interesting patterns of image loss in which there are both Gestalt-type and ensemble effects observed. One particularly interesting effect was observed by (Crane and Piantanida, 1983) by stabilising the line between a rectangle of green and one of red. Normally it is not possible to observe both green and red together (they are opponent colours) but on stabilising the dividing line between them three different percepts were observed by subjects:

1. an overall 'red/green' colored region they had never experienced before,
2. dots of red and green interspersed throughout the visual field,
3. islands of red in a sea of green or vice versa.

It is possible to give a simple model possessing these three activations by means of two coupled CNFTs, one coding for red, the other for green. The red and green neurons at a given position are coupled inhibitorily to each other. The resulting model is shown in Figure 11. The coupled equations for the membrane potentials u and v for the two sets of colours are

$$du/dt = -u + w_{uu} * \theta(u) - w_{uv} * \theta(v) + h_u + I_u \tag{18a}$$

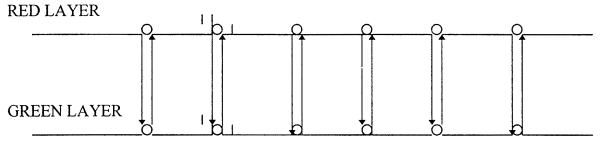


FIGURE 11. The architecture used for modelling red/green stabilised image percepts. See text for details.

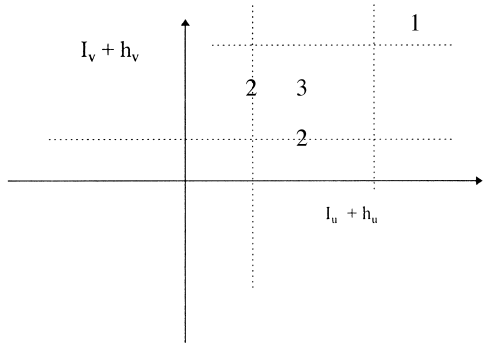


FIGURE 12. The regions of the ($I_u + h_u$, $I_v + h_v$)- plane associated with the possible solutions to the model of figure 11, as discussed in the text.

$$dv/dt = -v + w_{vv} * \theta(v) - w_{vu} * \theta(u) + h_v + I_v \quad (18b)$$

(where * denotes convolution over the cortical area, w_{uu} and w_{vv} are the usual Mexican hat functions and w_{uv}, w_{vu} are short range (excitatory) carriers of inhibition to their opposite colour). The inputs I_u and I_v are assumed to have been reduced on stabilisation of the image, so as to lead to a range of possible solutions to (18a) and (18b). It may be shown (Taylor, 1997b) that there are at least three forms of solution to (18):

1. ∞x∞ (in the notation of earlier in the section, where '∞' denotes an infinitely extended bubble),
2. interdigitating spatial oscillations,
3. solutions of the type $\{(-\infty, a_1] \cup [a_2, \infty\} \times [a_1, a_2]$ (giving the support of the bubble solution in the separate coordinates).

It is seen that these solutions correspond respectively to the experiences of subjects listed above. The range of the parameter plane ($I_u + h_u$, $I_v + h_v$) for which these solutions (i)–(iii) exist is shown in Figure 12 and Figure 13.

4.2.3. *Bubble Lifetimes.* Having seen that bubbles can help explain some of the phenomena of perception (and other features of cortical processing as well, as noted in the earlier references) we now turn to the important question as to how bubbles might disappear It is very unlikely that they persist for ever, and if they did so then they would present an ever increasing background of 'noise' interfering with current ongoing processing. There is even some evidence for their possessing a finite lifetime from work of Lu et al. (1992). They exposed subjects to a sequence of sounds with the inter-stimulus interval (ISI) being gradually increased. They discovered that the amplitude of the N100 response (100 ms after stimulus onset) reached saturation at suitably long ISI in both primary and secondary auditory cortex, with the rise to saturation corresponding to a decaying trace with lifetime of about 3 s in primary and 5 in secondary auditory cortex.

What mechanism could cause the decay of such a trace or more generally of the bubbles which they might represent? The most likely answer is that of adaptation of the responses of the neurons in the CNFT. Spike adaptation is a well-studied phenomenon (Connors et al., 1982; McCormick, 1986) arising from slow after-hyperpolarizing currents I_{AHP} and I_M, which can last for several seconds (Schwindt et al., 1988). These after-currents essentially raise the threshold for neuronal response. There may also be an effect from previous traces of neural activity which shunts out later activation (May et al., 1997).

The results of Lu et al. (1992) and of Uusitalo and

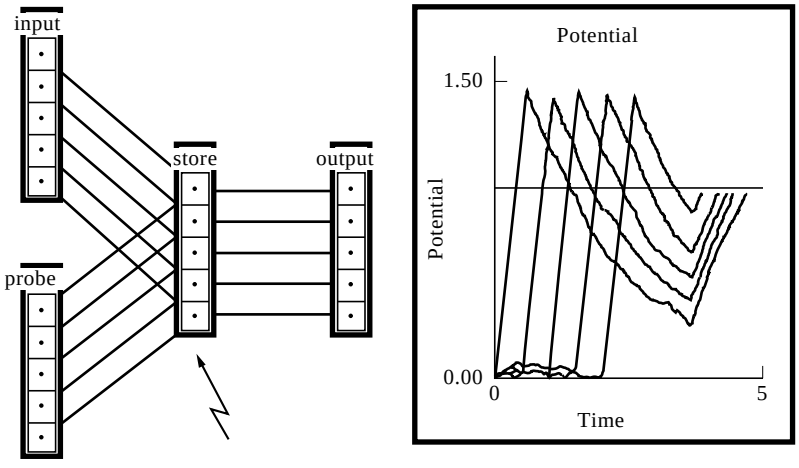


FIGURE 13. The model of recency in working memory (Hastings and Taylor, 1994). The nodes are dedicated to a given input item, and the resulting activity decays as corresponding to a leaky-integrator neuron. Probe re-activation has to lead to neuron activity above a criterial level before reponse. The resulting 'universal forgetting' formula for the reaction time to a new input as having been in the previously presented list fits known data well.

Williamson (1996) will be interpreted here in terms of the decay of a lasting trace of activity in the particular region from which the magnetic field was being measured (auditory or primary visual cortex respectively). This will be shown by using a very simple model of the response of a neuron, that of the leaky integrator (Hodgkin–Huxley) neuron, but with an after-hyperpolarizing current as observed in Schwindt et al. (1988). The equation for the membrane potential u of this neuron will be

$$\tau\, \mathrm{d}u/\mathrm{d}t = -u + I(t) - \lambda \int_0 \exp[-(t-t')/\tau']\, f[u(t')]\mathrm{d}t' \tag{19}$$

where $I(t)$ is the input current at time t, λ is the strength of an after-hyperpolarising current dependent on the response $f[u(t)]$ integrated over the time the neuron is active with an exponential decay of lifetime τ', and τ is the intrinsic lifetime of activity on the surface of the neuron (assumed to have no internal structure).

Under the experimental paradigm of Lu et al. (1992) the input is on for the time T and off for the time which we denote by the value ISI.

During the period that the input is on, u builds up its value driven by the input I. Let us suppose that the hyperpolarising lifetime τ' is much longer than the time T of duration of the input. This is valid for the long-lasting potassium-dependent hyperpolaristions mentioned by Schwindt et al. (1988), which we assume to be the one of equation (19). At the end of the input period the membrane potential begins to decay exponentially, so behaves as $\exp[-t/\tau]$ (multiplied by a suitable constant, and to within an additive constant).

We now assume that the term on the left-hand side of (19) can be neglected (in other words the time constant τ is relatively short) with the result that the time dependence of u at (or close to) the beginning of the next input, say at the N100 response, has the value (to within a constant initial value from $t = T$)

$$-\lambda \int_0 \exp[-(t-t')/\tau']\, f[u(t')]\mathrm{d}t' \tag{20}$$

where $t = T +$ ISI. For f chosen to be the semi-linear function $f(x) = (x)_+$ (the positive part of x) and assuming that the membrane potential is positive through the ISI (so that f is linear) then the membrane potential at $t = T =$ ISI, is from (20) equal to

$$constant - \lambda exp[-(T + ISI)/\tau] \tag{21}$$

For an ISI $= 0$ there will be no N100 so that (21) must then be equal to zero. Thus (21) becomes (for fixed T)

$$u(T + ISI) = A(1 - exp[-ISI/\tau]) \tag{22}$$

Formula (22) fits the results of Williamson and colleagues (1992, 1996). This justifies interpreting the value of the time constant (as the duration of the neural trace of activity in the cortical site from which they were measuring. A more complete analysis can be given, under other assumptions on the relation between τ and τ', with a similar result to that presented above.

There is the further question as to how neurons can possess such large and variable time constants, both across areas and across subjects, as have been measured by Williamson and colleagues (1992, 1996). We will now show that these variations are to be seen as arising from variable levels of recurrence, giving an increase to the intrinsically identical time constants of each neuron by an amount dependent on the amount of recurrence; this level of recurrence can vary from one cortical region (and from one subject) to another, and is crucially dependant on the cell density in the short-term memory stores.

In order to consider bubble decay we will consider in more detail the effect of the long-lasting after-hyperpolarisation current used in equation (19) on bubble lifetime. The two-dimensional expression replacing equation (19) is

$$\begin{aligned}\tau \partial u(x,t)/\partial t = &-u(x,t') + \int \mathrm{d}x'\, w(x-x')\, f[\,u(x',t)\,] \\ &+ \int \mathrm{d}y\, s(x,y)\, I(y,t) + h - \lambda \int_0^a \\ &\times \exp[-(t-t')/\tau']\, f[u(x,t')]\, \mathrm{d}t'\end{aligned} \tag{23}$$

where τ' is a measure of the lifetime of the adaptation current, and λ denotes its strength. It is now necessary to calculate the lifetime of a bubble created using equation (23). Let us first consider the one-dimensional case; that for two dimensions will follow straightforwardly.

A particular case of interest is when a bubble has initially been created by an input which is then removed. That could be due, for example, to the neural module acting as the source of the input having a shorter lifetime for the persistence of bubbles than the one under consideration. It would also occur if the bubble is created in a primary sensory module and the input itself has been modified.

To discuss this case it is appropriate to first reduce even further to a single recurrent neuron. For that case the membrane potential equation, from (23), is:

$$\begin{aligned}\tau \partial u(t)/\,\partial t = &-u(t) + w\theta[\,ut)\,] + h \\ &-\lambda \int_0^a \exp[-(t-t')/\tau']\, \theta[u(t')]\, \mathrm{d}t'\end{aligned} \tag{24}$$

where a step function response has been taken for the neuron. From equation

$$u(t) = u(0) + [1 - \exp(-t/\tau)][h + w] \tag{25}$$

$$-\lambda \int_0^t exp[-(t-t')/\tau]\mathrm{d}t' \int_0^{t'} \exp[-(t' - t')/\tau']\mathrm{d}t''$$

From (25), with $u(0) > 0$, $u(t)$ will remain positive initially in time. Moreover (25) reduces to the expression

$$\begin{aligned}u(t) = &\, u(0) + [1 - \exp(-t/\tau)][h + w - \lambda\tau\tau'] \\ &+ \lambda\tau(\tau')^2[\exp(-t/\tau') - \exp(-t/\tau)]/(\tau' - \tau)\end{aligned} \tag{26}$$

where the last term on the right hand side of (26) is replaced, for $\tau = \tau'$, by the expression $\lambda\tau\tau'\exp(-t/\tau)$. The last term in (26) may be neglected if $\tau' < \tau$, so that if

$$\lambda\tau\tau' > h + w + u(0) \tag{27}$$

then for suitably large t, $u(t)$ will become negative and the firing of the neuron will then cease. If no new input arrives then no further activity will ensue from the neuron.

The initial lifetime of the bubble is given by equating the right hand side of (26) to zero. Using the assumption that $\tau' \gg \tau$ in (26) gives the approximate value for the life-time T as

$$T = -\tau' 1n\{1 - u(0)/\tau[\lambda\tau\tau' - h - w]\} \tag{28}$$

where the factor $[\lambda\tau\tau' - h - w]$ is positive by (27). Equation (28) is the formula we wish to extend to the case of a one- and then a two-dimensional CNFT.

Firstly the case of a bubble solution infinitely extended in either dimension reduces to the above analysis with the constant w in the single neuron case being replaced by the quantities $w = \int w(x)\mathrm{d}x$, $w = \int w(x)\mathrm{d}x$ in the one and two dimensional cases respectively.

The relevant equation in one dimensions for a finite-sized bubble solution (Taylor, 1997a) has the extra adaptation term

$$-\lambda\tau\tau' \tag{29}$$

(dropping the term of $O(\tau')$ in (26)) and the added initial value $u(l(0), 0)$, where $l(t)$ is the size of the bubble at time t; the input term involving S has also to be dropped. The bubble will have a finite lifetime if the adaptation term is so negative that there exists a solution to the resulting equation for the asymptotic size of the bubble Amari, 1977:

$$h - \lambda\tau\tau' + W(21(\infty)) = 0 \tag{30}$$

where $W(x)$ is the first integral of the connection weight w over the bubble domain. Such a solution could arise if

$$\lambda\tau\tau' - h > W_m \tag{31}$$

where W_m is the maximum value of W. Thus if (31) is true then the bubble will have a finite lifetime given, under the same approximation as for the single neuron, by

$$T = \tau' 1n[-h(\lambda\tau\tau' - h\,W_m)] \tag{32}$$

This approximation should hold for both the one and two dimensional cases. In both cases we note that as W_m is increased, say by increase of cell density, the corresponding lifetime increases.

For the other extreme $\tau' > > \tau$ then τ and τ' must be interchanged in the lifetime formulae (28) and (32).

In conclusion, for the case $\tau > > \tau'$ the bubble lifetime is effectively proportional to τ, so dependent on whatever mechanism produces the bubble itself. In the opposite case the bubble lifetime is proportional to τ'. The latter quantity is expected to be an intrinsic characteristic of the single (pyramidal) neuron, so very likely constant throughout cortex. This characteristic difference between the bubble lifetimes in the two cases (in the former depending on the cortical area under consideration, in the latter constant throughout cortex) has interesting experimental consequences for cortical processing, which are explored in Taylor (1997a).

Finally the lifetime is seen to increase (logarithmically) as the lateral connection strength (so W_m in (32)) increases. Such an increase is slow, according to the formula (32), but this may be only a result of the assumption of a hard limiting threshold output function for the neurons. A smooth sigmoid response (which can still support bubble creation) can lead to a linear increase in lifetime. This may therefore explain the observed increase of lifetime as observed in Lu et al. (1992), as well as be the source of the buffer capability of working memory modules.

4.2.4. *The Buffer Stores.* A simple explanation of the recency effect, that more recent items in a short list are remembered better and faster, has been given by many groups; one of these, with a useful 'universal forgetting formula' was in Hastings and Taylor (1994). This models the short-term buffer store as a set of dedicated nodes which have decaying activity on them. A short list of items coded by these nodes has decaying activity on the store in which the strongest activity is that for the latest input. Recognition of the items presented then occurs from probe inputs which cause a temporal increase in activity until it reaches some criterial threshold for response. The form of this reactivation is shown in Figure 12. It is clear from the figure that the most recent input reaches the criterial threshold soonest, the earliest one taking the longest time. The resulting set of reaction times $RT(n, N)$ for the n'th item in a list of length N may be shown to be given by the 'forgetting' formula

$$RT(n, N) = a.1n\{b + c.\exp[d(n-N)]\} \tag{33}$$

where a, b, c, d are constants and in particular d is determined by the decay constant of the nodes. A very good fit to the experimental observations leads to a value of the lifetime of the activations on the nodes of about 1.5 s, which is in the same 'ball-park' as the lifetimes observed by Williamson et al. noted above.

We conclude that there is support for the existence of bubbles in cortex at the basis of phenomenal experience and that these can help to explain the somewhat activity-independent lifetimes observed in short-term memory tests and modeled by very simple dedicated nodes with lifetimes of about 100 times the decay constant of the single neurons themselves. This latter feature is explicable in terms of adaptation-driven bubble decay. The resulting lifetime of a bubble depends on the density of the cells in the area supporting it, with the longest

lifetime occurring for highest density. Finally we note that these bubble may be used to help explain the nature of the 'qualia'-like aspects of phenomenal experience, in particular its apparently intrinsic and non-relational characteristics Taylor (1997b).

5. THE THREE-STAGE MODEL

We now turn briefly to the third stage of the three-stage model of Figure 3. This stage receives conscious material from the second stage discussed above and uses it to achieve goals by planning and reasoning. We would expect that the regions of frontal cortex so involved would be the working memory sites of the prefrontal cortex, particularly areas 46 and 9, those observed to cause deficits in long-term planning and delayed tasks when lost due to accident or disease. There is also a clear change in personality and social responsiveness when there is loss of the mesial oribtofrontal regions. Such regions would also be implicated in the construction of higher order consciousness.

We have already given support, at the end of Section 3, to the suggested dissociability between the first two stages and that at the third, highest level assumed in the three-stage model of Figure 3. There is considerable further material for such dissociation between the stages from deficits of frontal patients, who appear to posses awareness of their deficits but not the concern that would be expected in their situations. Thus there is 'a dissociation between what the frontal patient knows or says, and how he or she behaves' (Stuss, 1991), a feature indicating dissociation between posterior knowledge and anterior action systems.

The earlier approach, in terms of the use of memory structures, termed the 'Relational Mind' model in Taylor (1973, 1991) can also be used to explain how content arises in higher consciousness. Thus there are representations of earlier experienced objects encoded in appropriate sites in frontal lobe and their re-activations by later input gives content to that input. These inputs arise from the posterior buffer sites discussed earlier, so that the coding of the frontal sites is guided by posterior activity and at the same time feedback to those sites would help their own representations. However the late onset of prefrontal myelination would prevent much use being made of such frontal sites in the first few years of life, a period when there is extensive coding of words and objects in posterior sites. Thus the prefrontal representations may only help in the development of more advanced and sophisticated concept representations than carried posteriorly.

There is now considerable experimental evidence from non-invasive instruments for the involvement of the above mentioned areas in so-called 'working memory' tasks, those requiring the holding of activity over long and variable periods of time for transformation and later response. Such tasks, for example in the encoding and later retrieval of words or pictures, have shown up a network of modules involving area 46 on both left and right as well as Broca's speech area 44/45. These new modules appear to have considerably longer lifetimes than those of the posterior cortex (Goldman-Rakic, 1992) so posing a question as to the mechanism behind such long activation. We will turn in the next section to consider a possible model for such long and adaptive lifetimes.

6. A NEURAL NETWORK FOR THE THREE-STAGE MODEL.

One of the clearest differences between the frontal and posterior cortices is the existence of strong cortico-thalamic recurrence between frontal sites and the medio-dorsal thalamus. There is also known crucial involvement of the basal ganglia in motor and cognitive processing, as defects brought about by loss of the basal ganglia show in patients with Parkinson's disease. The problem we are facing is that of modeling what Fuster (Fuster, 1993) calls 'active memory', that is persistent neural activities which can be acted upon and transformed into other ones so as to achieve some goal or other. It is useful to turn to consider what form representations of motor actions take in motor and related cortices. Strong support for population vector coding in motor and pre/supplementary motor cortices for output responses (coding of the direction of motor actions) has been presented in Georgopolous (1994). This also gives a mechanism for allowing the active memory regions of frontal cortex to be used as 'blackboards'. Material can be written on these by input causing the recurrent lateral cortical and recurrent thalamo-cortical loops to tend rapidly to the relevant attractors of the population coding; modulation by basal ganglia will act thereby to modify the thresholds and direct or change the attractors.

A neural model of such processing has been presented in the ACTION network of Taylor (1995) and Alavi and Taylor (1996), which has some similarity to models of Houk and Wise (1993). It has been applied to modeling delayed tasks and their deficits in frontal patients (Monchi and Taylor, 1995, 1997). There are also a further range of neural models of frontal components of working memory (Carpenter and Grossberg, 1993; Dominey and Arbib, 1992; Dominey et al., 1995: Guigon et al., 1994; Kirillov et al., 1993; Zipser et al., 1993). We will concentrate here on the ACTION network as being most closely related to the neuroanatomy of the frontal cortex.

6.1. The Action Network

The ACTION net is composed of the basal ganglia acting in a disinhibitory manner on the thalamo-cortical recurrent loops (with the presence of lateral cortico-cortical

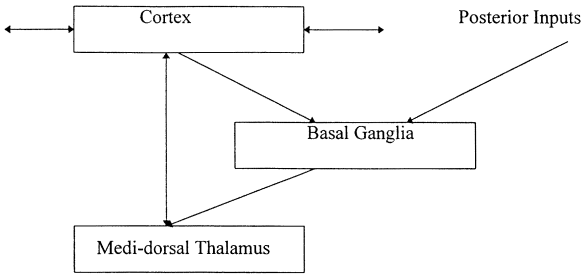


FIGURE 14. A schematic version of the ACTION network. See text for details.

connections present as well), so as to cause an attractor to be set up by an external input in the (possibly high) dimensional space given by the input features, as shown schematically in Figure 14. This generalises the two-dimensional population-vector action coding of the motor cortex to the other loops of the frontal system. The process of learning the correct connection weights for such an input set, so as to be able to write any material from posterior WM and other memory sites onto the frontal cortex, is presently under analysis (Taylor and Taylor, 1997), and has been considered in some detail in Dominey et al. (1995). These connections would allow the active memories to have written on them material from posterior buffer sites, from autobiographical/ episodic memory databases related to this material and from goal or other memory buffered in hippocampus.

The process of 'intermingling' of these activities with ongoing activity so as to achieve higher level consciousness is expected to require competitive processing for a similar reason to that discussed in connection with low-level consciousness and for which the NRT was employed earlier in the paper. That such competition can be supported by the ACTION network was suggested in Alavi and Taylor (1996), where the inhibitory nuclei in basal ganglia (striatum and globus pallidus) may be able to function in a similar manner to that of NRT. The competition would then involve threshold-changing processes, as performed by the disinhibitory action of striatum on thalamus.

There are five possible ACTION networks, corresponding to the 5 great frontal loops of Alexander and colleagues (Alexander et al., 1986) involving motor, premotor and supplementary motor cortex (action sequences), limbic (affect), orbitofrontal (social), frontal eye fields (eye movements) and dorsolateral prefrontal cortex (cognitive/ attentional). Each of the cortical regions is expected to have suitable connections for writing on them from posterior and limbic sites; this seems to be the case from neuro-anatomical knowledge, as mentioned earlier, and so supporting the 'well-connected character' required for the modules supporting consciousness noted in Table 1.

From the discussions so far it is clear that the re-excitation of episodic memories from hippocampal and mesio-orbital areas needs to be of a whole host of earlier memories, which can give the consciousness 'color' to experience. That might not best be achieved by a pattern completion or attractor network (Amit, 1990), since only a single pattern would result at one time. The most suitable memory structure is an associative matrix memory for a feedforward net, in which a given input pattern will excite as many outputs as previously stored patterns with strong overlap with the input. However the use of an attractor net may in fact help explain the length of time taken in memory search, as various of the attractors are activated and then rejected by comparison to some actively held template as part of the processing of the frontal lobes' ACTION network. This has been discussed from an interesting neural basis in Amit (1995).

7. DISCUSSION

Having constructed some of the possible neural machinery to support higher consciousness, it has now to be put to work. In particular it is important to show, though in the space available only very cursorily, how the frontal processes of attention, thinking, planning, prediction and schema learning and selection might be achieved. Also the place of emotions in such a galaxy of functions needs to be clarified. More general aspects of higher consciousness, such as intentionality and introspection, are considered elsewhere (Taylor, 1997a).

A number of these processes have already been hinted at in Taylor (1995). Thus attention can be decomposed into the component sub-processes of comparison, activation/inhibition and monitoring. All of these can be performed by the ACTION network (Alavi and Taylor, 1996), as can sequence learning and generation, at the basis of schema learning and production. Prediction is also involved in sequence generation, in which hippocampal activity, as in Gray's model (Gray, 1995; Kubota and Gabriel, 1995), should be included. Thinking has been discussed in Taylor (1995), using action transformations on representations written on prefrontal cortex so as to make such representations be closer, after comparison by the ACTION network, to a goal held in the Papez circuit. Planning can use similar techniques to discover transformations taking a given state to a goal state. Note the difference between thinking and planning; on this view the former uses whatever transformations that can be constructed so as to reach a goal—it is the sequence of states that are important—whilst the latter emphasises the set of transformations and not the states.

Emotions are considered as global activations from the limbic system when concerns (Frijda and Moffat, 1993), that is differences between desired and actual states, become large. Such comparisons (as concerns) are related to those of the model of Gray, but are used to give a global signal of affect to the cortical sites of

consciousness so as to cause modification of on-going behavior. This may be crucial in order to change schemata (Frijda and Moffat, 1993) as well as having an important effect in motivation and intelligence. Such an approach has been used in the neural models of delayed tasks and the Wisconsin card sorting task in Monchi and Taylor (1995, 1997) where there are three coupled ACTION networks with switching betweeen them according to valuation (concerns) by a net modeling the amygdala.

The above models of neural networks for consciousness have not only been sketched very briefly but also there has been omission of models for the limbic system, in particular the hippocampus and related areas. There is presently considerable interest in such areas and they clearly play an important role in determining the contents of consciousness. However this is something that the interested reader will have to look elsewhere (see, for example, Taylor, 1996c, 1998).

REFERENCES

Alavi, F. and Taylor, J. G. (1992). A simulation of the gated thalamocortical model (pp 929–932) In I. Aleksander and J. G. Taylor (Eds), *Artificial Neural Network 2*, North Holland: Amsterdam.

Alavi, F. and Taylor, J. G. (1993). Mathematical analysis of a global network for Attention, In J.G. Taylor (Ed.), *Mathematical Approaches to Neural Networks*, Elsevier.

Alavi, F., & Taylor, J. G. (1995). A global competitive neural network. *Biol Cybernetics*, *72*, 233–248.

Alavi, F. and Taylor, J. G. (1996). A basis for long range inhibition across cortex, In J. Sirosh, R. Miikulainen and Y. Choe (Eds.), *Lateral Interactions in the Cortex: Structure and Function*, Hpertext Book ISBN 0-9647060-8 (sited at http://www.cs.utexas.edu/users/nn/lateral-interaction-book/cover.html.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organisation of functionally segregated circuits linking basal ganglia and cprtex. *Annu Rev Neuroscience*, *9*, 357–381.

Amari, S.-I. (1977). Dynamical study of formation of cortical maps. *Biological Cybernetics*, *27*, 77–87.

Amit, D. (1990). *A Model of the Brain*, Cambridge: Cambridge University Press.

Amit, D. (1995). Psychology, neurobiology and modeling: the science of hebbian reverberations, In F. Ventriglia (Ed.), ch11 *Neural Modeling and Neural Networks*, Oxford: Pergamon Press.

Baars, B. J., (1988). *A Cognitive Theory of Consciousness*, Cambridge University Press.

Baars, B. J., & Newman, J. B. (1993). A neural attentional model for access to consciousness: a global workspace perspective. *Concepts in Neuroscience*, *4*, 255–290.

Baars, and B. J. and Newman, J. B. (1994). A neurobiological interpretation of global workspace theory, In A. Revonsuo and M. Kamppinenpp (Eds), *Consciousness in Philosophy and Cognitive Neuroscience*, (211–226) Hillsdale NJ: Lawrence Erlbaum.

Baddeley, A. D. and Hitch, G. (1974). Working memory, In G. A. Bower (Ed.), *The Psychology of learning and motivation*, *8*, 47–89, Academic Press, New York.

Baddeley, A. D. (1986). *Working Memory*, Oxford: Oxford University Press.

Ben-Yishai, R., Bar-Or, R.L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Nat Acad Sci (USA)*, *92*, 3844–3848.

Beurle, R. L. (1956). Properties of a mass of cells capable of regenerating pulses. *Trans. Roy. Soc. London B*, *240*, 55–94.

Bisiach E. (1988) The (haunted) brain and consciousness, In A. J. Marcel and E. Bisiach (eds.) *Consciousness in Contemporary Science*, New York: Oxford University Press, pp 101–120.

Cavanagh, and Mather, (1990). Motion: The long and short of it, *Spatial Vision*, 4, 103–129.

Carpenter, G., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in neuroscience*, *16*, 131–140.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E.E. (1996). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*, 604–608.

Connors, B. W., Gutnick, M. J., & Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, *48*, 1302–1320.

Crane, H. D., & Piantanida, T. P. (1983). On seeing Reddish Green and Yellowish Blue. *Science*, *221*, 1078–1080.

Dawson, M.R.W. (1991). The how and why of what went where in apparent motion: modeling solutions to the motion correspondence problem. *Psychological Review*, *98*, 569–603.

Dominey, P., & Arbib, M.A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, *2*, 153–173.

Dominey, P., Arbib, M., & Joseph, J.-P. (1995). A Model of corticostriatal plasticity for learning oculomotor associations and sequences. *J Cognitive Neuroscience*, *7*, 311–336.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981–985.

Edelman, G. (1989). *The Remembered Present*, New York: Basic Books.

Ermentrout, G. B. and Cowan, J. D. (1978). *Studies in mathematics*, The Math Assoc of America, 15, 67–117.

Frijda, N.H. and Moffat, D. (1993). A model of emotions and emotion communication, In *Proc ROMAN '93: 2nd IEEE International Workshop on Robot and Human Communication*, pp. 29–34.

Fuster, J.M. (1993). Frontal lobes. *Current Biology*, *3*, 160–165.

Georgopolous, A.P. (1994). New concepts in generation of movement. *Neuron*, *13*, 257–268.

Goldman-Rakic, P., (1992). *Working Memory and the Mind*, Scientific American, pp. 73–79.

Gray, J. A. (1995). The contents of consciousness: a neurophysiolgical conjecture. *Behavioural and Brain Sciences*, *18*, 659–676.

Guigon, E., Dorizzi, B., Burnod, Y. and Schultz, W. (1994). Neural correlates of learning in the prefrontal cortex of the monkey, *Cereb Cortex* (in press).

Harth, E. (1995). The sketchpad model. *Consciousness and Cognition*, *4*, 346–368.

Hastings, S. and Taylor, J.G. (1994). Modeling the articulatory loop, P. Morasso and M. Marinaro (Eds), *Proc Int Conf Art Neural Networks (ICANN'94)*, pp. 1452–1455.

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

Houk, J. C. and Wise, S. P. (1993). Outline for a theory of motor behaviour: Involving cooperative actions of the cerebellum, basal ganglia and cerebral cortex, In Rudomin, P., Arbib, M. A. and Cervantes, F. (Eds.), *From Neural Networks to Artificial Intelligence*, Springer, pp. 452–470.

Kilmer, (1996). Global inhibition for selecting modes of attention. *Neural Networks*, **9**(4), 567–573.

Kirillov, A. B., Myre, C. D., & Woodward, D. J. (1993). Bistability switches and working memory in a two-neuron inhibitory-feedback model. *Biol Cybern.*, *68*, 441–449.

Kolb, F.C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, *377*, 336–338.

Kopecz, K., & Schoner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biol. Cybern.*, *73*, 49–60.

Kubota, Y., & Gabriel, M. (1995). Studies of the limbic comparator:

limbic circuit training-Induced unit activity and avoidance behaviour in rabbits with anterior dorsal thalamic lesions. *Behavioral Neuroscience*, *109*, 258–277.

Leopold, and Logothetis, (1996). *Nature*, **379**, 549–553.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: a study of picture naming. *Psychological Review*, *98*, 122–142.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Phil. Quarterly*, *64*, 354–361.

Logothetis, and Schall, (1989). *Nature*, **245**, 761–763.

Lu, Z.-L., Williamson, S. L., & Kaufman, L. (1992). Human auditory primary and association cortex have different lifetimes for activation traces. *Brain Research*, *572*, 236–241.

Marcel, A. (1980). Conscious and preconscious recognition on polysemous words: Locating selective effects of prior verbal contexts, In Nickerson, R. S. ed. *Attention and performance VIII*, Hillsdale NJ: Erlbaum Assoc.

May, P., Himmelbach, M. and Taylor, J. G. (1997). (in preparation).

McCormick, D. (1986). Membrane properties and neurotransmitter actions, In: Sheperd (ed.) *The Synaptic Organization of the Brain*, 3rd edition, Oxford: Oxford University Press.

May, P., Tiitinen, H., Winkler, I., Naatanen, R. and Taylor, J. G. (1995). A neural network model of mismatch negativity and the underlying memory trace, Soc for Neuroscience Abstracts, san Diego.

Metzinger, T. (1995). The Problem of Consciousness, In (ed.) T. Metzinger, *Conscious Experience*, (pp. 3–40) Paderborn: Schoningh.

Monchi, O. and Taylor, J. G. (1995). A model of the prefrontal loop that includes the basal ganglia in solving the recency task, *Proc WCNN95*, Washington, INNS Press.

Monchi, O. and Taylor, J. G. (1997). A model of coupled frontal working memories for delayed response tasks and the wisconsin card sorting task, In (Ed.) P. P. Wang, (pp. 34–37) in *Int Joint Conf in Information Sciences, 2nd ICCIN'97*, Amsterdam: Elsevier.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Morton, J. (1979). Word recognition, In J. Morton and J. Marshall (eds.) *Psycholinguistics: Series 2. Structures and processes*, London: Elek, pp. 107–156.

Paulesu, E., Frith, C., & Frakowiak, R. S. J. (1993). The neural correlates of the verbal components of working memory. *Nature*, *362*, 342–345.

Petersen, R. S. and Taylor, J. G. (1996a). Reorganisation of somatosensory cortex after tactile training, pp. 82–88 in *Proc Conf on Neural Information Processing NIPS '96*, New York: Kaufmann.

Petersen, R. and Taylor, J. G. (1996b). Unmasking in somatosensory cortex, *Neural Information Processing* (in press).

Pritchard, R. M. (1963). Stabilized images on the retina. *Scientific American*, *204*, 72–78.

Salmon, E., Van der Linden, M., Collette, F., Delfiore, G., Maquet, P., Degueldre, C., Luxen, A., & Franck, G. (1996). Regional brain activity during working memory tasks. *Brain*, *119*, 1617–1625.

Schwindt, P. C., Spain, W. J., Foehring, R. C., Stafstrom, C. E., Chubb, M. C., & Crill, W. E. (1988). Multiple potassium conductances and their functions in neurons from cat sensorimotor cortex in vitro. *J Neurophysiol*, *59*, 424–449.

Skinner, J. E. and Yingling, C. D. (1977). Central gating mechanisms that regulate even-related potentials and behaviour (pp. 30–69) In Desmedt, J.E. (ed.), *Progress in clinical neurophysiology: Attention, voluntary contraction and event -related potentials*, Klinger, I.

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization. *Proc Nat Acad Sci*, *94*, 3408–3413.

Smith, E. E. and Jonides, J. (1995). Working memory in Humans: Neurophysiological Evidence, ch 66, pp. 1090–1020 in *The Cognitive Neurosciences*.

Somers, D. C., Nelson, S. B., & Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *J Neuroscience*, *15*, 5448–5465.

Steriade, M., Gloor, P., Llinas, R. R., Lopes da Silva, F. H., & Mesulam, M. M. (1990). Basic mechanisms of cerebral rythmic activities. *Electroenc. and Clin. Neurophysiol*, *76*, 481–508.

Stuss, D. T. (1991). Disturbance of self-awareness after frontal system damage, in *Awareness of Deficit After Brain Injury*, (eds.) G. P. Prigatano and D. L. Schachter, New York: Oxford University Press, pp. 63–83..

Takeuchi, A., & Amari, S.-I. (1979). Formation of topographic maps and columnar microstructures in nerve fields. *Biol Cybernetics*, *35*, 63–72.

Taylor, J. G. (1973). A model of thinking neural networks, Seminar, Institute for Cybernetics, Univ of Tübinjen.

Taylor, J. G. (1991). Can neural networks ever be made to think?. *Neural Network World*, *1*, 4–12.

Taylor, J.G. (1992). Towards a neural network model of the mind. *Neural Network World*, *2*, 797–812.

Taylor, J. G. (1993). A global gating model of attention and consciousness, In *Neurodynamics and Psychology*, ed. Oaksford, M. and Brown, G., Academic Press, New York.

Taylor, J. G. (1994). Goals, drives and consciousness. *Neural Networks*, *7*, 1181–1190.

Taylor, J. G. (1995). Modules for the mind of PSYCHE, *Proc World Congress on Neural Networks*, Washington 1995, Hillsdale NJ: L. Erlbaum/INNS Press.

Taylor, J. G. (1996). Breakthrough to awareness. *Biol Cybernetics*, *75*, 59–72.

Taylor, J. G. (1996). A global competition for consciousness? *Neurcomputing*, *11*, 271–296.

Taylor, J. G. (1996c). Modelling what it is like to be, In *Proc Arizona Conf on Scientific Basis of Consciousness*, (ed.), Hammeroff, S., Cambridge MA: MIT Press.

Taylor, J. G. (1997a). Neural 'bubble' dynamics in two dimensions I: Foundations; ibid II: Applications to vision, KCL/IME Preprints.

Taylor, J. G. (1997b). Cortical Activity and the Explanatory Gap, King's College preprint.

Taylor, J. G. (1998). *The Race for Consciousness*, Camrbidge MA: MIT Press.

Taylor, J. G., Jaencke, L., Shah, N. J., Noesselt, T., Schmitz, N., Himmelbach, T. and Mueller-Gaertner, H.-W. (1997). A Three-Stage Model of Awareness: Formulation and Intial Experimental Support, KFA Preprint, 1997 and Soc for Neurosci. Abstrs.

Taylor, N. and Taylor, J.G. (1997). (in preparation).

Taylor, J. G. and Villa, A. E. P. (1997). The Conscious I- A Neuro-Heuristic Approach to the Mind, to appear in The Brain of Homo Sapiens, Rome: Istituto della Enciclopedia Italiana.

Tootell, R. B. H., Reppas, J. B., Dale, A. M., Look, R. B., Sereno, M. I., Malach, R., Brady, T. J., & Rosen, B. R. (1996). Visual motion aftereffect in human cortical area MT revealed by functiona magnetic resonance imaging. *Nature*, *375*, 139–141.

Tootell, R. B. H. (1996b). Functional MRI maps of human visual cortical activity and consciousness, abstract 104, Consciousness Research Abstracts, J Consciousness Studies, 1996.

Ullman, S. (1979). *The Interpretation of Visual Motion*, Cambridge Mass: MIT Press.

Uusitalo, M. and Williamson, S. J. (1996). Memory Lifetimes of Human Visual Cortex, Human Brain Mapping Conference Abstract S565.

Vanni, S., Revonsuo, A., Saarinen, J., & Hari, R. (1996). Visual awareness of objects correlates with activity of right occipital cortex. *NeuroReport*, *8*, 183–186.

Villa, A. (1988). Influence de L'Ecorce Cerebrale sur L'Activite Spontanee et Evoque du Thalamus auditif du Chat, These, Univ de Lausanne.

Weiskrantz, (1986). *Blindsight*, Oxford: Clarendon Press.

Weiskrantz, L., Barbur, J. L., & Sahrie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proc Nat Acad Sci (USA), 92,* 6122–6126.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head—Direction cell ensemble: a theory. *J. Neuroscience*, *16*, 2112–2126.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *J. Neurosci., 13,* 3406–3420

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1227–1240, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00049-X

1997 SPECIAL ISSUE

Consciousness in Neural Networks?

EDMUND T. ROLLS

Department of Experimental Psychology, University of Oxford

(*Received* 31 *August* 1996; *accepted* 15 *January* 1997)

Abstract—*A combined neurophysiological and computational approach is reviewed that leads to a proposal for how neural networks in the temporal cortical visual areas of primates could function to produce invariant object representation and identification. A similar approach is then reviewed which leads to a theory of how the hippocampus could rapidly store memories, especially episodic memories including spatial context, and how later recall of the information to the neocortex could occur. Third, it is argued that the visual and memory mechanisms described could operate without consciousness, and that a different type of processing is related to consciousness. It is suggested that the type of processing related to consciousness involves higher-order thoughts ("thoughts about thoughts"), and evolved to allow plans, formulated in a language, with many steps, to be corrected. It is suggested that it would feel like something to be a system that can think linguistically (using syntax) about its own thoughts, and that the subjective or phenomenal aspects of consciousness arise in this way. It is further suggested that "raw sensory feels" arise in evolution because once some types of processing feel like something by virtue of a system capable of higher-order thoughts, it is then parsimonious to postulate that sensory and related processing, which has to be taken into account in that processing system, should feel like something. It is suggested that it is this type of processing, which must be implemented in neural networks, which is related to consciousness.* © *1997 Elsevier Science Ltd.*

Keywords—Consciousness, Hippocampus, Memory, Invariance, Visual recognition, Higher-order thoughts, Visual cortex.

1. INTRODUCTION

Advances are being made in understanding *how* the brain could perform some of the processing involved in perception and memory. These advances come in part from neurophysiological experiments in which the processing involved in vision and memory is analysed by recording the activity of single neurons in primates during these types of processing, and incorporating this information into computational models at the neuronal network level which provide an account of the ways in which many neurons in the networks found in different brain regions could perform the required computations. Examples of this approach are described first in this paper.

Having considered brain mechanisms involved in visual object recognition and memory, I then consider whether, once this processing is fully understood, we will have produced an account of the brain mechanisms underlying consciousness. I argue that we will not, and that it is a different type of information processing that is involved in consciousness. I outline a theory of what the processing is that is involved in consciousness, of its adaptive value in an evolutionary perspective, and of how processing in our visual and other sensory systems can result in subjective or phenomenal states, the "raw feels" of conscious awareness. These processes involved in consciousness must themselves be implemented in neural networks, but before considering how these processes are implemented, it is useful to be clear about what processing must be implemented.

2. NEURONAL NETWORKS INVOLVED IN INVARIANT VISUAL OBJECT RECOGNITION

2.1. Neurophysiology

The visual pathways project in primates by a number of

Acknowledgements: The author has worked on some of the experiments described here with G. C. Baylis, M. Booth, M. J. Burton, P. Georges-François, M. E. Hasselmo, C. M. Leonard, F. Mora, D. I. Perrett, R. G. Robertson, M. K. Sanghera, T. R. Scott, S. J. Thorpe, and F. A. W. Wilson, and their collaboration, and helpful discussions with or communications from M. Davies and C. C. W. Taylor (Corpus Christi College, Oxford), and M. Stamp Dawkins, are sincerely acknowledged. Some of the research described was supported by the Medical Research Council (PG8513579), and by The Human Frontier Science Program.

Requests for reprints should be sent to Professor E. T. Rolls, University of Oxford, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK; Tel.: +44-1865-271348; Fax: +44-1865-310447; e-mail: Edmund.Rolls@psy.ox.ac.uk

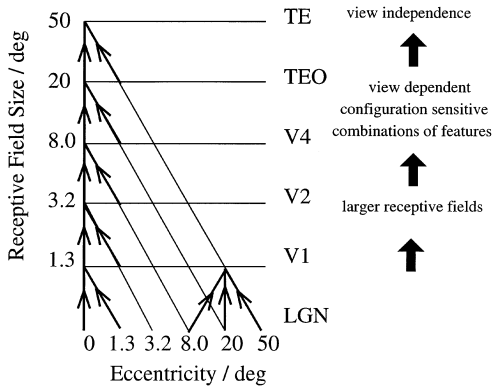


FIGURE 1. Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). LGN, Lateral geniculate nucleus. Area TEO forms the posterior inferior temporal cortex. The receptive fields in the inferior temporal visual cortex (e.g. in the TE areas) cross the vertical midline (not shown). (Reprinted from Wallis & Rolls, 1997.)

cortico-cortical stages from the primary visual cortex until they reach the temporal lobe visual cortical areas (see Figure 1, and for details of the neurophysiology summarized next, see Rolls, 1991, 1992, 1994b, 1995b, 1997). Along these pathways the receptive fields of neurons gradually become larger, as shown in Figure 1. (The receptive field of a neuron is the part of visual space within which appropriate visual stimuli can activate the neuron.) Part of the basis for this is the convergence onto neurons at any one stage of processing from a limited area of the preceding cortical area (see Figure 1). For this to result in neurons at the final stages of visual processing responding to the same object or stimulus independently of position on the retina, the appropriate connections must be set up in the hierarchy. Ways in which the appropriate synaptic weights to achieve this translation invariance could be learned are considered below.

The encoding that is provided of objects and faces at the end of this processing in the inferior temporal visual cortex is distributed, in the sense that the representation is not local or "grandmother cell" like, but instead many neurons are active to encode any one object (Rolls & Tovee, 1995; Rolls et al., 1996). Using an information-theoretic approach, it has been shown that the information available from the firing rates of a population of neurons about which visual stimulus (which of 20 equiprobable faces) has been shown on a single 500 ms presentation increases linearly with the number of neurons in the sample (Abbott et al., 1996; Rolls et al., 1997b). Because information is a logarithmic measure, this indicates that the number of stimuli encoded rises approximately exponentially, as the number of neurons in the sample increases. The consequence of this is that large numbers of stimuli, and fine discriminations between them, can be represented without (a receiving neuron) having to measure the activity of an enormous number of neurons. For example, the results of the experiments of Rolls et al. (1997b) indicate that the activity of 15 neurons would be able to encode 192 face stimuli (at 50% accuracy), of 20 neurons 768 stimuli, and of 25 neurons 3072 stimuli (Abbott et al., 1996; the values are for an optimal decoding case). This is strong evidence for distributed encoding. This type of encoding makes brain connectivity possible, in the sense that a receiving neuron can gain a great deal of information even when it does not receive vast numbers of inputs. Another interesting aspect of this encoding is that the information just described is available from the firing rates of the neurons, without taking into account the relative time at which the neurons fire. Thus temporal encoding is not an essential part of the code at this stage at least of visual information processing (see further Rolls et al., 1997b; Tovee & Rolls, 1995; Tovee et al., 1993). Another interesting aspect of the encoding is that much of the information from a population of neurons is available when the decoding is a simple neuronally plausible decoding involving a dot product of the neuronal activity in the current 500 ms (or 100 ms or 50 ms) presentation with that which occurred previously in the population of neurons to a particular stimulus (Rolls et al., 1997b). Such decoding could be performed by neurons which calculate their activation by a weighted sum of their input activity, which is common in neural network modelling. The fact that the information is available in a form in which it can be read out by this simple neuronally plausible dot product decoding with sampling from a limited set of neurons, and at the same time having the properties of a constant sparseness of the representation, and providing for generalization and graceful degradation, is probably what accounts for the fact that neurophysiologically interpretable information is available in the responses of *single* neurons about which stimulus has been seen (Rolls et al., 1997a; Tovee & Rolls, 1995; Tovee et al., 1993)[1]. This is one of the factors that allows single neuron recording to be so useful in understanding brain function—a correlation can frequently be found between the activity of even a single neuron and a subset of the stimuli being shown, of the motor responses being made, etc.

Some neurons in the temporal cortical visual areas have responses which are invariant not only for position on the retina, but also for the size, contrast, spatial frequency, position on the retina, and even angle of view

[1] The fact that the information increases approximately linearly with the number of neurons in the sample implies that the neurons convey almost independent information (if the stimulus set size is sufficiently large). If local encoding were used, the information would increase in proportion to the logarithm of the number of cells. If, for example, binary encoding were used (as, for example, numbers are encoded in a computer word), then the sparseness of the representation would fluctuate wildly, any receiving neuron would need to receive from all the input neurons, and generalization and graceful degradation would not occur.

(see Rolls, 1992, 1994b, 1995b, 1997; Rolls et al., 1996). It is clearly important that invariance in the visual system is made explicit in the neuronal responses, for this simplifies greatly the output of the visual system to memory systems such as the hippocampus and amygdala, which can then remember or form associations about *objects*. The function of these memory systems would be almost impossible if there were no consistent output from the visual system about objects (including faces), for then the memory systems would need to learn about all possible sizes, positions, etc., of each object, and there would be no easy generalization from one size or position of an object to that object when seen with another retinal size or position.

Other aspects of the neurophysiological findings which provide constraints on and guide the development of neural network theories about how the visual cortical areas involved in visual object recognition operate is that learning of new faces or objects can occur rapidly, within a few seconds; that the processing within any one cortical area is fast, with sufficient processing being completed within 30 ms in each cortical area in the hierarchy to subserve recognition; and that neurons in intermediate stages of processing (e.g. V2 and V4) respond to combinations of features present at earlier stages of processing (see Figure 1 and Rolls, 1992, 1994b, 1995b, 1997).

2.2. Computational Processes Involved in Invariant Visual Object Recognition

Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g. TE3, TEa and TEm), and anterior temporal cortical areas (e.g. TE2 and TE1). (This stream of processing has many connections with a set of cortical areas in the anterior part of the superior temporal sulcus, including area TPO.) There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g. 1° near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, e.g. 8° in V4, 20° in TEO, and 50° in inferior temporal cortex; Boussaoud et al., 1991) (see Figure 1). Such zones of convergence would overlap continuously with each other (see Figure 1). This connectivity would be part of the architecture by which translation invariant representations are computed (see Rolls, 1992, 1994b, 1995b, 1996a; Wallis & Rolls, 1997). Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g. sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back to many of the principal cells, which serves to decrease the firing rates of the less active neurons relative to the rates of the more active neurons (i.e. soft competition); and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (see Rolls, 1989c; Rolls & Treves, 1997). (A biologically plausible form of this learning rule that operates well in such networks is

$$\delta w_{ij} = k \cdot m_i (r_j' - w_{ij})$$

where k is a constant, δw_{ij} is the change of synaptic weight, r_j' is the firing rate of the jth axon, and m_i is a non-linear function of the output activation of neuron i which mimics the operation of the NMDA receptors in learning; see Rolls, 1989a, b, c; Rolls & Treves, 1997). Related approaches to self-organization in the visual system are described by Linsker (1986, 1988) and MacKay & Miller (1990).

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analysers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 s), the membrane of the postsynaptic neuron would still be in its "Hebb-modifiable" state, and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested (Rolls, 1992) that the short temporal window (e.g. 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Foldiak (1991) has proposed computing an average activation of the postsynaptic neuron to assist with the same problem. Another suggestion is that a memory trace for what has been seen in the last 300 ms appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared, as we have shown in masking experiments (see Rolls & Tovee, 1994; Rolls et al., 1994b). This continued firing could be implemented by local attractor networks in columns or modules in the cerebral cortex implemented by the local recurrent collaterals of the cortical

pyramidal cells (Rolls & Treves, 1997). Other invariances, for example, size, spatial frequency, and rotation invariance, could be learned by a comparable process. It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought. It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. Increasing complexity of representations could also be built in such a multiple layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons.

To test and clarify these hypotheses (see further Rolls, 1992, 1994b, 1995b, 1997) about how the visual system may operate to learn invariant object recognition, we have performed a simulation which implements many of the ideas just described, and is consistent with and based on much of the neurophysiology summarized above. The network simulated can perform object, including face, recognition in a biologically plausible way, and after training shows, for example, translation and view invariance (Wallis & Rolls, 1997; Wallis et al., 1993).

In the four-layer network, the successive layers correspond approximately to V2, V4, the posterior temporal cortex, and the anterior temporal cortex. The forward connections to a cell in one layer are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities to determine the exact neurons in the preceding layer to which connections are made. This schema is constrained to preclude the repeated connection of any cells. Each cell receives 100 connections from the 32 $\times$ 32 cells of the preceding layer, with a 67% probability that a connection comes from within four cells of the distribution centre. Figure 2 shows the general convergent network architecture used, and may be compared with Figure 1. Within each layer, lateral inhibition between neurons has a radius of effect just greater than the radius of feedforward convergence just defined. The lateral inhibition is simulated via a linear local contrast enhancing filter active on each neuron. (Note that this differs from the global 'winner-take-all' paradigm implemented by Foldiak, 1991). The cell activation is then passed through a non-linear activation function (e.g. sigmoid), which also produces contrast enhancement of the firing rates.

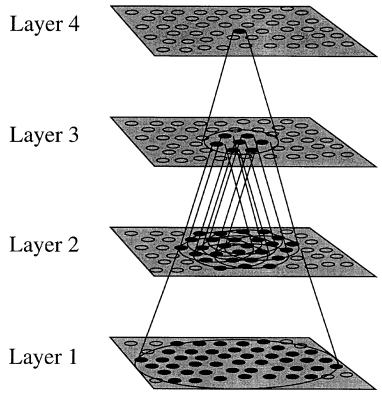


FIGURE 2. Hierarchical network structure used in the model of invariant visual object recognition. (Reprinted from Wallis & Rolls, 1997.)

So that the results of the simulation might be made particularly relevant to understanding processing in higher cortical visual areas, the inputs to layer 1 come from a separate input layer which provides an approximation to the encoding found in cortical visual area 1 (V1) of the primate visual system. These response characteristics of neurons in the input layer are provided by a series of spatially tuned filters with image contrast sensitivities chosen to accord with the general tuning profiles observed in the simple cells of V1.

The synaptic learning rule used in these simulations (VisNet) can be summarized as follows:

$$\delta w_{ij} = k m_i r_j'$$

and

$$m_i^t = (1-\eta) r_i^{(t)} + \eta m_i^{(t-1)}$$

where r_j' is the jth input to the neuron, r_i is the output of the ith neuron, w_{ij} is the jth weight on the ith neuron, η governs the relative influence of the trace and the new input (typically 0.4–0.6), and $m_i^{(t)}$ represents the value of the ith cell's memory trace at time t. In the simulations the neuronal learning was bounded by normalization of each cell's dendritic weight vector.

To train the network to produce a translation invariant representation, one stimulus was placed successively in a sequence of nine positions across the input, then the next stimulus was placed successively in the same sequence of nine positions across the input, and so on through the set of stimuli. The idea was to enable the network to learn whatever was common at each stage of the network about a stimulus shown in different positions. To train on view invariance, different views of the same object were shown in succession, then different views of the next object were shown in succession, and so on. It has been shown that this network, inspired by Fukushima's (Fukushima, 1980) neocognitron as well as by the neurophysiological data, can form cells in its final layer with translation, size and view invariant responses to stimuli presented on the 'retina' (Wallis & Rolls, 1997; Wallis et al., 1993).

These results show that the proposed learning mechanism and neural architecture can produce cells with responses selective for stimulus type with considerable

position, size or view invariance. The ability of the network to be trained with natural scenes is currently helping to advance our understanding of how representations of objects are built and encoded in the primate visual system.

This combined neurophysiological and computational approach is thus leading to biologically plausible theories about how the brain operates when it performs face or object recognition. In addition, there is now considerable evidence about what happens in our higher cortical visual areas when we recognize faces, and about how information about at least some classes of object in the world is represented in the visual system. Yet does this understanding of visual object recognition help us directly with the problem of consciousness, of why it is that it feels the way it does when we recognize a face? Would a computer which operated in the way described above be conscious during object recognition? I suggest that it would not be, and that for the object recognition processes to be conscious, including to feel like anything, the information from the type of visual processing system I describe would have to be projected to a different brain system, the nature of which will be described below. Before turning to that, some recent advances in understanding the brain processing that occurs when we store and then recall later everyday events are described, and I ask whether these memory processes are closer to consciousness.

3. THE HIPPOCAMPUS AND MEMORY

The hippocampus is implicated in a particular type of memory, the memory for recent events and episodes, in which there is frequently a spatial aspect or context (see for details Rolls, 1996b, d, 1997). In monkeys, a prototypical memory task impaired by damage to the hippocampal system is object-place memory, in which the locations of objects in space must be remembered (see Gaffan, 1994). This impairment is analogous to that shown by anterograde amnesic patients with damage to

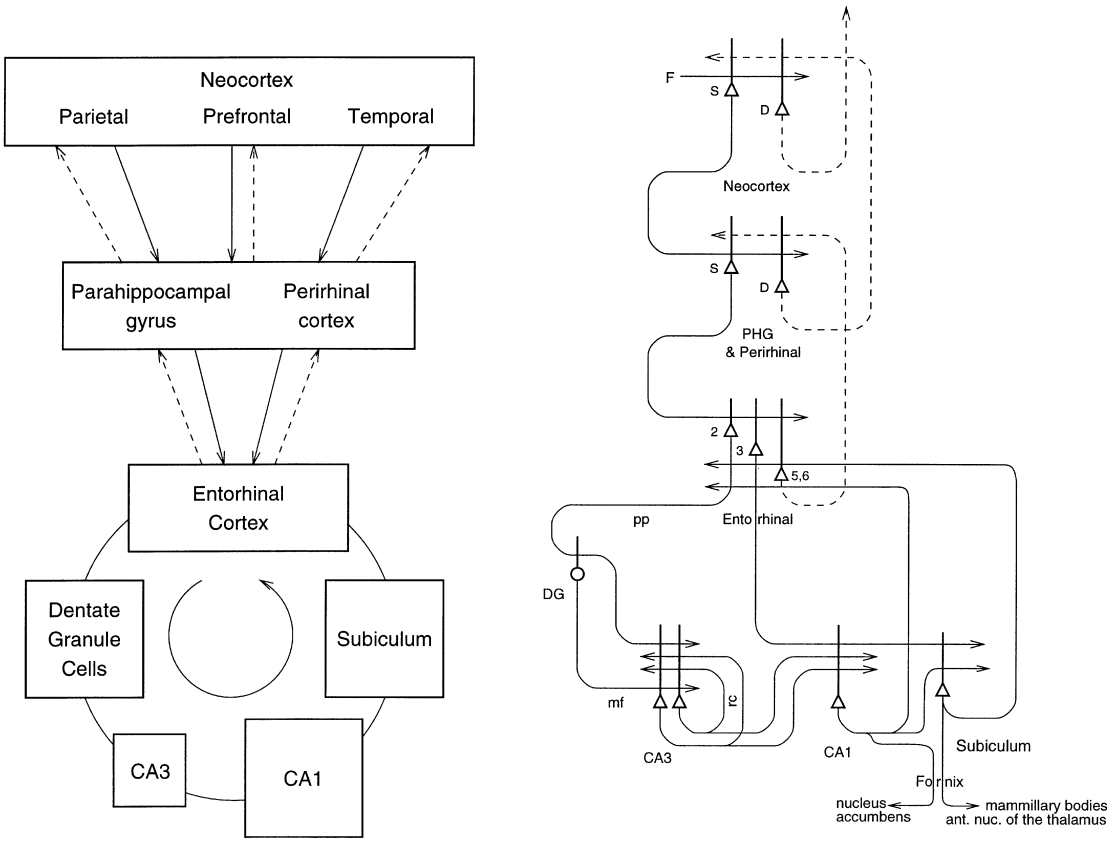


FIGURE 3. Forward connections (continuous lines) from areas of cerebral association neocortex via the parahippocampal gyrus and perirhinal cortex, and entorhinal cortex, to the hippocampus; and backprojections (dashed lines) via the hippocampal CA1 pyramidal cells, subiculum, and parahippocampal gyrus to the neocortex. There is great convergence in the forward connections down to the single network implemented in the CA3 pyramidal cells, and great divergence again in the backprojections. Left: block diagram. Right: more detailed representation of some of the principal excitatory neurons in the pathways. D, Deep pyramidal cells; DG, dentate granule cells; F, forward inputs to areas of the association cortex from preceding cortical areas in the hierarchy; mf, mossy fibres; PHG, parahippocampal gyrus and perirhinal cortex; pp, perforant path; rc, recurrent collateral of the CA3 hippocampal pyramidal cells; S, superficial pyramidal cells; 2, pyramidal cells in layer 2 of the entorhinal cortex; 3, pyramidal cells in layer 3 of the entorhinal cortex. The thick lines above the cell bodies represent the dendrites.

the hippocampus and nearby areas who cannot remember the locations of objects on a tray.

On the basis of these findings in humans and other animals, the hypothesis is suggested that the importance of the hippocampus in spatial and other memories is that it can rapidly form event or "episodic" representations of information originating from many areas of the cerebral cortex. In rats, hippocampal pyramidal cells (e.g. CA3 and CA1 neurons) respond when the rat is in a particular place in a spatial environment. In monkeys, it has been shown that there is a rich representation of space outside the monkey implemented by "spatial view" cells (see Rolls, 1996b, 1996d). These would provide an excellent representation of the spatial information needed to form a memory of where an object had been seen in space. It is suggested that an autoassociation network implemented by the CA3 cells of the hippocampus brings together the object information represented in temporal cortical visual areas, and spatial information represented in parietal areas, so that associations can be formed between objects and places (see Figure 3 and Rolls, 1989a, b, c, 1990a, 1996a,b).

A theory of how the networks shown in Figure 3 could operate, not only to store memories of events, but also to recall them to the neocortex via the backprojection pathways, has been developed (see Rolls, 1989a, b, 1996a; Rolls & Treves, 1997; Treves & Rolls, 1992, 1994). A way in which such recall could be useful in the cortex for building long-term semantic memories has been described by McClelland et al. (1995). A comparison of these approaches with others (for example by Burgess et al., 1994; and Hasselmo & Bower, 1993) is provided by Rolls (1996a), Rolls & Treves (1997) and Treves & Rolls (1994).

4. CONSCIOUSNESS

It would be possible to build a computer which would perform all the above functions of visual object recognition, memory storage and recall to the neocortex, and even emotion (Rolls, 1990b, 1995c), using the same computational principles described above, and yet we might not want to ascribe subjective or phenomenal states, which I shall call qualia, to this computer. We might not want to say that it feels like something to the computer when the computer is performing these functions. This raises the issue of in which networks in the brain would consciousness be represented. Because the topic of subjective or phenomenal feels or feelings (that it feels like something to be in that state) is of considerable current interest, and is for the present purposes the defining aspect of consciousness, one view on consciousness, influenced by contemporary cognitive neuroscience, is outlined next. However, this view is only preliminary, and theories of consciousness are likely to develop considerably. A reason for describing this view of consciousness is that we need to be clear about *what* must be implemented before considering *how* it could be implemented in neural networks.

A starting point is that many actions can be performed relatively automatically, without apparent conscious intervention. An example sometimes given is driving a car. Such actions could involve control of behaviour by brain systems which are old in evolutionary terms such as the basal ganglia. It is of interest that the basal ganglia (and cerebellum) do not have backprojection systems to most of the parts of the cerebral cortex from which they receive inputs (see, e.g. Rolls, 1994a; Rolls & Johnstone, 1992). In contrast, parts of the brain such as the hippocampus and amygdala, involved in functions such as episodic memory and emotion respectively, about which we can make (verbal) declarations (hence declarative memory, Squire, 1992) do have major backprojection systems to the high parts of the cerebral cortex from which they receive forward projections (Rolls, 1992; Rolls & Treves, 1997; Treves & Rolls, 1994; see Figure 3). It may be that evolutionarily newer parts of the brain, such as the language areas and parts of the prefrontal cortex, are involved in an alternative type of control of behaviour, in which actions can be planned with the use of a (language) system which allows relatively arbitrary (syntactic) manipulation of semantic entities (symbols).

The general view that there are many routes to behavioural output is supported by the evidence that there are many input systems to the basal ganglia (from almost all areas of the cerebral cortex), and that neuronal activity in each part of the striatum reflects the activity in the overlying cortical area (Rolls, 1994a; Rolls & Johnstone, 1992). The evidence is consistent with the possibility that different cortical areas, each specialized for a different type of computation, have their outputs directed to the basal ganglia, which then select the strongest input, and map this into action (via outputs directed, for example, to the premotor cortex) (Rolls & Johnstone, 1992; Rolls & Treves, 1997). Within this scheme, the language areas would offer one of many routes to action, but a route particularly suited to planning actions, because of the syntactic manipulation of semantic entities which may make long-term planning possible. A schematic diagram of this suggestion is provided in Figure 4. Consistent with the hypothesis of multiple routes to action, only some of which utilize language, is the evidence that split-brain patients may not be aware of actions being performed by the "non-dominant" hemisphere (Gazzaniga, 1988, 1995; Gazzaniga & LeDoux, 1978). Also consistent with multiple including non-verbal routes to action, patients with focal brain damage, for example to the prefrontal cortex, may emit actions, yet comment verbally that they should not be performing those actions (Rolls et al., 1994a). In both these types of patient, confabulation may occur, in that a verbal account of why the action was performed may be given, and this may not be related at all to the

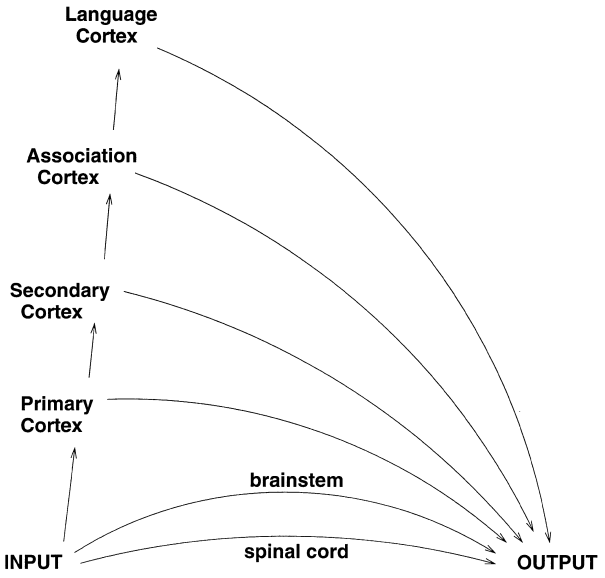


FIGURE 4. Schematic illustration indicating many possible routes from input systems to action (output) systems. Cortical information processing systems are organized hierarchically, and there are routes to output systems from most levels of the hierarchy.

environmental event which actually triggered the action (Gazzaniga, 1988, 1995; Gazzaniga & LeDoux, 1978). It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region such as those involved in some types of rewarded behaviour, the language system may subsequently elaborate a coherent account of why that action was performed (i.e. confabulate). This would be consistent with a general view of brain evolution in which as areas of the cortex evolve, they are laid on top of existing circuitry connecting inputs to outputs, and in which each level in this hierarchy of separate input–output pathways may control behaviour according to the specialized function it can perform (see schematic diagram in Figure 4). (It is of interest that mathematicians may have a hunch that something is correct, yet not be able to verbalize why. They may then resort to formal, more serial and language-like theorems to prove the case, and these seem to require conscious processing. This is a further indication of a close association between linguistic processing and consciousness. The linguistic processing need not, as in reading, involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements, and to perform syntactic operations on the semantic representations of states in the world, would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g. I think that he thinks that···), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others. This capability for higher-order thoughts would also allow reflection on past events, which would also be useful in planning. In contrast, non-linguistic behaviour would be driven by learned reinforcement associations, learned rules, etc., but not by flexible planning for many steps ahead involving a model of the world including others' behaviour. (For an earlier view which is close to this part of the argument, see Humphrey, 1980.) (The examples of behaviour from non-humans that may reflect planning may reflect much more limited and inflexible planning. For example, the dance of the honey-bee to signal to other bees the location of food may be said to reflect planning, but the symbol manipulation is not arbitrary. There are likely to be interesting examples of non-human primate behaviour, perhaps in the great apes, that reflect the evolution of an arbitrary symbol-manipulation system that could be useful for flexible planning; see Cheney & Seyfarth, 1990.) It is important to state that the language ability referred to here is not necessarily human verbal language (though this would be an example). What it is suggested is important to planning is the syntactic manipulation of symbols, and it is this syntactic manipulation of symbols which is the sense in which language is defined and used here.

It is next suggested that this arbitrary symbol-manipulation using important aspects of language processing and used for planning but not in initiating all types of behaviour is close to what consciousness is about. In particular, consciousness may *be* the state which arises in a system that can think about (or reflect on) its own (or other peoples') thoughts, that is, in a system capable of second- or higher-order thoughts (Rosenthal, 1986, 1990, 1993; compare Dennett, 1991). On this account, a mental state is non-introspectively (i.e. non-reflectively) conscious if one has a roughly simultaneous thought that one is in that mental state. Following from this, introspective consciousness (or reflexive consciousness, or self consciousness) is the attentive, deliberately focused consciousness of one's mental states. It is noted that not all of the higher-order thoughts need themselves be conscious (many mental states are not). However, according to the analysis, having a higher-order thought about a lower-order thought is necessary for the lower-order thought to be conscious. (A slightly weaker position than Rosenthal's on this is that a conscious state corresponds to a first-order thought that has the *capacity* to cause a second-order thought or judgement about it—Carruthers, 1996). This analysis is consistent with the points made above that the brain systems that are required for consciousness and language are similar. In particular, a system that can have second- or higher-order thoughts about its own operation, including its planning and linguistic operation, must itself be a language

processor, in that it must be able to bind correctly to the symbols and syntax in the first-order system. According to this explanation, the feeling of anything is the state which is present when linguistic processing that involves second- or higher-order thoughts is being performed.

It might be objected that this captures some of the process aspects of consciousness, what it is good for in an information processing system, but does not capture the phenomenal aspect of consciousness. I agree that there is an element of "mystery" that is invoked at this step of the argument, when I say that it feels like something for a machine with higher-order thoughts to be thinking about its own first- or lower-order thoughts. But the return point is the following: *if a human with second-order thoughts is thinking about his or her first-order thoughts, surely it is very difficult for us to conceive that this would* NOT *feel like something*? This is especially the case when the first-order thoughts are linguistic, and are about (grounded in) the real world.

It is suggested that part of the evolutionary adaptive significance of this type of higher-order thought is that it allows correction of errors made in first-order linguistic or in non-linguistic processing. Indeed, the ability to reflect on previous events is extremely important for learning from them, including setting up new long-term semantic structures. It was shown above that the hippocampus may be a system for such "declarative" recall of recent memories. Its close relation to "conscious" processing in humans (Squire has classified it as a declarative memory system) may be simply that it allows the recall of recent memories, which can then be reflected upon in conscious, higher-order, processing. Another part of the adaptive value of a higher-order thought system may be that by thinking about its own thoughts in a given situation, it may be able to better understand the thoughts of another individual in a similar situation, and therefore predict that individual's behaviour better (Humphrey, 1980).

As a point of clarification, I note that according to this theory, a language processing system is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have higher-order thoughts, and a first-order language processor (that might be perfectly competent at language) would not be conscious, in that it could not think about its own or others' thoughts. One can perfectly well conceive of a system which obeyed the rules of language (which is the aim of much connectionist modelling), and implemented a first-order linguistic system, that would not be conscious. (Possible examples of language processing that might be performed non-consciously include computer programs implementing aspects of language, or ritualized human conversations, e.g. about the weather. These might require syntax and correctly grounded semantics, and yet be performed non-consciously. A more complex example, illustrating that syntax could be used, might be: "If A does X, then B will probably do Y, and then C would be able to do Z." A first-order language system could process this statement. Moreover, the first-order language system could apply the rule usefully in the world, provided that the symbols in the language system (A, B, X, Y, etc.) are grounded (have meaning) in the world.) In line with the argument on the adaptive value of higher-order thoughts and thus consciousness given above, that they are useful for correcting lower-order thoughts, I now suggest that correction using higher-order thoughts of lower-order thoughts would have adaptive value primarily if the lower-order thoughts are sufficiently complex to benefit from correction in this way. The nature of the complexity is specific: that it should involve syntactic manipulation of symbols, probably with several steps in the chain, and that the chain of steps should be a one-off set of steps, as in a particular plan or sentence, rather than a set of well-learned rules. The first- or lower-order thoughts might involve a linked chain of "if⋯then" statements that would be involved in planning, an example of which has been given above. It is partly because complex lower-order thoughts such as these, which involve syntax and language, would benefit from correction by higher-order thoughts, that I suggest that there is a close link between this reflective consciousness and language. The hypothesis is that by thinking about lower-order thoughts, the higher-order thoughts can discover what may be weak links in the chain of reasoning at the lower-order level, and having detected the weak link, might alter the plan, to see if this gives better success. In our example above, if it transpired that C could not do Z, how might the plan have failed? Instead of having to go through endless random changes to the plan to see if by trial and error some combination does happen to produce results, what I am suggesting is that by thinking about the previous plan, one might, for example, using knowledge of the situation and the probabilities that operate in it, guess that the step where the plan failed was that B did not in fact do Y. So by thinking about the plan (the first- or lower-order thought), one might correct the original plan, in such a way that the weak link in that chain, that "B will probably do Y", is circumvented. To draw a parallel with neural networks: there is a "credit assignment" problem in such multistep syntactic plans, in that if the whole plan fails, how does the system assign credit or blame to particular steps of the plan? The suggestion is that this is the function of higher-order thoughts and is why systems with higher-order thoughts evolved. The suggestion I then make is that if a system were doing this type of processing (thinking about its own thoughts), it would then be very plausible that it should feel like something to be doing this. I even suggest to the reader that it is not plausible to suggest that it would not feel like anything to a system if it were doing this.

Two other points in the argument should be emphasized for clarity. One is that the system that is having

syntactic thoughts about its own syntactic thoughts would have to have its symbols grounded in the real world for it to feel like something to be having higher-order thoughts. The intention of this clarification is to exclude systems such as a computer running a program when there is in addition some sort of control or even overseeing program checking the operation of the first program. We would want to say that in such a situation it would feel like something to be running the higher-level control program only if the first-order program was symbolically performing operations on the world and receiving input about the results of those operations, and if the higher-order system understood what the first-order system was trying to do in the world. The second clarification is that the plan would have to be a unique string of steps, in much the same way as a sentence can be a unique and one-off string of words. The point here is that it is helpful to be able to think about particular one-off plans, and to correct them; and that this type of operation is very different from the slow learning of fixed rules by trial and error.

This analysis does not yet give an account for sensory qualia ("raw sensory feels"; for example, why "red" feels red), for emotional qualia (e.g. why a rewarding touch produces an emotional feeling of pleasure), or for motivational qualia (e.g. why food deprivation makes us *feel* hungry). The view I suggest on such qualia is as follows. Information processing in and from our sensory systems (e.g. the sight of the colour red) may be relevant to planning actions using language and the conscious processing thereby implied. Given that these inputs must be represented in the system that plans, we may ask whether it is more likely that we would be conscious of them or that we would not. I suggest that it would be a very special-purpose system that would allow such sensory inputs, and emotional and motivational states, to be part of (linguistically based) planning, and yet remain unconscious. It seems to be much more parsimonious to hold that we would be conscious of such sensory, emotional and motivational qualia because they would be being used (or are available to be used) in this type of (linguistically based) higher-order thought processing, and this is what I propose.

The explanation of emotional and motivational subjective feelings or qualia that this discussion has led towards is thus that they should be felt as conscious because they enter into a specialized linguistic symbol-manipulation system that is part of a higher-order thought system that is capable of reflecting on and correcting its lower-order thoughts involved, for example, in the flexible planning of actions. It would require a very special machine to allow this higher-order linguistically based thought processing, which is conscious by its nature, to occur without the sensory, emotional and motivational states (which must be taken into account by the higher-order thought system) becoming felt qualia. The qualia are thus accounted for by the evolution of the linguistic system that can reflect on and correct its own lower-order processes, and thus has adaptive value.

This account implies that it may be especially animals with a higher-order belief and thought system and with linguistic symbol manipulation that have qualia. It may be that much non-human animal behaviour, provided that it does not require flexible linguistic planning and correction by reflection, could take place according to reinforcement-guidance (using, e.g. stimulus-reinforcement association learning in the amygdala and orbitofrontal cortex, Rolls, 1990b, 1996c), and rule-following (implemented, e.g. using habit or stimulus-response learning in the basal ganglia, Rolls, 1994a; Rolls & Johnstone, 1992). Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals close to humans, with demonstrable syntactic manipulation of symbols, and the ability to think about these linguistic processes) would be different from other animals, and would have evolved qualia.

For processing in a part of our brain to be able to reach consciousness, appropriate pathways must be present. Certain constraints arise here. For example, in the sensory pathways, the nature of the representation may change as it passes through a hierarchy of processing levels, and in order to be conscious of the information in the form in which it is represented in early processing stages, the early processing stages must have access to the part of the brain necessary for consciousness. An example is provided by processing in the taste system. In the primate primary taste cortex, neurons respond to taste independently of hunger, yet in the secondary taste cortex, food-related taste neurons (e.g. responding to sweet taste) only respond to food if hunger is present, and gradually stop responding to that taste during feeding to satiety (see Rolls, 1989d, 1993, 1995a). Now the quality of the tastant (sweet, salt, etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety. The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex), and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it). The same argument arises for reward in general, and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is

reflected in the responses of neurons to visual stimuli (see Rolls, 1990b, 1995a, c). It is also of interest that reward signals (e.g. the taste of food when we are hungry) are associated with subjective feelings of pleasure (see Rolls, 1990b, 1993, 1995a, c). I suggest that this correspondence arises because pleasure is the subjective state that represents in the conscious system a signal that is positively reinforcing (rewarding), and that inconsistent behaviour would result if the representations did not correspond to a signal for positive reinforcement in both the conscious and the non-conscious processing systems.

Do these arguments mean that the conscious sensation of, for example, taste quality (i.e. identity and intensity) is represented or occurs in the primary taste cortex, and of the pleasantness of taste in the secondary taste cortex, and that activity in these areas is sufficient for conscious sensations (qualia) to occur? I do not suggest this at all. Instead, the arguments I have put forward above suggest that we are only conscious of representations when we have high-order thoughts about them. The implication then is that pathways must connect from each of the brain areas in which information is represented about which we can be conscious, to the system which has the higher-order thoughts, which, as I have argued above, requires language. Thus, in the example given, there must be connections to the language areas from the primary taste cortex, which need not be direct, but which must bypass the secondary taste cortex, in which the information is represented differently (see Rolls, 1989d, 1995a). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have higher-order thoughts about the pleasantness of the representation in the secondary taste cortex. There would also need to be pathways from the hippocampus, implicated in the recall of declarative memories, back to the language areas of the cerebral cortex (at least via the cortical areas which receive backprojections from the hippocampus, see Figure 3, which would in turn need connections to the language areas). A schematic diagram incorporating this anatomical prediction about human cortical neural connectivity in relation to consciousness is shown in Figure 5.

One question that has been discussed is whether there is a causal role for consciousness (e.g. Armstrong & Malcolm, 1984). The position to which the above arguments lead is that indeed conscious processing does have a causal role in the elicitation of behaviour, but only under the set of circumstances when higher-order thoughts play a role in correcting or influencing lower-order thoughts. The sense in which the consciousness is causal is then, it is suggested, that the higher-order thought is causally involved in correcting the lower-order thought; and that it is a property of the higher-order thought system that it feels like something when it is operating. As we have seen, some behavioural responses can be elicited when there is not this type of

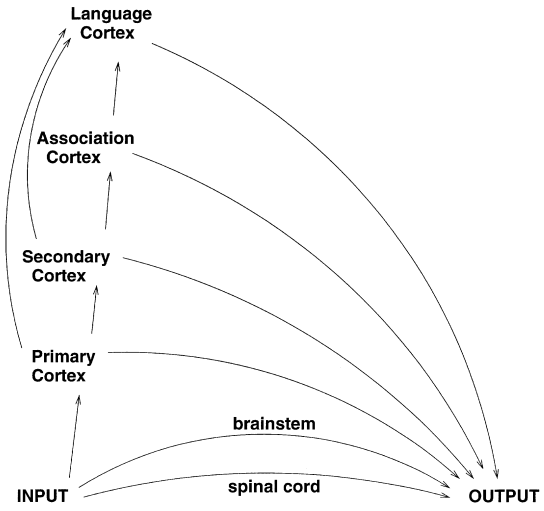


FIGURE 5. Schematic illustration indicating that early cortical stages in information processing may need access to language areas which bypass subsequent levels in the hierarchy, so that consciousness of what is represented in early cortical stages, and which may not be represented in later cortical stages, can occur. Higher-order linguistic thoughts (HOLTs) could be implemented in the language cortex itself, and would not need a separate cortical area. Backprojections, a notable feature of cortical connectivity, with many probable functions including recall (Rolls, 1989a, b, 1996a), probably reciprocate all the connections shown.

reflective control of lower-order processing, nor indeed any contribution of language. There are many brain processing routes to output regions, and only one of these involves conscious, verbally represented processing which can later be recalled (see Figure 4).

Some of the brain systems involved in this type of conscious processing that it is suggested has evolved to help the correction of plans are as follows. One module is a system that can implement syntax, because the many symbols (e.g. names of people) that are part of the plan must be correctly linked or bound. Such linking might be of the form: "if A does this, then B is likely to do this, and this will cause C to do this⋯". The requirement of syntax for this type of planning implies that an output to language systems in the brain is required for this type of planning (see Figure 4). Another building block for such planning operations in the brain may be the type of short-term memory in which the prefrontal cortex is involved. This short-term memory may be, for example, in non-human primates of where in space a response has just been made. A development of this type of short-term response memory system in humans to allow multiple short-term memories to be held in place correctly, preferably with the temporal order of the different items in the short-term memory coded correctly, may be another building block for the multiple step "if⋯then" type of computation so as to form a multiple step plan. Such short-term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of

non-human primates and humans (see Goldman-Rakic, 1996; Petrides, 1996), and may be part of the reason why prefrontal cortex damage impairs planning (see Shallice & Burgess, 1996).

It is of interest to comment on how the evolution of a system for flexible planning might affect emotions. Consider grief which may occur when a reward is terminated and no immediate action is possible (see Rolls, 1990b, 1995c). It may be adaptive by leading to a cessation of the formerly rewarded behaviour and thus facilitating the possible identification of other positive reinforcers in the environment. In humans, grief may be particularly potent because it becomes represented in a system which can plan ahead, and understand the enduring implications of the loss. (Thinking about or verbally discussing emotional states may also in these circumstances help, because this can lead towards the identification of new or alternative reinforcers, and of the realization that, for example, the negative consequences may not be as bad as feared.)

This account of consciousness also leads to a suggestion about the processing that underlies the feeling of free will. Free will would in this scheme involve the use of language to check many moves ahead on a number of possible series of actions and their outcomes, and then with this information to make a choice from the likely outcomes of different possible series of actions. (If, in contrast, choices were made only on the basis of the reinforcement value of immediately available stimuli, without the arbitrary syntactic symbol manipulation made possible by language, then the choice strategy would be much more limited, and we might not want to use the term free will, as all the consequences of those actions would not have been computed.) It is suggested that when this type of reflective, conscious, information processing is occurring and leading to action, the system performing this processing and producing the action would have to believe that it could cause the action, for otherwise inconsistencies would arise, and the system might no longer try to initiate action. This belief held by the system may partly underlie the feeling of free will. At other times, when other brain modules are initiating actions, the conscious processor may confabulate and believe that it caused the action, or at least give an account (possibly wrong) of why the action was initiated. The fact that the conscious processor may have the belief even in these circumstances that it initiated the action may arise as a property of it being inconsistent for a system which can take overall control using conscious verbal processing to believe that it was overridden by another system.

In the operation of such a free will system, the uncertainties introduced by the limited information possible about the likely outcomes of series of actions, and the inability to use optimal algorithms when combining conditional probabilities, would be much more important factors than whether the brain operates deterministically or not. (The operation of brain machinery must be relatively deterministic, for it has evolved to provide reliable outputs for given inputs.)

These are my initial thoughts on why we have consciousness, and are conscious of sensory, emotional and motivational qualia, as well as qualia associated with first-order linguistic thoughts. It is likely that theories of consciousness will continue to undergo rapid development, and current theories should not be taken to have practical implications.

5. DISCUSSION

Some ways in which the current theory may be different from other related theories follow. The current theory holds that it is higher-order *linguistic* thoughts (HOLTs) that are closely associated with consciousness, and this may differ from Rosenthal's higher-order thoughts (HOTs) theory (Rosenthal, 1986, 1990, 1993), in the emphasis in the current theory on language. Similarly, the theory differs from suggestions for a function of consciousness in "monitoring" (e.g. Marcel, 1988), in that a specification is given in the present theory of the type of correction being performed of first-order linguistic thought processes, and of the computational advantages of this. Language in the current theory is defined by syntactic manipulation of symbols, and does not necessarily imply verbal language. The reason that strong emphasis is placed on language is that it is as a result of having a multistep flexible "on the fly" reasoning procedure that errors which cannot be easily corrected by reward or punishment received at the end of the reasoning, need 'thoughts about thoughts', that is, some type of supervisory and monitoring process, to detect where errors in the reasoning have occurred. This suggestion on the adaptive value in evolution of such a higher-order linguistic thought process for multistep planning ahead, and correcting such plans, may also be different from earlier work. Put another way, this point is that credit assignment when reward or punishment are received is straightforward in a one layer network (in which the reinforcement can be used directly to correct nodes in error, or responses), but is very difficult in a multistep linguistic process executed once "on the fly". Very complex mappings in a multilayer network can be learned if hundreds of learning trials are provided. But once these complex mappings are learned, their success or failure in a new situation on a given trial cannot be evaluated and corrected by the network. Indeed, the complex mappings achieved by such networks (e.g. back-propagation nets) mean that after training they operate according to fixed rules, and are often impenetrable and inflexible. In contrast, to correct a multistep, single occasion, linguistically based plan or procedure, recall of the steps just made in the reasoning or planning, and perhaps related episodic material, needs to occur, so that the link in the chain which is most likely to be in error

can be identified. This may be part of the reason why there is a close relation between declarative memory systems, which can explicitly recall memories, and consciousness.

Some computer programs may have supervisory processes. Should these count as higher-order linguistic thought processes? My current response to this is that they should not, to the extent that they operate with fixed rules to correct the operation of a system which does not itself involve linguistic thoughts about symbols grounded semantically in the external world. If, on the other hand, it were possible to implement on a computer such a high-order linguistic thought supervisory correction process to correct first-order linguistic thoughts with symbols grounded in the real world, then this process would prima facie be conscious. If it were possible in a thought experiment to reproduce the neural connectivity and operation of a human brain on a computer, then prima facie it would also have the attributes of consciousness. It might continue to have those attributes for as long as power was applied to the system.

Another possible difference from earlier theories is that raw sensory feels are suggested to arise as a consequence of having a system that can think about its own thoughts. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution.

A property often attributed to consciousness is that it is unitary. The current theory would account for this by the limited syntactic capability of neuronal networks in the brain, which renders it difficult to implement more than a few syntactic bindings of symbols simultaneously (see McLeod et al., 1998; Rolls & Treves, 1997). This limitation makes it difficult to run several "streams of consciousness" simultaneously. In addition, given that a linguistic system can control behavioural output, several parallel streams might produce maladaptive behaviour (apparent as, e.g. indecision), and might be selected against. The close relation between, and the limited capacity of, both the stream of consciousness, and auditory–verbal short-term memory, may be that both implement the capacity for syntax in neural networks. Whether syntax in real neuronal networks is implemented by temporal binding (see von der Malsburg, 1990) is still an unresolved issue (see Rolls & Treves, 1997). (For example, the code can be read off from the end of the visual system without taking the temporal aspects of the neuronal firing into account, as described above; much of the information about which stimulus is shown is available in short times of 30–50 ms, and cortical neurons need fire for only this long during the identification of objects (Rolls & Tovee, 1994; Rolls et al., 1994b; Tovee & Rolls, 1995; Tovee et al., 1993) (these are rather short time windows for the expression of multiple separate populations of synchronized neurons); and oscillations, at least, are not an obvious property of neuronal firing in the primate temporal cortical visual areas involved in the representation of faces and objects (Tovee & Rolls, 1992).)

The current theory holds that consciousness arises by virtue of a system that can think linguistically about its own linguistic thoughts. The advantages for a system of being able to do this have been described, and this has been suggested as the reason why consciousness evolved. The evidence that consciousness arises by virtue of having a system that can perform higher-order linguistic processing is, however, and I think may remain, circumstantial. (Why must it feel like something when we are performing a certain type of information processing? The evidence described here suggests that it does feel like something when we are performing a certain type of information processing, but does not produce a strong reason for why it has to feel like something. It just does, when we are using this linguistic processing system capable of higher-order thoughts.) The evidence, summarized above, includes the points that we think of ourselves as conscious when, for example, we recall earlier events, compare them with current events, and plan many steps ahead. Evidence also comes from neurological cases, from, for example, split brain patients (who may confabulate conscious stories about what is happening in their other, non-language, hemisphere), and from cases such as frontal lobe patients who can tell one consciously what they should be doing, but nevertheless may be doing the opposite. (The force of this type of case is that much of our behaviour may normally be produced by routes about which we cannot verbalize, and are not conscious about.) This raises the issue of the causal role of consciousness. Does consciousness cause our behaviour?[2] The view that I currently hold is that the information processing which is related to consciousness (activity in a linguistic system capable of higher-order thoughts, and used for planning and correcting the operation of lower-order linguistic systems) can play a causal role in producing our behaviour (see Figure 4). It is, I postulate, a property of processing in this system (capable of higher-order thoughts) that it feels like something to be performing that type of processing. It is in this sense that I suggest that consciousness can act causally to influence our behaviour: consciousness is the property that occurs when a linguistic system is thinking about its lower-order thoughts. The hypothesis that it does feel like something when this processing is taking

[2] This raises the issue of the causal relation between mental events and neurophysiological events, part of the mind–body problem. My view is that the relation between mental events and neurophysiological events is similar (apart from the problem of consciousness) to the relation between the program running in a computer and the hardware in the computer. In a sense, the program causes the logic gates to move to the next state. This move causes the program to move to its next state. Effectively, we are looking at different levels of what is overall the operation of a *system*, and causality can usefully be understood as operating both within levels (causing one step of the program to move to the next), as well as between levels (e.g. software to hardware and vice versa).

place is at least to some extent testable: humans performing this type of higher-order linguistic processing, for example, recalling episodic memories and comparing them with current circumstances, who denied being conscious, would prima facie constitute evidence against the theory. Most humans would find it very implausible though to posit that they could be thinking about their own thoughts, and reflecting on their own thoughts, without being conscious. This type of processing does appear to be for most humans to be necessarily conscious.

Finally, I provide a short specification of what might have to be implemented in a neural network to implement conscious processing. First, a linguistic system, not necessarily verbal, but implementing syntax between symbols implemented in the environment would be needed. Then a higher-order thought system also implementing syntax and able to think about the representations in the first-order language system, and able to correct the reasoning in the first-order linguistic system in a flexible manner, would be needed. So my answer to the title of this paper is that consciousness can be implemented in neural networks (and that this is a topic worth discussing), but that the neural networks would have to implement the type of higher-order linguistic processing described in this paper.

REFERENCES

Abbott, L. A., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex, 6,* 498–505.

Armstrong, D. M., & Malcolm, N. (1984). *Consciousness and Causality*. Oxford: Blackwell.

Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *Journal of Computational Neurology, 306,* 554–575.

Burgess, N., Recce, M., & O'Keefe, J. (1994). A model of hippocampal function. *Neural Networks, 7,* 1065–1081.

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

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World.* Chicago: University of Chicago Press.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3,* 193–199.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Gaffan, D. (1994). Scene-specific memory for objects: a model of episodic memory impairment in monkeys with fornix transection. *Journal of Cognitive Neuroscience, 6,* 305–320.

Gazzaniga, M. S. (1988). Brain modularity: towards a philosophy of conscious experience. In A.J. Marcel & E. Bisiach (Eds.), *Consciousness in Contemporary Science* (Ch. 10, pp. 218–238). Oxford: Oxford University Press.

Gazzaniga, M. S. (1995). Consciousness and the cerebral hemispheres. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1392–1400). Cambridge, MA: MIT Press.

Gazzaniga, M. S., & LeDoux, J. (1978). *The Integrated Mind.* New York: Plenum.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society, Series B, 351,* 1445–1453.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neuroscience, 16,* 218–222.

Humphrey, N. K. (1980). Nature's psychologists. In B. D. Josephson & V. S. Ramachandran (Eds.), *Consciousness and the Physical World* (pp. 57–80). Oxford: Pergamon.

Linsker, E. (1986). From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences of the USA* **83**, 7508–7512, 8390–8394, 8779–8783.

Linsker, E. (1988). Self-organization in a perceptual network. *Computer, March 1988,* 105–117.

MacKay, D. J. C., & Miller, K. D. (1990). Analysis of Linsker's simulation of Hebbian Rules. *Neural Computation, 2,* 173–187.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Marcel, A. J. (1988). Phenomenal experience and functionalism. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in Contemporary Science* (pp. 121–158). Oxford: Oxford University Press.

Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Philosophical Transactions of the Royal Society, Series B, 351,* 1455–1462.

Rolls, E. T. (1989a). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne & W. O. Berry (Eds.), *Neural Models of Plasticity: Experimental and Theoretical Approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rolls, E. T. (1989b). The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron* (pp. 125–159). Wokingham, UK: Addison Wesley.

Rolls, E. T. (1989c). Functions of neuronal networks in the hippocampus and cerebral cortex in memory. In R. M. J. Cotterill (Ed.), *Models of Brain Function* (pp. 15–33). Cambridge: Cambridge University Press.

Rolls, E.T. (1989d). Information processing in the taste system of primates. *Journal of Experimental Biology, 146,* 141–164.

Rolls, E. T. (1990a). Functions of the primate hippocampus in spatial processing and memory. In D. S. Olton & R. P. Kesner (Eds.), *Neurobiology of Comparative Cognition* (pp. 339–362). Hillsdale, NJ: Lawrence Erlbaum.

Rolls, E.T. (1990b). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion, 4,* 161–190.

Rolls, E.T. (1991). Neural organisation of higher visual functions. *Current Opinion in Neurobiology, 1,* 274–278.

Rolls, E.T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society, 335,* 11–21.

Rolls, E. T. (1993). The neural control of feeding in primates. In D. A. Booth (Ed.), *Neurophysiology of Ingestion* (pp. 137–169). Oxford: Pergamon.

Rolls, E.T. (1994a). Neurophysiology and cognitive functions of the striatum. *Revue Neurologique (Paris), 150,* 648–660.

Rolls, E.T. (1994b). Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes, 33,* 113–138.

Rolls, E. T. (1995a). Central taste anatomy and neurophysiology. In R. L. Doty (Ed.), *Handbook of Olfaction and Gustation* (pp. 549–573). New York: Dekker.

Rolls, E.T. (1995b). Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research, 66,* 177–185.

Rolls, E. T. (1995c). A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1091–1106). Cambridge, MA: MIT Press.

Rolls, E.T. (1996a). A theory of hippocampal function in memory. *Hippocampus, 6,* 601–620.

Rolls, E. T. (1996b). The representation of space in the primate hippocampus, and episodic memory. In T. Ono, B. L. McNaughton, S. Molotchnikoff, E. T. Rolls, & H. Nishijo (Eds.), *Perception, Memory and Emotion: Frontier in Neuroscience* (pp. 567–579). Amsterdam: Elsevier.

Rolls, E.T. (1996c). The orbitofrontal cortex. *Philosophical Transactions of the Royal Society, Series B, 351,* 1433–1444.

Rolls, E. T. (1996d). The representation of space in the primate hippocampus, and its relation to memory. In K. Ishikawa, J. L. McGaugh & H. Sakata (Eds.), *Brain Processes and Memory* (pp. 203–227). Amsterdam: Elsevier.

Rolls, E. T. (1997). A neurophysiological and computational approach to the functions of the temporal lobe cortical visual areas in invariant object recognition. In L. Harris & M. Jenkin (Eds.), *Computational and Psychophysical Mechanisms of Visual Coding.* Cambridge: Cambridge University Press.

Rolls, E. T., & Johnstone, S. (1992). Neurophysiological analysis of striatal function. In G. Vallar, S. F. Cappa, & C. W. Wallesch (Eds.), *Neuropsychological Disorders Associated with Subcortical Lesions* (pp. 61–97). Oxford: Oxford University Press.

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

Rolls, E.T., & Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73,* 713–726.

Rolls, E. T., & Treves, A. (1997). *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rolls, E. T., Booth, M. C. A., & Treves, A. (1996). View-invariant representations of objects in the inferior temporal visual cortex. *Society for Neuroscience Abstracts, 22,* 760.5.

Rolls, E.T., Hornak, J., Wade, D., & McGrath, J. (1994a). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry, 57,* 1518–1524.

Rolls, E.T., Tovee, M.J., Purcell, D.G., Stewart, A.L., & Azzopardi, P. (1994b). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research, 101,* 474–484.

Rolls, E. T., Tovee, M., Treves, A., & Panzeri, S. (1997a). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience,* in press.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997b). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research, 114,* 149–162.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49,* 329–359.

Rosenthal, D. (1990). *A Theory of Consciousness* (ZIF Rep. 40). Bielefeld, Germany: Zentrum fur Interdisziplinaire Forschung.

Rosenthal, D. M. (1993). Thinking that one thinks. In M. Davies & G. W. Humphreys (Eds.), *Consciousness* (pp. 197–223). Oxford: Blackwell.

Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society, Series B, 351,* 1405–1411.

Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review, 99,* 195–231.

Tovee, M. J., & Rolls, E. T. (1992). Oscillatory activity is not evident in the primate temporal visual cortex with static stimuli. *Neuroreport, 3,* 369–372.

Tovee, M. J., & Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cognition, 2,* 35–58.

Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology, 70,* 640–654.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus, 2,* 189–199.

Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus, 4,* 374–391.

von der Malsburg, C. (1990). A neural architecture for the representation of scenes. In J. L. McGaugh, N. M. Weinberger and G. Lynch (Eds.), *Brain Organization and Memory: Cells, Systems and Circuits* (pp. 356–372). New York: Oxford University Press.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

Wallis, G., Rolls, E. T., & Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks, 2,* 1087–1090

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1241–1255, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00048-8

1997 SPECIAL ISSUE

From Brains to Neural Nets to Brains

ERICH HARTH

Physics Department, Syracuse University

(*Received* 10 *October* 1996; *accepted* 6 *January* 1997)

Abstract—*The paper traces theoretical work concerning the understanding and simulation of brain functions from early studies of artificial neural nets to present considerations of human consciousness. The emphasis is on work carried out since about 1963 at my laboratory in collaboration with my students. The discussion centers on sensory, especially visual, information processing, some of the cerebral mechanisms involved, and current approaches to an understanding of conscious perception. The sketchpad model, in which the ubiquitous feedback pathways in the brain play a dominant role, is described, together with a discussion of the meaning and applicability of scientific reductionism to the problem of consciousness. © 1997 Elsevier Science Ltd.*

Keywords—Neurodynamics, Receptive fields, Thalamocortical pathways, Reticular formation, Hillclimbing processes, Consciousness, Reductionism.

1. INTRODUCTION

I was asked to review my work and ideas for this Special Issue on Consciousness. My aim here will be to retrace my own circuitous route that took me from studying primitive structures of formal neurons to that hardest of hard questions, the nature of consciousness.

In the beginning, the problem of how mind arises from the physical brain seemed not too awesome a challenge to some of us. With enviable insouciance one theoretician had declared that the constellation of neurons active in a neural net—both its biological prototype and the electronic analogue—is quite simply … a thought! We have become more sophisticated both in our knowledge of the brain and in constructing powerful neural networks, but the question What, if anything, lies between the machine and human thought? has continued to be an unresolved issue.

The impetus for the many studies of neural nets in the late 1950s and early 1960s came mostly from a seminal paper by McCulloch & Pitts (1943) in which neurons were treated as simple threshold elements with all-or-none response. The authors showed that networks of such *formal neurons* were capable of carrying out any computation whose logical steps could be defined, and were thus functionally equivalent to universal Turing machines. The McCulloch–Pitts paper raised the hope that all manner of brain functions could be understood—and eventually synthesized—by systems composed of such primitive logical elements, relying only on the sheer mass of the net and, perhaps, the delicacy of its interconnections.

The ability to undergo permanent or semi-permanent changes is a further essential ingredient in a system that is to mimic the brain. The psychologist Donald Hebb (1949) suggested that this plasticity resided in the modifiable strengths of neuronal interconnections, the synaptic couplings between neurons. He further assumed that these synaptic changes, called *synaptic facilitation*, result from strictly local effects, that is, from the coincidence of a presynaptic action potential and a postsynaptic excitation of the target cell.

It remains to endow the net of real or simulated neurons with structure. Here, the controversy raged for a long time, whether the biological neural net was hard-wired in the sense that the entire connectivity was both determined and essential to its functioning, or whether genetic determination left some of the details of the circuitry to chance. The extreme assumption of a single, amorphous, randomly connected net was clearly in violation of biological facts, but was nevertheless often taken as the starting point of theoretical discussions and the study of artificial neural networks. At the other extreme, the assertion of a rigidly deterministic circuitry in which all detail is specified, has been contradicted by numerous brain studies.

Requests for reprints should be sent to: Erich Harth, Department of Physics, Syracuse University, Syracuse, NY 13244-1130, USA; E-mail: harth@suhep.phy.syr.edu.

There is, finally, the question what we wish the net to accomplish, or what output to expect, and in what code. Here, again, Hebb has shown the way. Meaning is to be attached not to the firing of single neurons (sometimes called grandmother cells), but to subpopulations of the neural net that Hebb termed *cell assemblies*. A cell assembly may arise through synaptic facilitations and is capable of sustaining neural activity within itself through reverberating circuits.

Hebb's ideas of synaptic facilitation and cell assemblies have been the cornerstone of practically all thinking about the brain's ability to record, associate, and recall sensory events, as well as the chief ingredient in learning machines we call neural nets. They appear over and over in most theories of brain functions from the most primitive to the elusive phenomena of human thought and consciousness.

Together with the concept of McCulloch–Pitts neurons, the Hebb rules make it appear that brain functions are mechanisms that are conceptually simple. This has given rise to bold attempts to simulate different aspects of the brain by use of artificial nets of formal neurons, and to a general euphoria about their future potential. A word of caution is in order. It has become increasingly clear just how ''impoverished'' artificial neurons are when compared to their biological archetypes. The proliferation of recognized neurotransmitters, neuromodulators, and neurohormones since about the mid 1970s, all of which affect the behavior of neurons but take their information from widely different sources, has made us aware of the enormous functional complexity that exists already at the level of individual cells.

2. SIMPLE NEURAL NETS AND BRAIN FUNCTIONS

The late 1050s saw a number of papers (Allanson, 1956; Rochester et al., 1956; Beurle, 1959) investigating the dynamics of populations of primitive neuron-like elements, and incorporating the ideas proposed by McCulloch & Pitts (1943) and Hebb (1949).

In 1958 Frank Rosenblatt (1958) initiated a new approach in neural net technology. Rather than studying the dynamic properties of multiply connected networks of many formal neurons, he conceived a circuit consisting of an array of input units, connected through a set of intermediate neurons to an array of output units. The inputs and outputs are both binary words. He called this device a perceptron. Initial coupling strengths are chosen arbitrarily, and, for a particular task, a truth table prescribes the ''correct'' output for every possible input. In a training procedure many different inputs are presented to the perceptron. A simple training rule specifies the changes to be made to the coupling strengths according to whether the net gave a correct or incorrect response. After training, the perceptron will in general produce correct answers a large fraction of the time.

Limitations of the original elementary perceptrons were pointed out by Minsky & Papert (1969); these were largely overcome by adding one or more layers of hidden units to the device. The simple reinforcement rules for elementary perceptrons could no longer be applied since the role of the hidden units in forming a particular output is not readily ascertained. A procedure called back propagation (Rumelhart et al., 1986) solves this problem, but is time-consuming and biologically implausible. Neither the elementary perceptron, nor its more sophisticated hidden-layer successor, contained neural feedback loops; hence no reverberatory activity.

2.1. Netlets

My own contribution to the field began with a paper presented in 1965 at a conference on Automata Theory held in Ravello (Harth, 1966). I carried out computer simulations of a randomly connected net of 100 formal neurons whose structure was given by a 'sparse' 100×100 asymmetric coupling matrix. Before learning, the net was amorphous and isotropic except for statistical fluctuations. In addition, 25 input units were randomly and diffusely connected into the net.

Stimuli were defined as 25-five digit binary words. The resulting state of the net—a 100-bit word—represented the response. The application of Hebbian learning rules brought out two related results. When the net was trained on the simultaneous presentation of two different stimuli, then the subsequent presentation of just one of them would produce a response that contained elements of the responses to both stimuli (association). Similarly, the presentation of parts of a stimulus would tend to produce a response closer to that of the previously learned entire stimulus (pattern completion).

2.2. Associations and Reflexes

Harth & Edgar (1967) extended this study to association nets of 1000 neurons. These were defined by such parameters as thresholds, average number of efferent junctions per neuron, average coupling coefficients, and ratio of inhibitory to excitatory neurons. Although containing many loops, reverberatory activities were kept down by sparse connectivity matrices and high firing thresholds. The response to a given input vector is the totality of neurons that reach firing level following the stimulus. We demonstrated that, upon applying rules of synaptic facilitation, systems of this kind are able to perform a remarkable variety of ''cognitive'' tasks such as simple conditioned reflexes, dual-conditioned reflexes, bidirectional, and reverse conditioned reflexes, as defined by Asratian (1965).

We assumed that subsets of the network acted as Hebbian cell assemblies with fixed and arbitrarily assigned meanings. Associations between different cell assemblies were accomplished by strengthening their

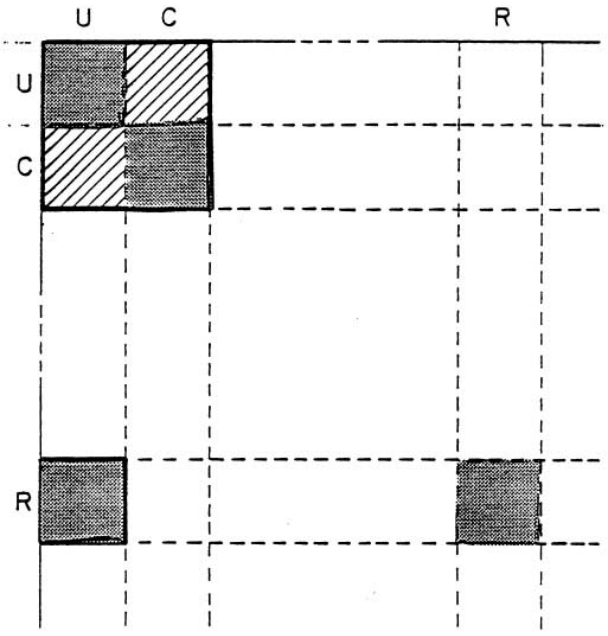


FIGURE 1. Matrix of coupling coefficients in simulation of a conditioned reflex. Each small square represents a *netlet* of about 100 neurons. *U*—unconditioned stimulus; *C*—conditioned stimulus: *R*—unconditioned response. Shaded areas denote domains of enhanced coupling coefficients within netlets *U*, *C*, and *R*, and the initially strong connections from *U* to *R*. Cross-hatched areas are enhanced due to learning.

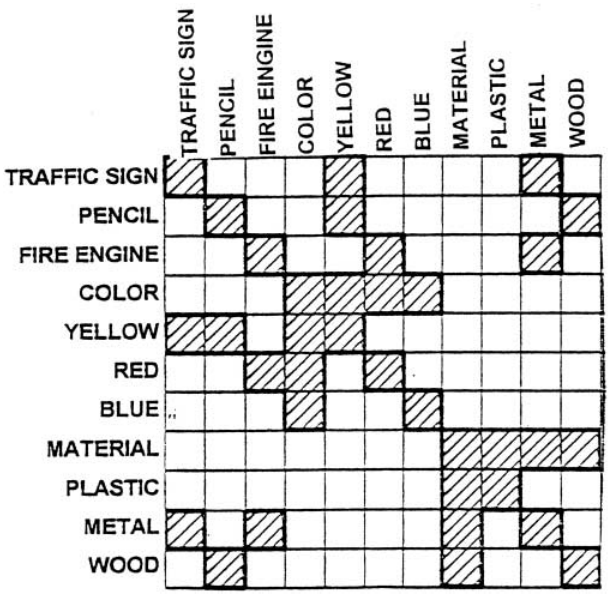


FIGURE 2. Connectivity matrix (defined as in Figure 1) in a computer simulation experiment. The 11 netlets are assigned meanings as indicated, and cross-hatched areas are enhanced.

interconnections. The previously observed effect of pattern completion means that the association of an assembly **A** with an assembly **B** will trigger a portion of **B** whenever **A** alone is activated, and vice versa.

In this approach the simple conditioned reflex is characterized schematically in Figure 1, which shows a part of the matrix of coupling coefficients. The domains *U*, *C*, and *R* are cell assemblies representing the unconditioned and conditioned stimuli and the response, respectively. Here, each small square in Figure 1 is an $n \times n$ matrix of coupling strengths, where n is the number of neurons in the cell assembly. Before training, the only inhomogeneities in the net are the shaded area indicating enhanced coupling coefficients within *U*, *C*, and *R*, and from *U* to *R*, causing activity in *R* whenever *U* is triggered, but no activity in *R* when *C* is active. The cross-hatched areas represent synaptic facilitation due to learning, in this case the frequent coincidences of stimuli *U* and *C*. Thus, triggering the conditioned stimulus *C* will cause activity to extend into *U* and from there into *R*.

Figure 2 is a schematic diagram of the connectivity matrix in a classification task. Eleven cell assemblies are shown here, representing different objects and properties as labeled. The initially random, homogeneous and isotropic connectivity matrix is shown after associations, i.e., enhancement of couplings between netlets. The resulting matrix is seen to be macroscopically symmetrical, but microscopically asymmetric. Cross-hatched areas are domains of enhanced synaptic strengths resulting from binary associations: FIRE ENGINE with RED, FIRE ENGINE with METAL, etc. The net is thus instructed that fire engines are red and of metal, that pencils are yellow and of wood, traffic signs yellow and of metal, that red, yellow, and blue are colors and plastic, wood and metal materials. With properly chosen parameters of thresholds and average normal and enhanced synaptic strengths, it will take more than one cell assembly to cause significant activity in an associated assembly. Taking advantage of the non-linearity of responses, the diagram in Figure 2 shows that the net will respond with RED when the input is FIRE ENGINE plus COLOR, but shows weak activities in the other assemblies. Similarly, the question ‘‘What is the material of pencils?’’ is answered by WOOD, ‘‘What is yellow and made of metal?’’ by TRAFFIC SIGN, and so on.

In a computer simulation, each of the cell assemblies was made up of 75 formal neurons. The question ‘‘What is the color of traffis signs?’’ produced the activities shown in Table 1. The answer is given by the strong activity of 16 out of 75 in the YELLOW cell assembly.

2.3. Neurodynamics

The next two papers (Harth et al., 1970b; Anninos et al., 1970) examine the dynamics of structured neural nets made up of randomly connected subsystems that are non-randomly connected to one another. We introduced here the concept of randomness-in-the-small and structure-in-the-large (Anninos et al., 1970). For the random subsystems we chose the term netlets. In a sense, the cell assemblies arising in the originally homogeneous

TABLE 1
Cell assembly responses to ''What is the color of traffic signs?''

Traffic sign	Pencil	Fire eng.	Color	Yellow	Red	Blue	Material	Plastic	Wood	Metal
75	**3**	0	**75**	**16**	2	3	1	0	6	0

connectivity matrix (Figs 1 and 2) become netlets in the trained net.

We classified netlets according to their excitability, that is their ability to sustain activity in the absence of input. The parameters defining a netlet were, among others, fraction of inhibitory neurons, thresholds, and average coupling strengths. In the formalism used the results were independent of the total number of ''neurons'' in the netlet, except for the size of the statistical fluctuations. Dynamics of these structures were obtained from statistical arguments as well as by computer simulations, some of which are shown in Figure 3. Here, the abscissa α_n refers to the fraction of neurons active in the nth iteration, and the ordinate is the activity at $n+1$. (The drop-off of the curves as α_n approaches 1.0 is due to an assumed refractory period.) Thus, a very high initial activity will find most neurons refractory in the next iteration, and the activity in the net will soon terminate. The curve labeled $\eta = 2$ (a measure of the netlet's excitability) shows two stationary points A and B (in addition to the one at the origin). We deduced that this gives rise to a hysteresis effect: a momentary input that triggers more than about 15% of the neurons in the netlet causes sustained activity at the level of point B, that is with around 30% of the neurons active in each iteration.

The dynamics were extended to the case of netlets having a steady excitatory or inhibitory input from the outside. Points O, A, and B are now functions of the input as shown in the schematic *phase diagram* of Figure 4. For changing input levels the net now exhibits *hysteresis*, with the arrows in Figure 4 indicating the spontaneous shift in activity level. For a combination of excitatory and inhibitory inputs, stationary states form a surface in three dimensions (Figure 5).

The calculations and simulations were extended (Wong & Harth, 1973) to netlets of different temporal characteristics describable by finite difference equations of order two and greater. This formalism can take into account any combination of refractory periods, summation times, and effective delays.

The dynamic characteristics here described were simulated for netlets of 1000 neurons, but turned out to be robust down to very small populations. Shaw et al. (1982) have shown that small netlets can act as functional units in the central nervous system, and that the cooperative phenomena described above were reliably performed with net sizes down to about 30 neurons.

With the approach of randomness-in-the-small and structure-in-the-large, neural systems can be constructed to perform specific tasks. Harth et al. (1975) and

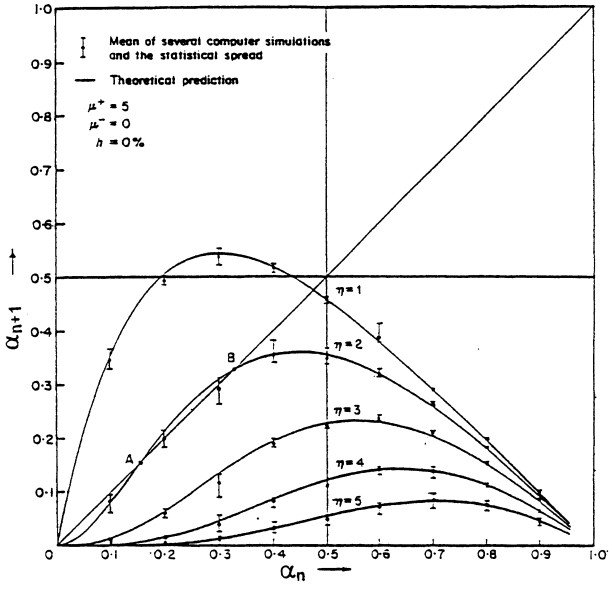


FIGURE 3. Dynamics of neural netlet, showing activity at time $n+1$ as function of activity at n. Points are results of repeated computer simulations and solid lines are computed from a finite difference equation. Different curves are for different excitabilities of the netlets. Points *O*, *A*, and *B* are stationary points (for details, see Anninos et al., 1970).

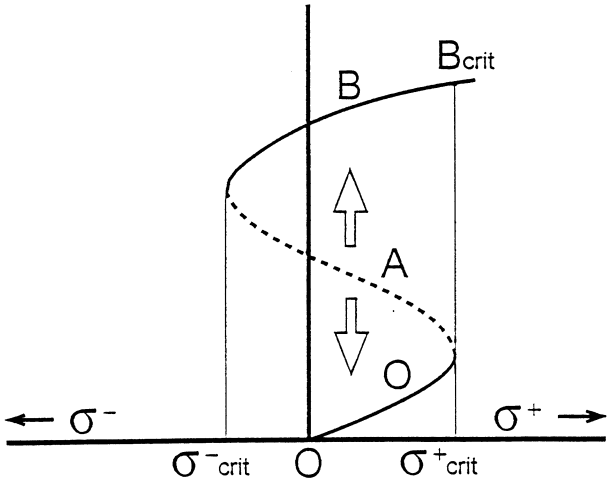


FIGURE 4. Hysteresis of netlets with steady input σ. The unstable stationary state *A* is indicated by the dotted line. It is seen that, when excitatory input exceeds σ^+_{crit}, activity will rapidly increase to B_{crit}. The input will now have to be reduced to the negative (inhibitory) value σ^-_{crit}, before the activity is extinguished. Two netlets of lower excitability are shown for comparison.

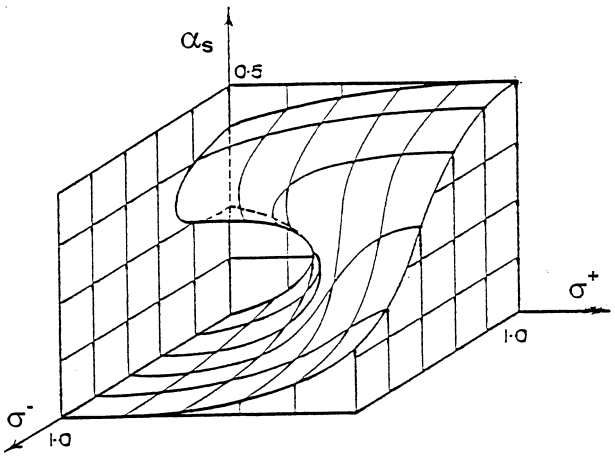


FIGURE 5. Surface of stationary netlet activities in the presence of excitatory (σ^+) and inhibitory (σ^-) input.

Csermely et al. (1973) attempted to account for the oscillatory behavior in the escape motion of the mollusk *Tritonia gilberti*. In the anatomical studies reported by Willows (1967), the system is composed of four small neural populations (Figure 6), where TFN refers to trigger group neurons, GEN to general excitor neurons, and DFN and VFN to dorsal and ventral flexion neurons, respectively. In the mechanism proposed by Willows, the trigger neurons excite the GEN pool causing alternating activities in DFN and VFN (Figure 7).

In our simulation of *T. gilberti*'s motion, we have taken the connectivity proposed by Willows, replacing each of his neuron pools with a random netlet. No further details of connectivity are thus assumed. This contrasts with the approach of Selverston (1980), in which he proposes to account for a similar problem (the motion in the lobster stomatogastric ganglion) by invoking all details of neural connectivity. I discussed the fundamental differences in the two approaches in a commentary to Selverston's paper (Harth, 1980).

2.4. Visual Perception

Vision is the most extensively studied of the senses, and, in humans, the most valuable. A much simplified

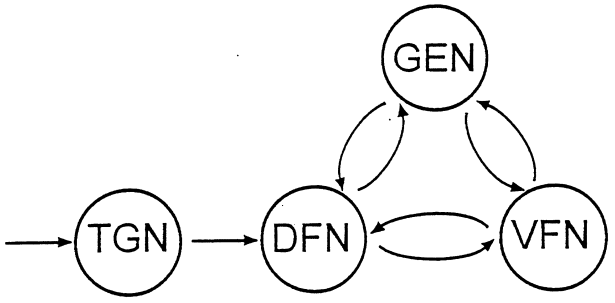


FIGURE 6. Connectivity between neuron populations in *Tritonia gilberti,* according to Willows (1967).

FIGURE 7. Activities in DFN and VFN populations of swimming *T. gilberti* (after Willows, 1967).

schematic diagram of the visual pathway is shown in Figure 8. Here, visual information gathered by the retina is first conveyed to neurons in the lateral geniculate nucleus (LGN), a structure in the thalamus. From there about a million neural fibers go to a region in the cortex called V1, the first cortical area devoted to vision. A prominent feature—but one rarely discussed by physiologists—is the fact that perhaps 10 times as many fibers descend from V1 and send information back to the LGN. From V1 the visual information travels along several parallel pathways to centers called V2, V3, V4, and many others. Some 30 different visual centers have been identified in monkey cerebral cortex (Felleman & Van Essen, 1991). Again, practically all connections between different cortical centers are reciprocal.

The ubiquity of such loops of connectivity in the central nervous system blurs the distinction between higher and lower cortical centers. There is no strict hierarchy of sensory processing. The ''higher'' centers receive information from the ''lower'' centers, but these sources are modified, augmented, and perhaps censored by the ''higher'' centers. The prominence of these return pathways, especially the massive fiber bundle from V1 to LGN, has been one of the great puzzles of cerebral architecture. It appears that the cortex is more bent on introspection and confabulation than on forming an unbiased view of the outside world.

2.5. Receptive Fields

To understand the functioning of the visual system, we would like to know what patterns of neural activity in the brain follow different visual stimuli. This has been until very recently extremely difficult to determine, and so, experimenters have often chosen to investigate a related problem: What is the visual stimulus pattern to which a given single neuron exhibits its strongest response? The stimulus, usually obtained by trial and error, is called the receptive field of that cell.

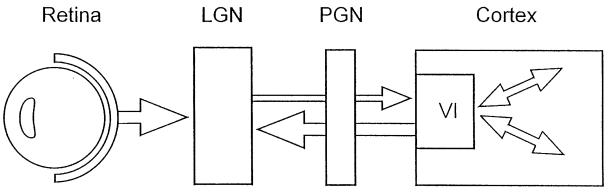


FIGURE 8. Schematic of visual pathways from retina to visual cortex.

In a slight departure from the netlet concept, Finette et al. (1978) applied the dynamic principles we had developed for homogeneous and isotropic netlets, to structures with random but anisotropic connectivity. The attempt here was to account for the transition from small, essentially circular, visual receptive fields in mammalian retina and lateral geniculate nuclei (LGN) to the elongated fields observed by Hubel & Wiesel (1962) in area V1 of visual cortex. It is customary to assume that these elongated fields are the result of multiple inputs to V1 cells from LGN cells whose fields lie along an oriented line.

By contrast, Finette et al. (1978) found that locally anisotropic connectivity within V1 can give rise to resonant activity having the shapes of Hubel and Wiesel fields. Although attempts to detect such anisotropies in rat visual cortex (Paldino & Harth, 1977) remain inconclusive, the ease with which these elongated responses arise, and the comparative simplicity of the neural circuitry required, make this model still a plausible candidate for explaining cortical receptive fields.

The study of receptive fields offers an interesting exercise in solving the *problem of the inverse*. One is attempting to determine a visual pattern for which a particular neuron (that is being monitored with a microelectrode) shows the highest response. Clearly, it is impractical to try all possible stimuli, hence the answer involves some reasonable guesses on the part of the experimenter. There is never a certainty, however, that some modification of the input may not have produced an even higher response.

2.6. Alopex

In the early 1970s work began in my laboratory on a procedure for determining visual receptive fields that would eliminate the guesswork and biases of the experimenter. Receptive fields were to be mapped by a *hill-climbing* procedure in which the cell under observation generates its own receptive field. We accomplished this by using an algorithm that maximizes the cell response (Harth & Tzanakou, 1974). In this procedure, called Alopex, an animal viewed a computer-generated visual pattern displayed on a CRT while a neuron in the visual pathway was being monitored (Figure 9). The Alopex procedure changes all pixel intensities on the CRT before each successive presentation depending on the correlation (or anticorrelation) between the preceding changes in pixel intensity and cell response. The algorithm can be represented by the simple equation

$$x^{(n)} = x^{(n-1)} + g^{(n)} + c[x^{(n-1)} - x^{(n-2)}]\cdot[R^{(n-1)} - R^{(n-2)}] \quad (1)$$

In eqn (1) $x^{(n)}$ is the intensity of the ith pixel in the stimulus array in the nth iteration, $g^{(n)}$ a random variable (noise), c an adjustable parameter, and $R^{(n)}$ the strength

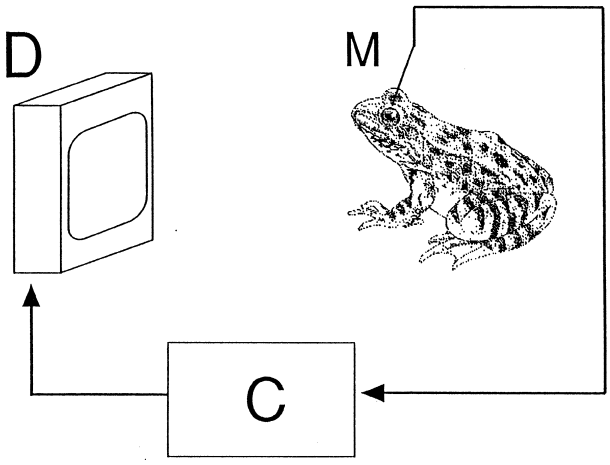


FIGURE 9. Determination of visual receptive field by Alopex method. *D*—CRT display; *M*—microelectrode; *C*—computer.

of the cell's response in the nth iteration. R is thus a scalar quantity that is to be maximized in this process. This is often called a *cost function*.

The method was applied successfully to frog retinal ganglion cells (Tzanakou et al., 1979) and cat cortical cells (Micheli-Tzanakou, 1984). In Figure 10 we show the ''optimal'' pattern obtained by the Alopex procedure together with the receptive field arrived at by conventional methods for a neuron in V1 of cat.

The ease with which the Alopex algorithm reversed the sensory process in our studies of visual receptive fields, spawned another idea (Harth, 1976). Could it be that similar processes take place within the brain, whereby feature-detecting neural circuits in the cortex—such as the ''face detectors'' first reported by Gross et al. (1969)—enhance or generate at more peripheral cerebral locations, those patterns to which they are evidently tuned? If such a mechanism existed, it would clearly add another dimension to visual perception: the visual stimuli would no longer be transported in one direction only, along the brain's conveyor belts, being progressively fragmented and having various features extracted in what I have called the *slaughter-house paradigm* (Harth, 1993). Instead, visual perception

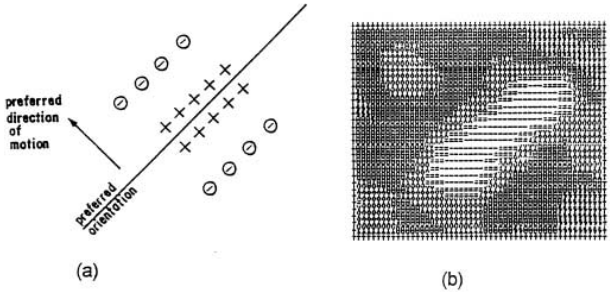


FIGURE 10. Cortical visual receptive field of cat cortical neuron determined by standard procedures (a), and by the Alopex method (b) (from Micheli-Tzanakou, 1984).

would now appear as a self-referent process in which the incoming stimulus is continuously recreated and reanalyzed by higher cortical centers.

2.7. Feedback Pathways

Neural connections in the brain are generally reciprocal. If a region *A* sends fibers to *B*, then, very likely, there will be return paths from *B* to *A*. Thus, we speak not only of *afferent* fibers, but also of descending, or *corticofugal* fibers. The massive return paths from cortical visual area V1 to the visual thalamic relay nucleus LGN has been known to neuroanatomists for nearly a century. Livingston (1958) quotes Brower (1933) who commented: ''We accept that there is also a centrifugal side in the process of sensation, of vision, of hearing, and so on. I believe that a further analysis of the descending tracts to pure sensory centers will also help physiologists and psychologists to understand some of their experiences''.

Through all the years since then this goal has been elusive. There have been many speculations concerning the function of these return paths, but little consensus. Rothblatt & Pribram (1972) speculated that there should be pattern-specific top-down modification of sensory inputs, and were among the first to raise the question by what neural mechanism this could be accomplished.

Apart from the cortico-thalamic tracts, it has been recognized for a long time that brainstem structures also communicate with all afferent sensory pathways. In an early symposium on the brainstem reticular formation, Livingston (1958) talks about ''the reticular formation … rippling its fingers in some kind of play of … perceptive … composition''.

In more recent work, Yingling & Skinner (1977) report that neurons of the nucleus reticularis thalami (NRT), which is interposed between thalamic sensory relay nuclei and sensory cortex, are inhibitory on thalamic relay cells. This inhibitory feedback is topographically organized between NRT and thalamic cells, but the NRT cells, in turn, receive ''*non-selective, generalized*'' inhibitory input (Steriade et al., 1986) from the mesencephalic reticular formation (MRF).

Thus, if cortico-thalamic pathways and thalamic input from the brainstem affect the pattern of incoming sensory stimuli—and it is difficult to avoid this conclusion—then these feedbacks are very likely to be feature-specific.

This reciprocity of connections in sensory pathways is almost universal and has given rise to many speculations. Von Holst & Mittelstaedt (1950) speak of a *principle of reafference*, by which the brain may generate its own input without sensory stimulation. Edelman (1992) points to the *re-entry* of information passed back and forth between sensory *maps*. The resulting reverberations are both conservative and creative, in the sense that they may preserve a stimulus long enough to be contemplated by ''higher'' cortical centers, and perhaps committed to memory, and creative in that neural activity patterns *evolve* and *link* sensory information to past experiences in novel combinations. Edelman believes that the modulation of images through extensive cross-talk between sensory centers bestows a kind of unity on sensory experience.

The problem of preserving spatial mappings was discussed by Harth et al. (1970a) in connection with lateral inhibition. This phenomenon is as ubiquitous in sensory systems as the reciprocity between maps. What effect does lateral inhibition have on the information that is passed back and forth between two maps, or passed sequentially from one map to the next?

It turned out that under repetitive mapping with a Gaussian spread function, a sharply defined spot will rapidly blur and melt into background noise.

In Figure 11(a) we observe the repeated mappings of an initial delta function by a Gaussian spread function that results in the expected progressive deterioration of the signal. In Figure 11(b) the spread function is altered by adding a negative Gaussian of smaller amplitude but greater width, which mimics the well known characteristic spread function for lateral inhibition. This results in a remarkable phenomenon. After some initial changes, the signal stabilizes and becomes essentially insensitive to further blurring.

A graphic confirmation of this principle is shown in Figure 12. Here, a black spot is imaged repeatedly by a photocopier whose spread function also showed an inhibitory surround. The succession of mappings at first introduced increasing amounts of noise. But, instead of eventually overwhelming the signal, the pattern stabilizes and soon becomes immune to further corruption. The information is now carried by a stable pattern of noise.

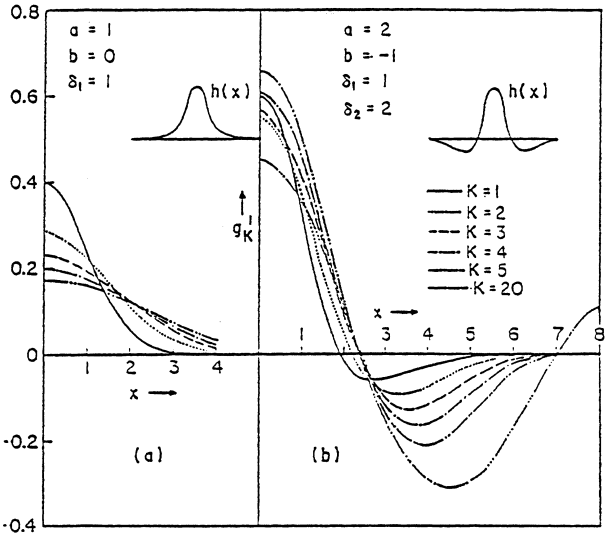


FIGURE 11. Successive mappings of a delta function through (a) a Gaussian point spread function, (b) through a function with positive Gaussian center and negative Gaussian surround.

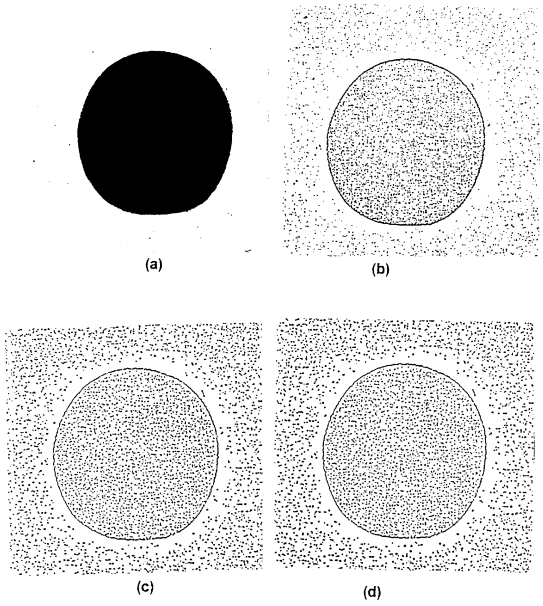


FIGURE 12. Successive mappings by a photocopier. (a) original; (b) 4th copy; (c) 13th copy; (d) 15th copy.

In a subsequent paper (Harth & Pertile, 1972), we showed that there exists a class of *ideal* mapping functions that would preserve information for an infinite number of successive mappings, and that one of these has the form $(\sin kx)/x$, which is similar to the spread function in lateral inhibition except for having further negative and positive peaks of diminishing amplitudes.

I believe that this form of signal stabilization due to lateral inhibition plays a role in most any sensory system, together with mechanisms for altering the signal in significant ways, as proposed for example by Edelman (1992). The first must exist for the second to function.

2.8. Mental Images

Humans are able to construct visual images in the absence of visual input. We speak of *mental images*, and the controversy has been whether these images have the spatio-temporal characteristics of true visual stimuli, or whether they are *neural codes* appearing at high level cortical areas. In the first alternative, we would expect them to appear at more peripheral regions in the visual pathway where neural representations are still retinotopically organized.

This appears to be the case. In a series of PET scan studies Kosslyn et al. (1993) have shown that mental imagery is accompanied by marked activity in V1, and that the spatial extent of the activity depends on the size of the object imagined. They propose that mental imagery possesses characteristics resembling images of real objects, and that these images form the inputs to higher cortical levels as in visual perception. In an analogous description, Llinás et al. (1994) call the brain a ''closed system capable of self-generated activity'' and a ''reality-emulating system'', though they take the emulation to occur at higher cortical levels as a result of temporal scanning mechanisms. This point is admittedly still somewhat in dispute, since some data suggest that V1 activity is not essential for conscious cognition.

3. THE SKETCHPAD MODEL

The ideas presented here are taken chiefly from my recent work (Harth, 1995, 1996a), but incorporate many elements of the earlier papers.

The notion of internal sketchpads has its origin in the observation that thought processes, especially creative ones, often involve a cyclic interplay between cerebral and external events: the artist sketches before putting paint to canvas, the composer needs a piano to compose his music, and all of us occasionally talk to ourselves.

Why do we bother to externalize for our own perusal something we already know? The answer is, I believe, not hard to find. Our nervous system finds it difficult to store complex images of, say, a painting or a symphony, in a way that would allow us to contemplate and judge the work in its entirety. Hence, the creation of such works requires temporary external storage of the nascent idea, sensory access to the image, and the ability to modify, elaborate and to perfect. This can be described as a hill-climbing procedure, in which the changes in the external image, or artifact, are guided by internal aesthetic or other criteria. Such processes are self-referent and involve positive feedback.

I have proposed that *mental* images, though more restrictive in their capacity to hold information, serve functions that are very similar to the artist's sketchpad or the composer's piano.

I came across this same idea in a most unlikely source. Trân Duc Thao (1984) quotes from Karl Marx's *Das Kapital*:

> … what distinguishes the worst architect from the best of bees is this, that the architect raises his structure in the imagination before he erects it in reality.

If, indeed, top-down mechanisms in the brain are to generate ''peripheral'' neural activity patterns that have the spatio-temporal characteristics of visual stimuli, then we may suppose that they can function as trial patterns, to be viewed and further developed, very much like patterns we put on a sketchpad.

In developing this idea further, I proposed (Harth, 1976) that a feedback to the LGN act as a *cost function* that guides a pattern-generating optimization process. I have taken the example of the Alopex algorithm and our receptive field studies (Tzanakou et al., 1979) as general guide for this model. Accordingly, the feedback would have to be a *global* function, i.e., one that carries a simple *figure of merit* expressing the degree to which the

peripheral pattern satisfies some internal requirements. In the limit, a simple scalar quantity would suffice. I believe this assumption of a *global* re-entry into the sensory pathway is somewhat related to Edelman's (1992) concept of *global mapping*.

In this first attempt (Harth, 1976), I viewed the LGN as the site of the internal sketchpad, and the cortico-thalamic feedback pathways as the carriers of the cost function.

There are several difficulties with this scheme. Rather than carrying *global* information, the cortico-thalamic pathways are specific, connecting cells in V1 with their retinotopic equivalents in the LGN (Yingling & Skinner, 1977). These are more likely to cause reverberations, i.e., a temporary stabilization of the stimulus pattern, while the ''nonselective, generalized'' input from MRF to NRT has more of the *global qualities* of a cost function.

3.1. Brainstem Control of Sensory Messages

The effect of brainstem neurons on thalamocortical transmission was considered in a revised sketchpad model (Harth & Unnikrishnan, 1985; Harth et al., 1986, 1987, 1990).

The diagram in Figure 13 illustrates the anatomical features on which the model is based. A *map* of thalamic relay cells (*G*) is connected reciprocally to visual cortex V1. These connections are excitatory, and send excitatory collateral inputs to PGN cells (*P*). These are inhibitory on thalamic relay cells and are, in turn, inhibited by fibers from the MRF. The *global* character of the latter is here represented by a single ascending line.

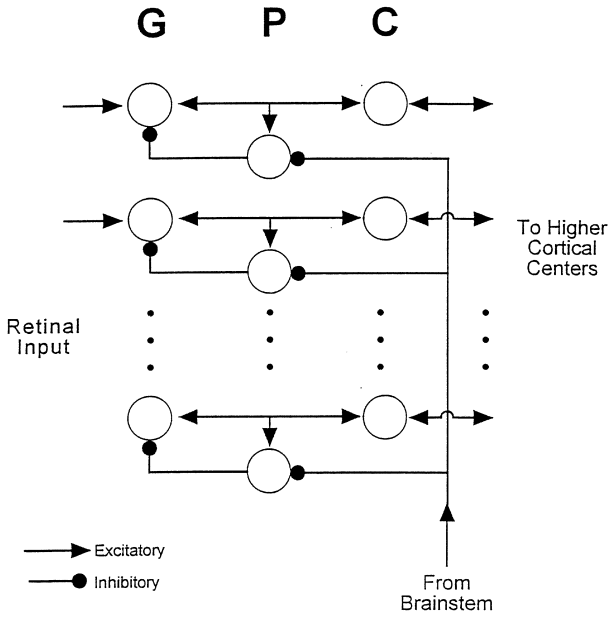


FIGURE 13. Schematic of connectivity in visual pathway, showing thalamocortical loop with collateral inputs to PGN, and PGN inhibition of LGN relay cells. *G*—thalamic relay neurons; *P*—neurons in PGN; *C*—cortical cells.

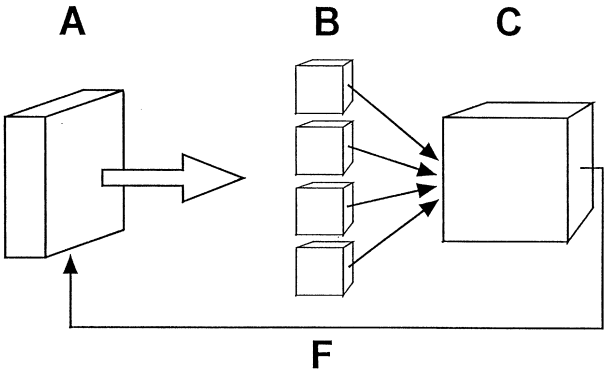


FIGURE 14. Schematic of computer simulation of *sketchpad* mechanism. *A*—hillclimbing algorithm modifying a peripheral pattern; *B*—array of feature analyzers producing responses $R_1 - R_4$; *C*—a non-linear integrator producing *a cost function F*.

On the basis of this diagram, one would expect that a retinal input pattern would be quickly extinguished by the PGN cells, except when strong brainstem input inhibits the PGN cells. With the addition of random noise and the iterative character made possible by the retinotopic feedback from V1, this circuit is capable of carrying out the optimization algorithm defined by eqn (1).

3.2. Sketchpad Simulations

I present here one of the many results of computer simulations reported in previous publications (e.g., Harth, 1995). In Figure 14, box *A* represents an optimizer that carries out an algorithm such as the one defined by eqn (1), using the scalar variable *F* as *cost function*, and displaying the resulting pattern in an iterative fashion. These patterns are *viewed* by four feature analyzers *B*, whose responses R_1 to R_4 reflect the similarity between the observed pattern and the fixed template assigned to

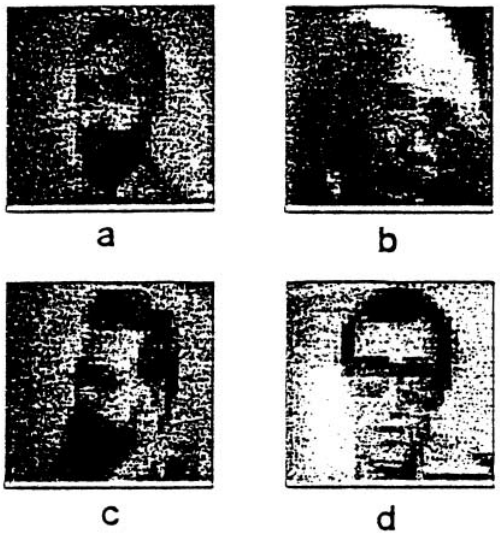


FIGURE 15. Templates used in sketchpad simulation.

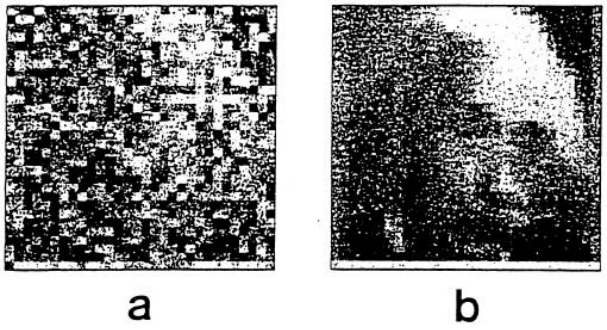


FIGURE 16. (a) Initial stimulus pattern; (b) pattern after hillclimbing convergence.

that analyzer. The analyzer responses are summed nonlinearly in C to make up the cost function F.

Figure 15 shows four such templates, and Figure 16(a) an initial state chosen for the system. This pattern is almost random; it was obtained by superimposing random noise on the Einstein template. With the application of the hillclimbing algorithm, the pattern changed gradually and eventually became almost identical with the Einstein template [Figure 16(b)].

Figure 17 shows the evolution of the analyzer responses and the combined response F. Note that the response R_2 of the Einstein analyzer is initially higher than that of the others, and begins to increase sharply after an initial period of *uncertainty*. The process speeds up as the cost function F becomes dominated almost exclusively by R_2.

An optimization algorithm of the type shown here is the most general way in which peripheral pattern generation can be achieved in the sensory system by top-down control. It is also the one requiring the least complexity of the necessary neural circuitry. The known connectivity between LGN, PGN, V1, and BRF (Figure 13) appears to be eminently suited to carry out the necessary algorithm, but similar processes may take place at higher visual areas, as well as in other sensory modalities.

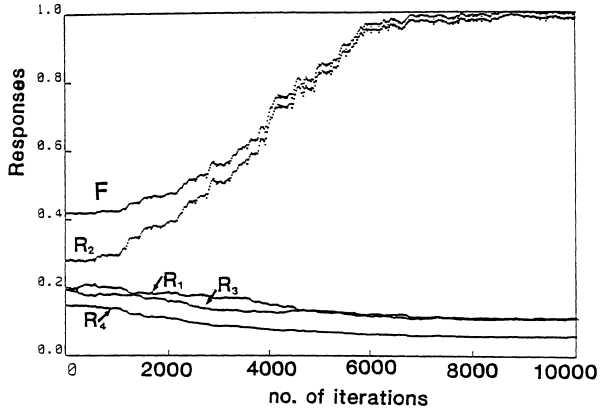


FIGURE 17. Responses $R_1 - R_4$ of feature analyzers and *cost function F* in sketchpad simulation.

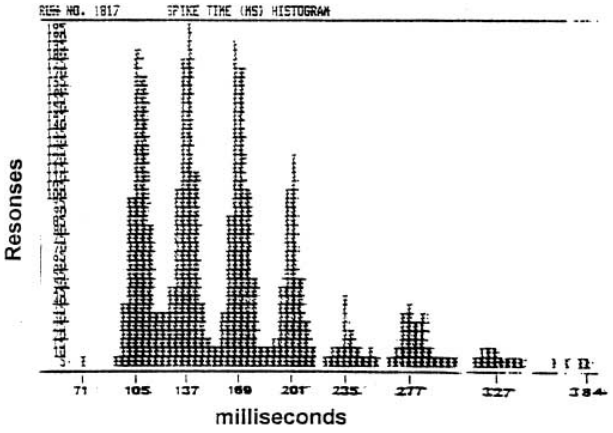


FIGURE 18. Periodic (ca. 30 Hz) post-stimulus histogram in frog retinal ganglion cell (from Stiles et al., 1985).

3.3. Summary of Model

Mental processes are appropriately called *reflective*, and concomitant neural activity is cyclic, self-referent. What distinguishes the thinking brain from computers, and from most artificial neural networks, is the fact that in thought processes there are no final, or *output states*.

In summarizing the model, let me begin, arbitrarily, by considering the extensive, scattered activities in cortical areas, including sensory analyzers, association areas, and ''higher'' areas such as the *working memory* (Baddely, 1986), or the *global workspace* (Newman & Baars, 1993). These activities are in part elicited by current sensory stimuli, in part due to entirely endogenous processes, including neural noise. Chaotic processes may well be involved, allowing minute fluctuations to cause macroscopic activity changes (Harth, 1983; Freeman, 1987; Freeman & Barrie, 1994).

From all this scattered activity, a *global* function is constructed and, by way of brainstem nuclei, relayed to visual and/or other sensory projection areas. A hillclimbing process will then strengthen whatever features happen to dominate the global cost function.

There will be competition between such cortical *fancy* and ongoing sensory stimuli, with the latter generally dominating the process, except in the cases of hallucinations and dreams, and such creative activities we referred to above.

The transformation from the extended activities in the association areas and working memory to specific mental images may be likened to the collapse of a wave function in quantum mechanics, although I hesitate to refer to that much overused term. No quantum mechanical effect is here implied.

I want to stress two significant characteristics of the model.

1. Unification of sensory elements and stored memories and associations is achieved by reciprocal linking, in which those sensory elements are enhanced which

cause resonant activity at higher levels. (I leave open for the time being the question of how a cost function is constructed from what goes on in the working memory, and how this is conveyed to the brainstem reticular formation.)
2. The self-referent processes in which higher cortical levels *observe* their own reflections, make unnecessary the introduction of the pesky homunculus.

3.4. Limitations of the Model

In a critique of the sketchpad model, Francis Crick (private communication) has pointed out that hillclimbing processes such as Alopex are inherently slow, and can therefore not account for mental phenomena like pattern recognition. Although some receptive field experiments (Tzanakou et al., 1979) produced significant convergence in less than 100 iterations, the simulation studies (Harth et al., 1990) extended over thousands of iterations. This would translate into many seconds or even minutes of neural processing time. Another objection to hillclimbing processes in general is their frequent failure to arrive at the *global* maximum by becoming trapped on secondary peaks.

I commented on these objections as follows (Harth, 1996a):

1. In the computer simulations (Harth, 1980) the initial patterns were either very noisy [Figure 16(a)] or completely random. Perception would, in general, have to deal with much less ambiguous inputs.
2. Unlike the recognition of unambiguous stimuli, many cognitive processes are slow, This is especially true for creative processes, for which the sketchpad model is most applicable.
3. Convergence times may be shortened considerably if we allow a hierarchy of feature analyzers together with a similar hierarchy of hillclimbing algorithms.
4. Recognition of ambiguous stimuli may require only a slight increase in global cost function, rather than complete convergence.
5. Other processes, such as matching activity in the thalamus–PGN–cortex–brainstem complex that was suggested by Taylor (1992, 1996), Newman et al. (1996), and Baars (1996), may contribute to solving the inverse problem.
6. The random variable used in the hillclimbing procedure (the quantity g in eqn (1)) will tend to *shake* the process off secondary maxima. On the other hand, we know that thought processes are not unerringly goal-seeking, but that ''getting sidetracked'' is a frequent, and probably useful feature of mentation.

4. CONSCIOUSNESS

How does subjective awareness arise in the brain, and how can we understand it in terms of a network of neurons?

Nobody likes to define consciousness, but everybody expects others to define it. I succumbed to this pressure, fully aware of the pitfalls of any such attempt, when I wrote: ''Consciousness is … a *feeling* of subjective and perceptual unity. It is as though a unique, but probably fictitious center, the *I*, were the recipient of all the sensory inputs, past and present, and as though all messages pertaining to an object of sensory reality were reunited somewhere in the mind to constitute a consciously perceived facsimile of the same reality'' (Harth, 1995).

4.1. The Binding Problem

One of the attributes I wanted to stress is that *consciousness is a feeling*. The brain is a network of neurons, and it feels. Whether a network of artificial neurons can also have feelings is one of the hotly debated issues. The other point refers to what is now generally called the *binding problem:* what is it that confers unity, meaning, and the property of being *owned* to the activities of countless neurons that are scattered over large areas of the nervous system? Why is this thought *my* thought, as William James mused? [I want to avoid here the much overworked dichotomy of the problem into an *easy* and a *hard question* (Chalmers, 1996)].

Binding poses an almost impossible task for the brain. It appears at first that the contents of consciousness would have to be brought together both spatially and temporally to form a single, localized, and instantaneous *global* expression. It is not clear, how anything other than a single bit of information can be so compressed.

If, on the other hand, the information is spread spatially and/or temporally, we are left with the feeling that—unless there is another observer, a *homunculus*, to integrate these data—we are back to the simplistic assertion that ''a thought is *simply* a constellation of active neurons'' containing either a spatial or temporal code. (Whenever ''*simply*'' or ''*nothing but*'' are used in a definition, I suspect that a great deal of complexity is being swept under the rug.)

True spatial *binding* appears to be out. Physiologists have looked in vain for a *convergence zone* at the highest cortical levels such as the prefrontal cortex, although convergence of a sort exists in the brainstem reticular formation through which all afferent and efferent information must pass.

This leaves us with temporal binding. Eckhorn et al. (1988) proposed that coherent oscillations of around 40 Hz are the ''feature linking mechanism'' in the brain. Such periodic oscillations were found in the visual system of cats and monkeys (Doty & Kimura, 1963), and of frogs (Stiles et al., 1985). Figure 18 shows a post-stimulus spike histogram obtained in my laboratory for frog retinal ganglion cells.

The idea of binding by temporally linking scattered neural activities was taken up by Crick & Koch (1990)

and Llinás et al. (1994). I have pointed out (Harth, 1995) that the difficulty in all these schemes comes from the assumption of a final, or *output state* arising spontaneously or triggered by sensory stimuli. To make such final states account for consciousness, somehow conflicts with the notion of locality in physics. Each neuron knows only what transmitters are coming across its synapses (if we can speak of a single cell as *knowing* anything). How, then does global knowledge arise from local ignorance? Does neuron *A* know that neurons *B*, *C*, and *D are also* firing at this moment? If not, *who* or *what* does? This problem has occasioned some investigators to resort to some form of non-local quantum mechanics (Penrose, 1994; Hameroff & Penrose, 1996).

In connection with the models of temporal binding by phase-locked 40 hz activities, Freeman & Barrie (1994) pointed out that cortical neurons, unlike peripheral sensory neurons, generally do not exhibit periodic activity, Instead, ''local neighborhoods … can be viewed as an equivalence class, in which the burden of firing is rotated at random among the members''. This is reminiscent of the dynamics of neural *netlets* (Harth et al., 1970b).

4.2. The Computer Analogy

The whole idea of binding derives, I believe, from the erroneous analogy between brains and computers, which assigns in both cases initial and final states with intervening computation. In the case of computers these states are written in a code, designed and read by an external intelligence. In the case of the brain, the sensory stimuli are not really input states, since—as we have seen—they are subject to modification by the very centers for which they are intended. It is even less appropriate to speak of output states because activity in the cortex is continuous, widespread, and ever-changing. Nor is meaning assigned to these activities by some external *user*. We can simulate many brain functions on a computer, but the brain does not *compute* any more than a telescope computes the trajectory of light passing through it (Harth, 1986).

In my sketchpad model, by contrast, no final cortical states can be defined, since the cortico-thalamic–reticular system is self-referent. It appears also, that the cortical activities following a given sensory stimulus are not invariant relative to that stimulus, but change with every subsequent presentation as a function of the ever-changing store of memories and associations (Freeman & Barrie, 1994). The picture I have drawn is perhaps complementary to Taylor's *relational model* (Taylor, 1996), in which sensory inputs derive their meaning and significance from cortically stored memories and associations. Here, these higher cortical activities play a role in conscious perception through their interaction with and modification of sensory primitives at peripheral levels.

5. REDUCTIONISM RE-EXAMINED

The problem of consciousness further raises these fundamental questions: Is the problem amenable to a scientific approach? and Can we apply the methodology of reductionism? Are the two questions perhaps the same? I addressed that problem at a recent conference (Harth, 1996b).

Much has been written about the scientific method, but the concept of reductionism appears to be at its core. To physicists this term has a lucid meaning that has to do with the flow of causality in the inanimate world, which is from the small to the large, or *bottom-up*. Hence, to understand a phenomenon, we must go the other way, *reducing* it to its more elementary constituents. We understand the formation of an ice crystal, for example, by examining the properties of water molecules and their mutual interactions. The water molecule itself, or its atomic constituents, are the same whether part of an icicle or a cloud of steam. There are, to be sure, situations in which large-scale phenomena affect events in the microworld—the cold weather causing the water to freeze—but the explanation of that is again *reduced* to molecular events.

The enormous success of what I will call *classical* or *atomistic reductionism*—explaining the large by looking at the small—suggests to many that the method should be universally applicable, even to its ultimate challenge: consciousness. Thus, Francis Crick (1994) states that,

> You, your joys and sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associate molecules.

In a similar vein, the philosopher John Searle (1995) writes,

> … our entire mental life is caused by the behavior of neurons, and all they do is increase or decrease their rate of firing.

And this from the physicist Steven Weinberg (1995):

> Phenomena like mind and life do emerge … the rules they obey are not independent truths, but follow from scientific principles at a deeper level … The reductionist world view rules out other sorts of understanding.

In 1895 Sigmund Freud argued for a *Project for a Scientific Psychology*. With true reductionist fervor he proposed that human actions and emotions were to be understood by studying the underlying neural processes. He soon abandoned that approach.

Now, a hundred years later, *emergentism* is still very popular among neuroscientists and philosophers. Mental phenomena are viewed as *emerging* out of the staggering complexity of the human brain, but their roots are sought in the electro-chemical pulsations of the nerve cell.

But the *where* and *how* of this emergence had remained obscure. Although we have learned much about brain function and neural dynamics, no satisfactory causal connection has been established between brain processes and the manifestations of the mind. The best we have been able to do is point to some neural activities that are observed when we experience certain emotions or perform certain mental tasks. Physiologists speak of the *neural correlates* of mental phenomena. The finding of such correlates has been made possible by modern non-invasive techniques such as PET scans, in which shortlived radioactive substances trace enhanced blood flow and, presumably, heightened neural activity. The subject is fully conscious and alert, and can be given a variety of mental tasks. But correlation alone does not produce understanding. Apart from the fact that such correlations are evanescent, and at best approximate, no such finding produces the much desired *Aha*! reaction.

Our frustration has to do, I believe, with the fact that we are looking for understanding in the wrong place. A cursory look at life processes will convince us that we are facing here a situation that differs fundamentally from what goes on in the world of physics. Where top-down control was virtually absent in inanimate nature—or at least transparently physical in its elementary details—it plays a dominant role where life is involved.

The downward sweep of causality is nowhere as apparent as in the interplay between an individual life form, its DNA blueprint, and its role in the ecological struggle for existence. We may explain most physical properties of the individual through methods of atomistic reductionism, going all the way back to its submicroscopic blueprint, the DNA molecule and studying the copying, reading, and expressing of the genetic messages. But if we wish to understand *why* a given genome has a certain nucleotide sequence, we must look far *above* the molecular level. Unlike the *emergent* phenomenon of the ice crystal, the genome structure is a *devolved principle*. It was shaped in the battle for survival by complex ecological factors extending over eons of time.

The complexity of the situation increases drastically with the introduction of mental processes into the causal chain. Human speech requires precise timing in the activation of many muscle groups, but if we want to understand what is being said and why it is being said, we must analyze the *global* sources (semantic coding, cultural background, motivation, etc.) rather than the neuronal firing patterns. Similarly, the detailed neural processes in dreaming are under top-down control by memory structures that involve the whole gamut of personal experience and values. To explain a given human activity other than a simple reflex, we must invoke phenomena that are themselves emergent principles of a higher order. If, instead, we tried delving into details of the underlying cortical neurodynamics, reducing behavior to the action of neurons that can only ''increase or decrease their rate of firing'', we would soon be lost in a causal jungle that is both impenetrable and unenlightening.

The absurdity of pushing atomistic reductionism beyond its useful range becomes more and more apparent as we further increase the complexity of the system. In a society of a few hundred million individuals the four elementary interactions recognized in physics are of little help in tracing and solving societal problems. What guides the dynamics of a nation—what really *moves* it—are non-physical entities: political philosophies, the economy, the appeal, persuasion (and duplicity) of its leaders. Clearly, we are no longer able to construct seamless causal chains made up entirely of elementary physical interactions, though they undoubtedly exist.

I believe that the concept of reductionism has to be broadened. In fact, the Latin root of the word merely means to *lead back*. The *reductio ad absurdum* leads us back to the original statement (and proves it wrong); it does not dissect an argument into mutually contradictory elements.

Mental activities, then, are what I have called *devolved* phenomena. They are—in the sense of Taylor's (1996) *relational mind*—tied inextricably to extensive memory structures and associations that, in turn, are derived from a lifetime of interaction with our environment. Searching through the underbrush of neural connectivity may uncover regions of correlated activity, or even mechanisms, but will not satisfy our curiosity regarding the nature of thought. If mental phenomena are under top-down control, then, to understand them, we must look at higher levels.

This new type of *non-atomistic* reductionism otherwise conforms to all requirements of a scientific approach. It is just immeasurably more complex than what the ''*nothing-but*'' school has been trying to tell us. Our one-hundred-billion-neuron brains with their trillions of modifiable synapses, their largely unexplored pharmacopoeia of neurotransmitters, neuromodulators and neurohormones, are the necessary instruments on which the mental symphonies are played, but there would be little music without the greater complexity of the society of humans with its language, traditions, its art, and its history. This shows us also how very much further we have to go before we can simulate by computers the true nature of consciousness.

REFERENCES

Allanson, J. T. (1956). Some properties of a randomly connected neural network. In C. Cherry (Ed.), *Information theory* (pp. 303–313). New York: Academic Press.

Anninos, P. A., Beek, B., Csermely, T. J., Harth, E., & Pertile, G. (1970). Dynamics of neural structures. *Journal of Theoretical Biology, 26*, 121–148.

Asratian, E.A. (1965). *Compensatory adaptations, reflex activity, and the brain*. English transl. by S.A. Corson, Oxford: Pergamon Press.

Baars, B. J. (1996). Sensory projection areas of the cortex may provide a conscious global workspace for a massively distributed unconscious brain. Paper presented at Tucson II conference: Toward a science of consciousness.

Baddely, A. D. (1986). *Working memory*. London: Oxford University Press.

Beurle, R. L. (1959). Storage and manipulation of information in the brain. *I.E.E. Journal, 5*, 75–82.

Brower, B. (1933). Centrifugal influence on centripetal systems in the brain. *Journal of Nervous & Mental Diseases, 77*, 621.

Chalmers, D. J. (1996). Facing up to the problem of consciousness. In Hameroff et al. (Eds.), *Toward a science of consciousness* (pp. 5–28). Cambridge, MA: MIT Press.

Csermely, T. J., Harth, E. & Lewis, N. S. (1973). The netlet theory and cooperative phenomena in neural nets. *Journal of Dynamic Systems, Measurement and Control, Transactions of the ASME*, 315–320.

Crick, F. H. C. (1994). *The astonishing hypothesis*. New York: Scribners.

Crick, F. H. C., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in Neuroscience, 2*, 263–275.

Doty, R. W., & Kimura, D. S. (1963). Oscillatory potentials in the visual system of cats and monkeys. *Journal of Physiology (London), 168*, 205–218.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biological Cybernetics, 60*, 121–130.

Edelman, G. M. (1992). *Bright air, brilliant fire*. New York: Basic Books.

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

Finette, S., Harth, E., & Csermely, T. J. (1978). Anisotropic connectivity and cooperative phenomena in visual cortex. *Biological Cybernetics, 30*, 231–240.

Freeman, W. J. (1987). Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics, 56*, 139–150.

Freeman, W. J., & Barrie, J. M. (1994). Chaotic oscillations and the genesis of meaning in cerebral cortex. In G. Buzsáki et al. (Eds.), *Temporal coding in the brain*. Berlin–Heidelberg: Springer Verlag.

Gross, C. G., Bender, D. B., & Rocha-Miranda, E. C. (1969). Visual receptive fields in inferotemporal cortex of the monkey. *Science, N.Y., 166*, 1303–1305.

Hameroff, S. R., & Penrose, R. (1996). In S. R. Hameroff, A. W. Kasniak, & A. C. Scott (Eds.), *Toward a science of consciousness* (pp. 507–540). Cambridge, MA: MIT Press.

Harth, E. (1966). Brain models and thought processes. In E. R. Caianiello (Ed.), *Automata theory* (pp. 201–217). New York: Academic Press.

Harth, E. (1976). Visual perception: a dynamic theory. *Biological Cybernetics, 22*, 169–180.

Harth, E. (1980). Must neural mechanisms be Newtonian? *The Behavioral and Brain Sciences, 3*, 550–551.

Harth, E. (1983). Order and chaos in neural systems: an approach to the dynamics of higher brain functions. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 782–789.

Harth, E. (1986). Does the brain compute? *Behavioral and Brain Sciences, 9*, 98–99.

Harth, E. (1993). *The Creative loop*. Reading, MA: Addison Wesley.

Harth, E. (1995). The sketchpad model. *Consciousness & Cognition, 4*, 346–368.

Harth, E. (1996a). Selfreferent mechanisms as the neural basis of consciousness. In S. R. Hameroff, A. W. Kaszniak, & A. C. Scott (Eds.), *Toward a science of consciousness* (pp. 611–631). Cambridge, MA: MIT Press.

Harth, E. (1996b). The pros and cons of mind-brain reductionism. Paper presented at Tucson II conference: Toward a science of consciousness.

Harth, E., Beek, B., Pertile, G., & Young, F. (1970). Signal stabilization and noise suppression in neural systems. *Kybernetik, 7*, 113–122.

Harth, E., Csermely, T. J., Beek, B., & Lindsay, R. D. (1970). Brain functions and neural dynamics. *Journal of Theoretical Biology, 26*, 93–120.

Harth, E., & Edgar, S. L. (1967). Association by synaptic facilitation in highly damped neural nets. *Biophysical Journal, 7*, 689–717.

Harth, E., Lewis, N. S., & Csermely, T. J. (1975). The escape of Tritonia. Dynamics of a neuromuscular control mechanism. *Journal of Theoretical Biology, 55*, 201–228.

Harth, E., Pandya, A.S., & Unnikrishnan, K.P. (1986). Perception as an optimization process. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 662–665). IEEE Computer Society, Washington, DC.

Harth, E., Pandya, A. S., & Unnikrishnan, K. P. (1990). Optimization of cortical responses by feedback modification and synthesis of sensory afferents. A model of perception and REM sleep. *Concepts in Neuroscience, 1*, 53–68.

Harth, E., & Pertile, G. (1972). The role of inhibition and adaptation in sensory information processing. *Kybernetik, 10*, 32–37.

Harth, E., & Tzanakou, E. (1974). Alopex: a stochastic method for determining visual receptive fields. *Vision Research, 14*, 1475–1482.

Harth, E., & Unnikrishnan, K.P. (1985). Brainstem control of sensory information: a mechanism for perception. *International Journal of Psychophysiology, 3*, 101–119.

Harth, E., Unnikrishnan, K.P., & Pandya, A.S. (1987). The inversion of sensory processing by feedback pathways: a model of visual cognitive functions. *Science, 237*, 184–187.

Hebb, D.O. (1949). *The organization of behavior*. New York: Wiley.

Von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Naturwissenschaften, 37*, 464–476.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, London, 160*, 106–154.

Kosslyn, S.M., Alpert, N.M., Thompson, W.L., Maljkovic, V., Weise, S.B., Chabris, C.F., Hamilton, S.E., Rauch, S.L., & Buonanno, F.S. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience, 5*, 263–287.

Livingston, R. B. (1958). Central control of afferent activity. In H.H. Jasper, L. D. Proctor, R. S. Knighton, W. C. Moshay, & R. T. Costello (Eds.), *The reticular formation of the brain* (pp. 177–185). New York: Little Brown.

Llinás, R., Ribary, U., Joliot, M., & Wang, X.-L. (1994). Content and context in temporal thalamocortical binding. In Buzsáki et al. (Eds.), *Temporal coding in the brain.* Berlin–Heidelberg: Springer Verlag.

McCulloch, M. C., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Micheli-Tzanakou, E. (1984). An automated system for visual studies. *Journal of Medical Systems, 8*, 507–519.

Minsky, M., & Papert, S. (1969). *Perceptrons: an introduction to computational geometry*. Cambridge, MA: MIT Press.

Newman, J., & Baars, B.J. (1993). A neural attentional model for access to consciousness: a global workspace perspective. *Concepts in Neuroscience, 4*, 255–290.

Newman, J., Baars, B.J., & Taylor, J.G. (1996). A neurocognitive model for consciousness and attention. Paper presented at Tucson II conference: Toward a science of consciousness.

Paldino, A. M., & Harth, E. (1977). Some quantitative results on Golgi impregnated axons. *Journal of Comparative Neurology, 176*, 247–261.

Penrose, R. (1994). *Shadows of the mind*. London: Oxford University Press.

Rochester, N., Holland, J.H., Haibt, L.H., & Duda, W.L. (1956). Tests of a cell assembly theory of the action of the brain using a large

digital computer. *IRE Transactions on Information Theory, PGIT-2*, 80–93.

Rosenblatt, F. (1958). The perceptron, a probabilistic model for information storage and organization in the brain. *Psychological Review, 62,* 386–408.

Rothblatt, L., & Pribram, K. H. (1972). Selective attention: input filter or response selection. *Brain Research, 39,* 427–436.

Rumelhart, D. E., Hinton, G. E., & Williams, R. S. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533–536.

Searle, J. R. (1995, November 2). The mystery of consciousness. *The New York Review*.

Selverston, A. (1980). Are central pattern generators understandable? *Behavioral and Brain Sciences, 3,* 535–571.

Shaw, G. L., Harth, E., & Scheibel, A. B. (1982). Cooperativity and brain function: assemblies of 30 neurons. *Experimental Neurology, 77,* 324–358.

Steriade, M., Domich, L., & Oakson, G. (1986). Reticularis thalami neurons revisited: activity changes during shifts in states of vigilance. *Journal of Neuroscience, 6,* 68–81.

Stiles, M., Tzanakou, E., Michalak, R., Unnikrishnan, K.P., Goyal, P., & Harth, E. (1985). Periodic and nonperiodic burst responses in frog (*Rana pipiens*) retinal ganglion cells. *Experimental Neurology, 88,* 176–197.

Taylor, J. G. (1992). Towards a neural network model of the mind. *Neural Network World, 2,* 797–812.

Taylor, J. G. (1996). The relational mind. Paper presented at Tucson II conference: Toward a science of consciousness.

Trân Duc Thao, (1984). *Investigations into the origin of language and consciousness*. Dordrecht: D. Reidel.

Tzanakou, E., Michalak, R., & Harth, E. (1979). The Alopex process: visual receptive fields by response feedback. *Biological Cybernetics, 35,* 161–174.

Weinberg, S. (1995, October 5). Reductionism redux. *New York Review of Books*, pp. 3–9.

Willows, A. O. D. (1967). Behavior elicited by stimulation of single, identifiable brain cells. *Science, 157,* 570–574.

Wong, R., & Harth, E. (1973). Stationary states and transients in neural populations. *Journal of Theoretical Biology, 40,* 77–106.

Yingling, C. D., & Skinner, J. E. (1977). Gating of thalamic input to cerebral cortex by nucleus reticularis thalami. In J. E. Desmedt (Ed.), *Attention, voluntary contraction and event-related cerebral potentials* (pp. 70–96) *Prog. clin. Neurophysiol.* vol. 1. Basel: Karger.

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1257–1268, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00058-0

1997 SPECIAL ISSUE

The Transition from Automatic to Controlled Processing

JEFFREY A. GRAY,[1] CATALIN V. BUHUSI[2] AND NESTOR SCHMAJUK[2]

[1]Department of Psychology, Institute of Psychiatry, De Crespigny Park and [2]Department of Psychology, Duke University

(*Received* 12 *September* 1996; *accepted* 12 *March* 1997)

Abstract—*The transition from automatic (unconscious) to controlled (conscious) processing is described in terms of a neural network model of classical conditioning (Schmajuk et al., 1996). In the framework of the network, an environmental stimulus is processed in controlled or conscious mode when Novelty and attention to the stimulus are large, and in automatic or unconscious mode otherwise. In the model, indirect dopamine (DA) agonists, such as amphetamine or nicotine, enhance the DA representation of Novelty, thereby increasing attention and engaging conscious processing of environmental stimuli. By contrast, DA receptor antagonists, such as haloperidol, reduce the DA representation of Novelty, thereby decreasing attention, and engaging unconscious processing of the stimuli. © 1997 Elsevier Science Ltd.*

Keywords—Consciousness, Dopamine, Latent inhibition, Nucleus accumbens, Schizophrenia, Controlled processing, Automatic processing, Neural network.

1. INTRODUCTION

The starting point for the speculations contained in this paper lies in the cognitive abnormalities characteristic of the acute phase of schizophrenic psychosis. Patients in this stage of the illness typically display a variety of positive symptoms (Crow, 1980), such as hallucinations, delusions, fragmentation, intensification or loss of selectivity of attention, etc. These symptoms have been interpreted as reflecting a breakdown in the normal control of current information processing by stored regularities of previous input (Hemsley, 1987, 1993, 1994; Gray et al., 1991a). One consequence of this breakdown is that processing which, in normal individuals, takes place in automatic, non-conscious mode has to be dealt with in controlled or conscious mode (Schneider and Shiffrin, 1977; see discussion in Gray et al., 1991b). We have for some years been studying the disruption of latent inhibition (LI; see below) as a relatively simple experimental model of Hemsley's (1987) postulated breakdown in the control of current information processing by stored regularities of previous input (Gray et al., 1991a, b, 1995), and have recently developed a neural network model of LI (Schmajuk et al., 1996; Schmajuk, 1997). We speculate here how LI, the neural machinery that underlies LI, and our neural network model of LI may relate to the transition from automatic to controlled processing.

In so speculating, we make no claims to deal with the Hard Question about consciousness (Chalmers, 1996), i.e. the issue of how it is that brain processes, the information processing that the brain performs, and/or the transactions between the organism and its environment determined by such information processing – how any or all of these give rise to conscious experiences. One of us (Gray, 1995a, b) believes that this is indeed a Hard Question that awaits a radically new theory for its resolution; the others (Schmajuk and Axelrad, 1995; Buhusi and Schmajuk, 1996), that the Hard Question will be resolved simply by the accumulation of new data and models within existing theoretical models. We all believe, however, that the best (probably the only) thing that can be done right now is to continue with the accumulation of such data and the development of such models. It is in that spirit that the present paper is written.

2. LATENT INHIBITION AND SCHIZOPHRENIA

If a stimulus is repeatedly presented without other consequence (preexposure) and is subsequently used as the conditioned stimulus (CS) in a standard Pavlovian conditioning paradigm, the preexposed (PE) CS develops a weaker association with the unconditioned stimulus (US), as measured by the strength of the ensuing conditioned response (CR), than does a non-preexposed (NPE) CS. This difference between the CRs evoked by PE and NPE CSs, respectively, is the phenomenon of latent inhibition. Most often, different groups of subjects are conditioned with PE and NPE CSs, respectively. In this case, LI is measured as the difference in the efficacy

of conditioning between the preexposed and non-preexposed groups. This between-subjects approach may, however, be complemented by within-subject designs, in which all subjects receive both PE and NPE CSs and LI is measured as the difference in the efficacy of conditioning between the two CSs (e.g. N. S. Gray et al., 1995; Killcross and Robbins, 1993).

First described by Lubow and Moore (1959), LI has subsequently been the subject of considerable investigation (Lubow, 1989) in many species, including man. The experiments in Gray's laboratory have used both rats and human subjects. In the former case, we typically assess LI using an off-the-baseline conditioned emotional response (CER) procedure in animals licking for water (e.g. Feldon and Weiner, 1991; Weiner et al., 1984). In this procedure, after initial baseline training to lick, CS preexposure and CS–footshock pairings are both conducted without access to water, and the CS is then presented to the rat while it is again licking; CR magnitude is measured by the degree of lick suppression during the CS. LI consists of lower suppression of licking in response to the CS in the PE as compared to the NPE animals.

Another procedure used to study LI is nictitating membrane or eyeblink conditioning, in rabbits (Solomon and Moore, 1975) and rats (Schmajuk et al., 1994). In these experiments, animals in the PE condition are exposed to the CS for several days. Animals in the control, NPE condition remain in the conditioning chamber for an equivalent amount of time without CS presentations for the same number of days. On the first conditioning day, animals in the PE condition are exposed to additional trials of CS-alone presentations, while animals in the NPE condition again remain in the conditioning chamber for an equivalent time. Subsequently on the same day, when subjects in both conditions are trained in a delay conditioning paradigm, PE animals exhibit slower acquisition of the CR than NPE animals.

With human subjects, our usual procedure is based upon that of Ginton et al. (1975). Subjects first listen to a tape recording of nonsense syllables, with instructions to count the number of times they recur. In the PE condition, bursts of low-intensity white noise (the CS) are randomly superimposed on the recording. Subsequently, still listening to the tape recording, subjects are asked to predict when a counter display will be incremented; increments (the US) are preceded for all subjects by the white noise CS, and the number of trials taken to detect this contingency is the measure of conditioning.

The significance of disrupted LI for the deficit in the ability to ignore irrelevant stimuli extensively documented in schizophrenia (for review, see Hemsley, 1987) was initially pointed out by groups in Massachusetts (Solomon et al., 1981; Solomon and Staton, 1982) and Tel Aviv (Weiner et al., 1981, 1984). These groups both reported that LI is abolished in the rat by systemic administration of the indirect dopamine (DA) agonist, amphetamine. Neurochemically, the relevance to schizophrenia of this observation derives from the fact that indirect DA agonists, including amphetamine (Meltzer and Stahl, 1976), have psychotomimetic effects, while DA receptor antagonists, such as haloperidol, are used therapeutically as anti-psychotics (Carlsson, 1988; Haracz, 1982; Meltzer and Stahl, 1976; Swerdlow and Koob, 1987). Psychologically, its relevance lies in the fact that at least some positive psychotic symptoms (Crow, 1980) can be regarded as reflecting a diminished tendency to screen out redundant stimuli from processing (Anscombe, 1987; Frith, 1979), a description that seems also to apply well to an amphetamine-treated rat learning about a PE CS that is disregarded by a placebo-treated control.

In the 15 years since publication of these seminal observations, LI and its relation to schizophrenia have become the focus of increasingly intense interest (Dunn et al., 1993; Ellenbroek and Cools, 1990; Killcross et al., 1994; Weiner, 1990). This interest is largely based upon the assumption that the blockade of LI by amphetamine in the rat presents a viable experimental model of positive psychotic symptoms in man. We have recently embedded this assumption in a model of the neural substrate of LI, as well as in a general neuropsychological theory of positively-symptomatic schizophrenia (Gray et al., 1991a, b; Weiner, 1990). This theory has in part been formulated mathematically as a neural network (Schmajuk et al., 1996; Buhusi and Schmajuk, 1996), considered in more detail below; and also related to clinically observed symptoms (Hemsley, 1993 Hemsley, 1994) and, in outline, to the aberrations of conscious experience that these symptoms reflect (Gray, 1993, 1995a, b).

At the psychological level, our central hypothesis (Gray et al., 1991a; Hemsley, 1987; Weiner, 1990) states that the cognitive deficits of acute, positively-symptomatic schizophrenia stem from disruption of the normal ability to use past regularities as a guide to current information processing. Within this framework, we treat LI as a case in which the initial regularity, *to-be-CS leading to no consequence*, is followed by the requirement to learn a conflicting regularity, which in our standard conditioned suppression paradigm with rats is *CS-shock* or, in our human paradigm (Baruch et al., 1988a, b), *CS-counter increment*. In the absence of pathology or drug treatment, this sequence of conflicting regularities normally leads to retarded learning of the CS–US association (i.e. to LI). In animals which fail to show LI (e.g. after amphetamine treatment), the influence of the past regularity is lost, and they are controlled by the most recent (i.e. the CS–US) regularity. In terms of symptomatology, Hemsley has, in a series of papers (Hemsley, 1987, 1993, 1994), provided support for the general theory by showing how, given this analysis, blockade of LI can give rise to such positive

psychotic symptoms, well documented in the early stages of schizophrenia, as hyper-vigilance; and similarly how blockade of the closely-related Kamin blocking effect (Jones et al., 1992; Kamin, 1969) can give rise to delusional beliefs (elaborated in detail by Garety and Hemsley, 1994).

In support of this theory, work in Gray's laboratory has demonstrated loss of LI in normal human subjects after oral administration of a low dose of amphetamine (N. S. Gray et al., 1992a; Thornton et al., 1996); and, critically, in medicated schizophrenics, provided that they are tested in the first two weeks of the current psychotic episode (Baruch et al., 1988a; N. S. Gray et al., 1992b), before treatment with neuroleptics is able to bring symptoms under control. LI is also absent in unmedicated schizophrenics, the loss of LI in this case lasting up to a year (N. S. Gray et al., 1995). In addition, a series of studies have reported that LI is reduced in normal subjects who obtain high scores on questionnaire measurements of the tendency towards psychosis (Baruch et al., 1988b; De la Casa et al., 1993; Lipp and Vaitl, 1992; Lubow et al., 1992). At the same time, extensive studies of the rat in a number of laboratories including our own (for review, see Gray et al., 1995) have done much to elucidate the neural basis of LI and its disruption.

3. THE NEURAL BASIS OF LATENT INHIBITION

As noted above, LI is abolished by increased DA release, as caused for example by systemic administration of the indirect DA agonist, amphetamine. Behaviourally, loss of LI occurs because conditioning is increased in PE subjects, but not in NPE subjects, by indirect DA agonists. This effect is blocked by concomitant administration of DA receptor antagonists, such as haloperidol. In addition, these compounds, given on their own, induce LI under conditions (too few preexposure trials) in which undrugged animals fail to display LI; this ''potentiated LI'' is due to decreased conditioning induced by DA antagonists in PE but not NPE subjects (for review, see Gray et al., 1995). It has been proposed (Solomon et al., 1981; Solomon and Staton, 1982; Gray et al., 1991a) that these effects are due to alterations in dopaminergic transmission specifically at the terminals of the mesolimbic dopaminergic projection from nucleus (n.) A 10 in the ventral tegmental area to n. accumbens in the ventral striatal region of the basal ganglia. Thus, on this hypothesis, increased intra-accumbens DA release facilitates, while blockade of DA transmission reduces, conditioning only in subjects that have had preexposure to the CS. Experiments using systemic administration of DA agonists and antagonists have further demonstrated that both blockade and potentiation of LI can be obtained if the relevant drug is administered only at the time of conditioning, preexposure having taken place in the normal state (Joseph et al., 1993; Peters and Joseph, 1993; Moran et al., 1996). Thus, the hypothesis proposed by Gray et al. (1991a, 1995) holds more specifically that an increase in accumbens dopamine transmission at the time of conditioning reverses, and a decrease in this transmission boosts, the effects of CS preexposure; this hypothesis is consistent with Hemsley's (1987) view of schizophrenic positive psychotic symptoms as reflecting a failure in the normal integration of current information processing with previous regularities of experience.

The experiments with the rat have provided evidence in support of this hypothesis. These experiments have shown *inter alia* the following:

1. In our standard CER paradigm (see above), a non-preexposed CS associated with the footshock US comes to elicit conditioned DA release in n. accumbens, but not in the dorsal striatum (target of the nigrostriatal dopaminergic projection). However, a preexposed CS (to which conditioning is substantially reduced) fails to elicit such conditioned DA release in n. accumbens (Young et al., 1993). Thus, accumbens DA release maps well on to the changes in conditioning that constitute LI.
2. Destruction of dopaminergic terminals within n. accumbens potentiates LI (causing this phenomenon to appear with a number of PE trials too few to cause LI in normal animals); blockade of accumbens post-synaptic receptors by local injection of the DA receptor antagonist, haloperidol, at the time of conditioning only similarly potentiates LI (S. Peters, G. Grigoryan, A. Young, M. H. Joseph and J. A. Gray, in preparation). In both cases, these effects are due to changes in the preexposed animals only. Thus, in accord with the hypothesis under test, decreased intra-accumbens DA transmission at the time of conditioning boosts the effects of CS preexposure, even though this has taken place in the normal state.
3. The blockade of LI by systemic injection of the indirect DA agonists, amphetamine and nicotine, is in both cases reversed by intra-accumbens injection of haloperidol at the time of conditioning only (S. Peters *et al*., in preparation). Thus, the effect of these agonists must be mediated by DA transmission in n. accumbens, showing that (in agreement with the hypothesis) an increase in such transmission reverses the effect of CS preexposure even though this has taken place in the normal state.

Other experiments, also offering support for the neuropsychological model of LI advanced by Gray et al. (1991a, b), have provided evidence (reviewed in Gray et al., 1995) that:

4. In general, but not always, LI depends upon the integrity of the hippocampal formation, and of the retro-hippocampal region (subicular area plus entorhinal cortex) reciprocally connected to the hippocampal formation.
5. The roles of the dopaminergic projections to n.

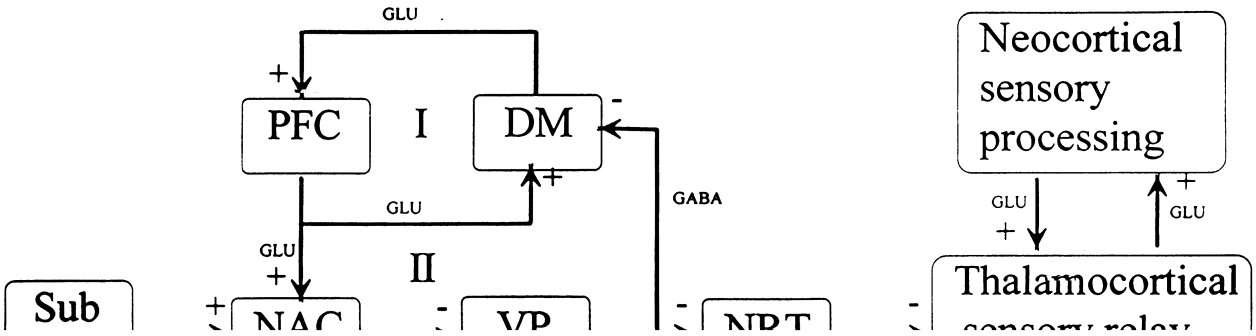


FIGURE 1. Diagram of different brain regions involved in latent inhibition. Connections from the subiculum (sub) and entorhinal cortex (ERC) ("retrohippocampal region") to the n. accumbens (NAC) and from the NAC to the nucleus reticularis thalami (NRT) and thalamocortical sensory pathways. PFC, prefrontal cortex; DM, dorsomedial thalamic area; VP, ventral pallidum; A 10, dopaminergic nucleus A 10 in the ventral tegmental area (VTA). Neurotransmitters: GLU, glutamate; DA, dopamine; GABA, gamma-aminobutyric acid. +, −: excitation and inhibition. I, II and III: feedback loops.

accumbens and the hippocampal system in LI are interconnected; thus, for example, the loss of LI after damage to the hippocampal formation or to the retrohippocampal region is in both cases restored by systemic administration of haloperidol (Christiansen and Schmajuk, 1993; Yee et al., 1995).

Although the n. accumbens is usually thought of as belonging to the motor programming circuitry of the basal ganglia (in which, indeed, it clearly plays an important role; Swerdlow and Koob, 1987; Gray et al., 1991a), it appears to play a key role in the regulation of LI. Recently, Lavin and Grace (1994) have studied what happens to the outputs from the n. accumbens. Using electrophysiological and tract-tracing techniques, these workers have demonstrated that the inhibitory GABA-ergic output from n. accumbens synapses, in the ventral pallidum, upon further GABA-ergic inhibitory neurons that project to the nucleus reticularis thalami (NRT). The NRT is unusual among thalamic nuclei in that it too consists mainly of inhibitory GABA-ergic neurons; these project to a number of the surrounding thalamic nuclei whose job is to relay impulses originating in peripheral sense organs to the appropriate sensory regions of the cerebral cortex (Jones, 1975). The possible role of the NRT in the selection of stimuli for attention and conscious processing was first pointed out by Crick (1984), and has been incorporated into a neural network model by Taylor and Alavi (1992). Note that, since the pallidal output to these neurons is itself inhibitory, its activation has the effect of disinhibiting these sensory relay pathways, i.e. increasing the entry to the cerebral cortex of those stimuli that are currently engaging the thalamocortical loops. Figure 1 presents this circuitry in diagrammatic form.

Let us consider how the circuitry of Figure 1 would be likely to work under the conditions of an experiment in which an indirect DA agonist, such as amphetamine or nicotine, is used to block LI by causing DA release in n. accumbens. As we have seen, the basic phenomenon of LI consists in the fact that the preexposed CS is slow to enter into an association with a Pavlovian US. Let us interpret this, speculatively, as reflecting a lack of access to conscious processing by the preexposed CS. If, however, presentation of this CS is accompanied by enhanced DA release in n. accumbens (as induced pharmacologically, by activation of the retrohippocampal input to n. accumbens, or during acute psychosis), LI is overcome, indicating *ex hypothesi* that the preexposed CS has regained the capacity to engage conscious processing. The circuitry of Figure 1 constitutes a mechanism by which this effect can be produced. DA release within n. accumbens inhibits (by acting on DA D2 receptors; Robertson and Jian, 1995) the GABA-ergic pathway to the ventral pallidum, thus disinhibiting the pallidal GABA-ergic pathway to NRT, which in turn inhibits the GABA-ergic projections from NRT to the ascending thalamocortical sensory relay projections, so disinhibiting the latter. In this way, accumbal DA release should lead to an intensification of processing in whatever thalamocortical sensory relay projections were already operative in the prior instant of time. In the LI experiment, this intensification of sensory processing will allow the preexposed CS (which otherwise would not have been fully processed) to enter more readily into association with the US. This line of argument is consistent

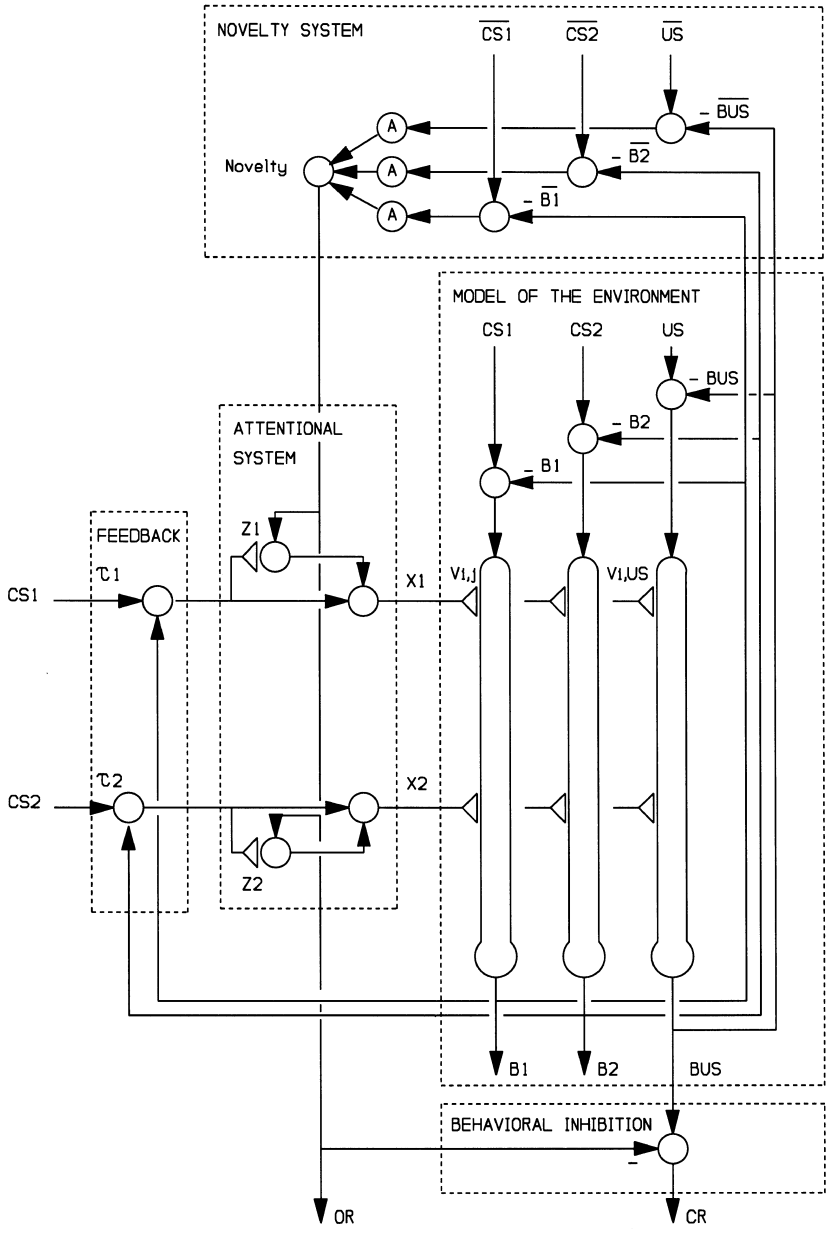


FIGURE 2. Diagram of the SLG network. CS_i, conditioned stimulus; τ_i, trace CS_i; X_i, internal representation of CS_i, z_i, attentional memory; $V_{i,US}$, CS_i–US association; $V_{i,j}$, CS_i–CS_j association; $\underline{B}_k$, aggregate prediction of event k; B_{US}, aggregate prediction of the US; $\overline{CS}_k$, average value of CS_k; $\bar{B}_k$, average prediction of CS_k; $\overline{US}$, average value of the US; $\bar{B}_{US}$, average prediction of the US; CR, conditioned response; OR, orienting response; A, Absolute values. Arrows represent fixed connections. Open triangles represent variable synapses.

with the suggestion that, in acute schizophrenia, conscious (or controlled) processing (Schneider and Shiffrin, 1977) is dominant.

This analysis of how the circuitry of Figure 1 would be expected to function when a supra-normal pulse of DA is injected into it in n. accumbens needs to be understood in the context of its likely continuing activity under more normal conditions. A striking feature of this circuitry is the imbalance between, on the one hand, the accumbens output to the thalamocortical sensory loops (3 steps) and, on the other, the dorsomedial thalamic-prefrontal motor loop (2 steps). This imbalance has the consequence that, assuming that these pathways are activated either by the same initial output from n. accumbens to the ventral pallidum or by two such outputs but simultaneously, the one set of thalamocortical loops will always be excited when the other is inhibited. This arrangement seems ideally suited to produce a rapid oscillation between the taking in of information from the environment and the organisation of a step in a motor program (see Gray et al., 1991a). Given other considerations as to the time-scale of events in these loops (Gray, 1982, 1995a), we may suppose that these oscillations between sensory-dominant and motor-dominant processing occur

every 50 ms or so. Figure 1 suggests, furthermore, that (in agreement with the hypothesis proposed by Weiner, 1990) the switch between the one and the other form of processing is controlled by the feedback loop from the ventral pallidum via A 10 back to n. accumbens. When the inhibitory step from n. accumbens to ventral pallidum is operated (leading to sensory-dominated processing), this simultaneously disinhibits the ventral pallidal input to A 10 and so inhibits the A 10 to n. accumbens pathway. In this way accumbens inhibition of the ventral pallidum is reinstated, switching dominance back to the thalamo-prefrontal loop. Interpreted without further gloss, this understanding of the circuitry of Figure 1 would seem to predict that conscious experience (assumed here to be linked to processing in the sensory but not the motor component of these linked loops, together with widespread activity in the overall comparator system, as described by Gray, 1995a) would be interrupted by recurrent periods of 50-ms blanks in every 100 ms; which is clearly not the case. We need therefore to suppose that, once the sensory processing that fills conscious experience is initiated in the thalamocortical loops, it continues until it is replaced, approximately every 100 ms, by a new set of selected elements. Many of these will presumably be the same as elements in the previous set, so resulting in an apparently seamless stream of consciousness.

4. A NEURAL NETWORK THEORY OF LATENT INHIBITION

In order to offer a formal description of the transition from automatic to controlled processing, or from conscious to unconscious modes, the present section introduces the neural network theory of LI offered by Schmajuk et al. (1996); (see also Schmajuk, 1997). Figure 2 shows a detailed diagram of the network, henceforth called the SLG model. The SLG network is a *real-time model* which describes the unbroken temporal dynamics of behavioural variables and, therefore, its output can be compared to behaviour as it unfolds in real time. Furthermore, the dynamics of its intervening variables can be contrasted with neural activity, providing a basis for the study of the physiological foundations of behaviour. This dynamic description is formalized by a set of differential equations that depict changes in the values of neural activities and connectivities as a function of time. A formal description of the model is presented in Appendix A.

The SLG network incorporates different types of memory: (1) a trace short-term memory (STM) of CS_i, (2) an attentional long-term memory (LTM) for CS_i, (3) associative CS_i–CS_k and CS_i–US LTMs, and (4) intermediate-term memories of CS_i, US, and their predictions.

Trace STM. Trace STM, $\tau_i(t)$, increases over time to a maximum during CS_i presentation and then decays back to zero (see Hull, 1943; Grossberg, 1975). In addition, we assume that $\tau_i(t)$ increases also when CS_i is predicted by other CSs (see eqn (1) in Appendix A).

Internal representation. The internal representation of CS_i, X_i, is proportional to τ_i and to τ_i multiplied by its attentional LTM, z_i (see eqn (2)). Internal representations, X_i, determine both the read-in and read-out of associative LTM (see below). The magnitude of the internal representation of CS_i, X_i, modulates both (a) the rate of change of CS_i–CS_j and CS_i–US associations (see eqns (4) and (4′)), and (b) the magnitude of the CR (see eqns (3) and (9)). That is, X_i controls both storage and retrieval of CS_i–CS_j and CS_i–US associations. Simultaneous control of memory read-in and memory read-out by X_i is a most important feature of the model.

Attentional LTM. Attentional LTM, z_i, represents the past experience that CS_i, or its prediction B_i, is accompanied by a certain amount of Novelty $\Sigma_k|\bar{\lambda}_k - \bar{B}_k|$ (defined below) (see eqn (5)). This attentional LTM is established when Novelty temporally overlaps with the STM of CS_i, τ_i, or its prediction B_i. LI is the consequence of the depressed X_i that results from the decreased attention (small positive z_i) or the inattention (negative z_i) to the preexposed CS. The magnitude of the LI effect depends on the time needed to increase X_i by reversing inattention (negative z_i) into attention (positive z_i) during conditioning.

Associative LTM. Associative LTM, $V_{i,k}$, represents the past experience that CS_i is followed by event k (CS_k or the US) (see eqns (4) and (4′)). This associative LTM is established when event k temporally overlaps with the internal representation of CS_i, X_i. By controlling the magnitude of X_i (see eqn (2)), attentional LTM z_i controls the rate of change of V_i, US associations (see eqn (4)). In addition, by controlling the magnitude of X_i, z_i also controls the size of B_{US} (see eqn (3′)) and the strength of the CR (see eqn (9)).

Cognitive mapping. The network storing CS–CS and CS–US associations in Figure 2 is a recurrent autoassociative network (Kohonen, 1977) capable of combining multiple associative values, a process similar to what Tolman (1932) called inference. Tolman hypothesized that a large number of associative values can be combined into a *cognitive map*. The internal representation of CS_i, X_i, becomes associated with the US in proportion to the difference $(\lambda_{US} - B_{US})$, where B_{US} is the aggregate prediction of the US by all CSs with representations active at a given time, and λ_{US} represents the intensity of the US. More generally, the aggregate prediction of CS_k by all CSs with representations active at a given time, B_k, is given by $B_k = \Sigma_i\ X_i\ V_{i,k}$. $V_{i,k}$ increases when $X_i(t)$ is active and event k is underpredicted, i.e. $(\lambda_k - B_k)$ is greater than zero. $V_{i,k}$ decreases when $X_i(t)$ is active and event k is overpredicted, i.e. $(\lambda_k - B_k)$ is smaller than zero. $V_{i,k}$ remains unchanged whenever $X_i(t)$ is zero or event k is perfectly predicted, i.e. $(\lambda_k - B_k)$ equals zero.

Novelty. Novelty of event k is computed as $|\bar{\lambda}_k - \bar{B}_k|$, i.e. the absolute value of the difference between λ_k, the running average of the observed value of event k, and $\bar{B}_k$, the running average of the aggregate prediction of event k. $\bar{\lambda}_k$ and $\bar{B}_k$ are computed by neural elements whose outputs are proportional to present and past values of λ_k and B_k (see eqns (6) and (7)).

Novelty is given by the sum of the novelty of all stimuli present or predicted at a given time, $\Sigma_k|\bar{\lambda}_k - \bar{B}_k|$ (see eqn (8)). We assume that CS_k can be predicted by other CSs, the context (CX), or itself. Therefore, either repeated presentations of CS_k in a given context or simply repeated presentations of CS_k lead to a decrease in CS_k novelty. Whereas CS_k–CS_k associations decrease CS_k novelty in a context-independent manner, CS_j–CS_k or CX–CS_k associations decrease CS_k novelty in a context-dependent way. Because, according to the SLG model, decrements in novelty are responsible for LI, CS_k–CS_k associations are responsible for context-nonspecific LI, whereas CS_j–CS_k or CX–CS_k associations are responsible for context-specific LI (see Good and Honey, 1993).

Figure 2 shows that Novelty becomes associated with trace $\tau_i(t)$. This association is represented by the attentional memory, z_i, which modulates the amplitude of internal representation X_i, thereby modulating the rate of learning. Therefore, the SLG model suggests that attentional memory z_i is CS_i-specific but regulated by the non-specific Novelty (see eqn (5) in Appendix A).

Orienting response (OR). Sokolov (1960) proposed that the strength of the OR might be an index of the amount of processing afforded to a given stimulus, and that this amount of processing is proportional to the novelty of the stimulus. In line with Sokolov's idea, we assume that Novelty, $\Sigma_k|\bar{\lambda}_k - \bar{B}_k|$, determines the magnitude of the OR (see eqn (9)).

Behavioural inhibition. Gray (1971) suggested that a behavioural inhibition system, activated by signals of punishment or non-reward, innate fear stimuli, or novel stimuli, gives rise to behavioural inhibition, increased arousal, and increased attention to environmental stimuli. In the same vein, we suggest that CR amplitude (a) increases proportionally to the magnitude of the prediction of the US, and (b) decreases in proportion to the magnitude of the OR, $\mathrm{CR} = \Sigma_i X_i V_{i,k} - \mathrm{OR}$ (see eqn (10)).

LI and the transition from controlled (conscious) to automatic (unconscious) processing. According to the model, at the beginning of acquisition of classical conditioning, Novelty and z_i are large because both CS_i and the US are unexpected, but both Novelty and z_i decrease as conditioning progresses and both CS_i and the US become expected. Pearce and Hall (1980) suggested that this transition from high to low levels of attention was akin to the concept of passage from controlled or conscious to automatic or non-conscious modes proposed by Schneider and Shiffrin (1977). At the beginning of conditioning, CS_i is processed in controlled or conscious mode, and later in training CS_i is processed in automatic or unconscious mode.

In terms of the SLG model, LI is the consequence of the depressed representation of CS_i, X_i, that results from the decreased Novelty that follows preexposure and from the decreased attention (small z_i) to the preexposed CS. This decreased attention can be interpreted, speculatively, as reflecting a lack of access to conscious processing by the preexposed CS.

In summary, in the framework of the SLG model, CS_i is processed in controlled or conscious mode when Novelty and z_i are large, and in automatic or non-conscious mode when Novelty and z_i are small.

5. A NEURAL SYSTEM INTERPRETATION OF THE NETWORK MODEL

In this section we consider how to translate the variables in the neural network model of latent inhibition into an equivalent set of functions on real-brain structures. In attempting to do this, we limit ourselves to the variable Novelty. Although we have begun the task of making appropriate translations for other variables within the neural network model, it is Novelty and its real-brain equivalent that are central to our present theme, the transition from automatic to controlled processing. Let us now consider how the critical variable Novelty in the SLG model can be mapped onto the circuitry of Figure 1. Under the assumption that Novelty is represented by the DA input to the accumbens, increases in Novelty result in increased DA release within n. accumbens which inhibits the GABA-ergic pathway to the ventral pallidum, thus disinhibiting the pallidal GABA-ergic pathway to NRT, which in turn inhibits the GABA-ergic projections from NRT to the ascending thalamocortical sensory projections, so disinhibiting the latter. In this way, increased Novelty increases accumbal DA release, thus leading to an intensification of processing of the CSs that activate the thalamocortical sensory relay projections, and to more conscious processing. After CS preexposure, decreases in Novelty result in a reduction of the attentional memory and the processing of the preexposed CS, and in less conscious processing.

Weiner et al. (1988) demonstrated that two administrations of amphetamine, given prior to preexposure and prior to conditioning on successive days, are effective in impairing LI. In contrast to the effect of amphetamine administration, Weiner et al. (1987) reported that haloperidol administered during both the preexposure and conditioning phases facilitates LI, but that this effect is absent when drug administration is limited to the preexposure phase.

The top left panel in Figure 3 shows mean percentage of avoidance responses for preexposed and non-preexposed groups under saline or haloperidol administration (during preexposure and conditioning) as reported by Weiner et al. (1987). Whereas animals receiving

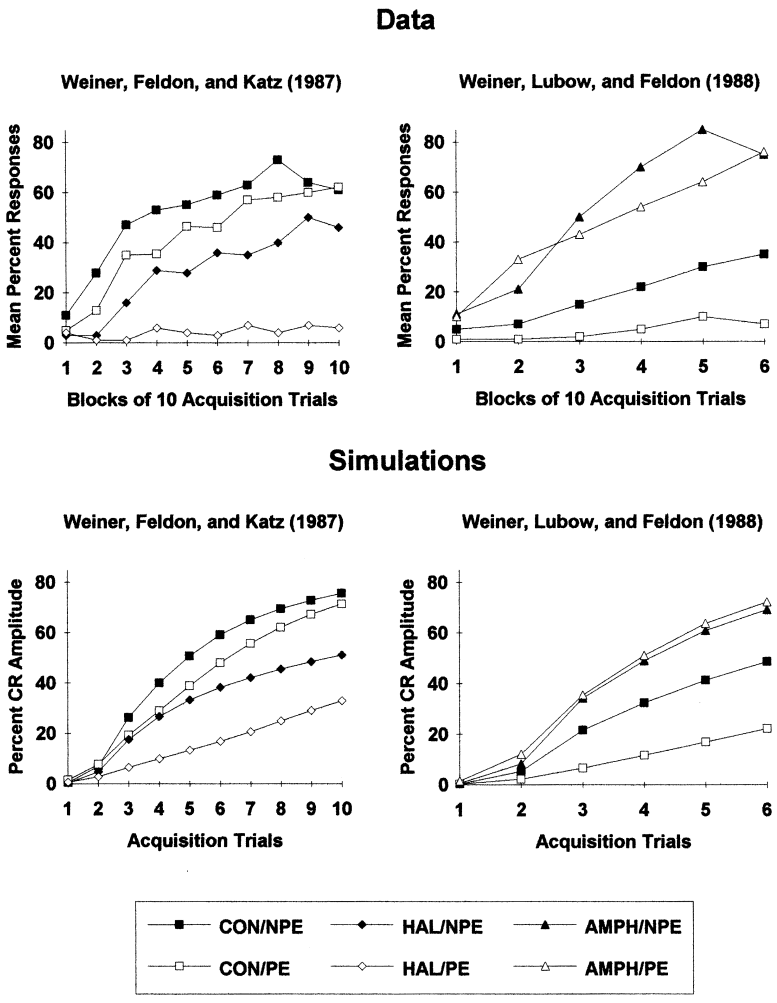


FIGURE 3. Effect of the administration of amphetamine and haloperidol on LI. Top left panel: (data from Weiner et al., 1987): Mean percentage avoidance responses for saline-control (CON) and haloperidol-treated (HAL), preexposed (PE) and non-preexposed (NPE) rats. Lower left panel: (simulation) Percentage peak CR after 5 CS preexposure trials followed by 10 conditioning trials for saline-control (CON) and haloperidol-treated (HAL), preexposed (PE) and non-preexposed (NPE) rats. Top right panel: (data from Weiner et al., 1988) Mean percentage avoidance responses for saline-control (CON) and amphetamine-treated (AMPH), preexposed (PE) and non-preexposed (NPE) rats. Lower right panel: (simulation) Percentage peak CR after 295 context-only preexposure trials, 5 CS preexposure trials followed by 6 conditioning trials for saline-control (CON) and amphetamine-treated (AMPH), preexposed (PE) and non-preexposed (NPE) rats. Because experimental data show that CS–US classical associations (interpreted as fear of the US) modulate the generation of avoidance responses (Rescorla, 1967; Rescorla and LoLordo, 1965; Weisman and Litner, 1969), we compare percent CR amplitude with the mean percent of avoidance responses.

saline (CON/NPE and CON/PE) show little difference in behaviour, PE animals treated with haloperidol show a clear retardation in the acquisition of avoidance (HAL/NPE vs. HAL/PE). In addition, animals receiving saline (CON/NPE and CON/PE) learn faster than animals receiving haloperidol (HAL/NPE vs. HAL/PE). The bottom left panel in Figure 3 presents computer simulations obtained with the SLG network showing peak CR amplitude after 5 CS preexposure trials followed by 10 conditioning trials. Although simulated HAL/PE animals show more learning than the experimental animals, in agreement with the experimental data, haloperidol administration facilitates LI.

The top right panel in Figure 3 shows mean percentage of avoidance responses for preexposed and non-preexposed groups under saline or amphetamine administration (during preexposure and conditioning) as reported by Weiner et al. (1988). Whereas animals treated with amphetamine (AMP/NPE and AMP/PE) show little difference in behaviour, PE animals receiving saline show a clear retardation in the acquisition of avoidance (CON/NPE vs. CON/PE). In addition, animals receiving amphetamine (AMP/NPE and AMP/PE) learn faster than animals receiving saline (CON/NPE and CON/PE). The bottom right panel in Figure 3 presents computer simulations obtained with the SLG network showing CR amplitude after 295 context preexposure trials representing apparatus familiarization and 5 CS preexposure trials followed by 6 conditioning trials. In agreement with the experimental data, amphetamine administration impairs LI.

In addition to describing the effect of haloperidol and amphetamine administration on LI, the model also describes LI facilitation produced by context preexposure that precedes CS preexposure (compare LI in normal preexposed animals without context preexposure (left panels) to LI in normal preexposed animals with context preexposure (right panels)).

In the model, indirect DA agonists, such as amphetamine or nicotine, enhance the DA representation of Novelty reduced by CS preexposure, thereby increasing attention, overcoming LI, and engaging conscious processing of the preexposed CS. As mentioned, CS_i is processed in controlled or conscious mode when Novelty and z_i are large, and in automatic or non-conscious mode when Novelty and z_i are small. Therefore, administration of DA agonists force the system to process CS_i in the controlled, conscious mode during conditioning. By contrast, DA receptor antagonists, such as haloperidol, further reduce the DA representation of Novelty already reduced by CS preexposure, thereby additionally decreasing attention, intensifying LI, and engaging unconscious processing of the preexposed CS.

6. CONCLUSION

In the preceding sections we have sketched out a neural network model of the basic behavioural phenomenon of LI. As indicated in detail elsewhere (Schmajuk et al., 1996; Schmajuk, 1997), this model makes a very good job of accounting for the substantial corpus of empirical data concerning LI, as summarised in the book by Lubow (1989), as well as accommodating (and therefore, in effect, predicting) a number of experimental observations published since the model was completed. We have also, but more tentatively, indicated some possible ''translations'' between the building blocks of the neural network (Figure 2), on the one hand, and systems in the real brain that have been implicated in LI experimentally (as reviewed briefly above; and see Weiner, 1990; Gray et al., 1991a, 1995). In doing so, we have concentrated here on only a few of these ''translations'', particularly those most closely related to our present theme regarding the transition from automatic to conscious processing. We shall expand the list of such possible translations in a future paper (Schmajuk, Buhusi and Gray, in preparation).

These developments are part of normal science in the relevant fields: that is, they use a standard process of reverse engineering in the attempt to account for observations on the brain and behaviour that have been gathered in standard laboratory procedures. However, in terms of current scientific understanding, the relationship between observations of this kind, on the one hand, and conscious experience, on the other, is undefined and must remain so until we are in possession of a theory that provides more than a series of brute correlations to link the two sides of this divide (see Nagel, 1993).

While awaiting the advent of that theory, it is nontheless worth speculating on just what might turn out to be linked to what in the brain-consciousness conundrum. This paper has attempted to do this for the transition between automatic and controlled processing, a transition which appears to go badly awry in schizophrenia.

REFERENCES

Anscombe, F. (1987). The disorder of consciousness in schizophrenia. *Schizophrenia Bulletin*, *13*, 241–260.

Baruch, I., Hemsley, D. R., & Gray, J. A. (1988a). Differential performance of acute and chronic schizophrenics in a latent inhibition task. *Journal of Nervous and Mental Disease*, *176*, 598–606.

Baruch, I., Hemsley, D. R., & Gray, J. A. (1988b). Latent inhibition and psychotic proneness in normal subjects. *Personality and Individual Differences*, *9*, 777–783.

Buhusi, C., & Schmajuk, N. A. (1996). Attention, configuration and hippocampal function. *Hippocampus*, *6*, 621–642.

Carlsson, A. (1988). The current status of the dopamine hypothesis of schizophrenia. *Neuropsychopharmacology*, *1*, 179–203.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.

Christiansen, B. A., & Schmajuk, N. A. (1993). Latent inhibition: effects of haloperidol and hippocampal lesions. *Society of Neuroscience Abstracts*, *19*, 798.

Crick, F. (1984). The function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences USA*, *81*, 4586–4590.

Crow, T. J. (1980). Positive and negative schizophrenic symptoms and the role of dopamine. *British Journal of Psychiatry*, *137*, 383–386.

De la Casa, L. G., Ruiz, G., & Lubow, R. E. (1993). Latent inhibition and recall/recognition of irrelevant stimuli as a function of preexposure duration in high and low psychotic-prone normals. *British Journal of Psychology*, *84*, 119–132.

Dunn, L. A., Atwater, G. E., & Kilts, C. D. (1993). Effects of antipsychotic drugs on latent inhibition: sensitivity and specificity of an animal behavioural model of clinical drug action. *Psychopharmacology*, *112*, 315–323.

Ellenbroek, B. A., & Cools, A. (1990). Animal models with construct validity for schizophrenia. *Behavioural Pharmacology*, *1*, 469–490.

Feldon, J., & Weiner, I. (1991). The latent inhibition model of schizophrenic attention disorder: haloperidol and sulpiride enhance rats' ability to ignore irrelevant stimuli. *Biological Psychiatry*, *29*, 635–646.

Frith, C. D. (1979). Consciousness, information processing and schizophrenia. *British Journal of Psychiatry*, *134*, 225–235.

Garety, P.A. & Hemsley, D.R. (1994). *Delusions: investigations into the psychology of delusional reasoning*. Maudsley Monographs, No. 36. Oxford: Oxford University Press.

Ginton, A., Urca, G., & Lubow, R. E. (1975). The effects of preexposure to a non-attended stimulus on subsequent learning: latent inhibition in adults. *Bulletin Psychonomic Society*, *5*, 5–8.

Good, M., & Honey, R. (1993). Selective hippocampus lesions abolish contextual specificity of latent inhibition and conditioning. *Behavioural Neuroscience*, *107*, 23–33.

Gray, J. A. (1971). *The psychology of fear and stress.* London: Weidenfeld and Nicholson.

Gray, J. A. (1982). *The neuropsychology of anxiety*. Oxford: Oxford University Press.

Gray, J. A. (1993). Consciousness, schizophrenia and scientific theory. In G.R. Bock & J. Marsh (Eds), *Experimental and theoretical studies of consciousness* (pp. 263–272). Ciba Foundation Symposium 174. Chichester: John Wiley and Sons.

Gray, J. A. (1995a). The contents of consciousness: a neuropsychological conjecture. *Behavioral and Brain Sciences*, *18*, 659–676.

Gray, J. A. (1995b). Consciousness and its (dis)contents. *Behavioral and Brain Sciences, 18,* 703–722.

Gray, J. A., Feldon, J., Rawlins, J. N. P., Hemsley, D. R., & Smith, A. D. (1991a). The neuropsychology of schizophrenia. *Behavioral and Brain Sciences, 14,* 1–20.

Gray, J. A., Hemsley, D. R., Feldon, J., Gray, N. S., & Rawlins, J. N. P. (1991b). Schiz bits: misses, mysteries and hits. *Behavioral and Brain Sciences, 14,* 56–84.

Gray, J. A., Joseph, M. H., Hemsley, D. R., Young, A. M. J., Warburton, E. C., Boulenguez, P., Grigoryan, G. A., Peters, S. L., Rawlins, J. N. P., Taib, C.-T., Yee, B. K., Cassaday, H., Weiner, I., Gal, G., Gusak, O., Joel, D., Shadach, E., Shalev, U., Tarrasch, R., & Feldon, J. (1995). The role of mesolimbic dopaminergic and retrohippocampal afferents to the nucleus accumbens in latent inhibition: implications for schizophrenia. *Behavioural Brain Research, 71,* 19–31.

Gray, N. S., Pickering, A. D., Hemsley, D. R., Dawling, S., & Gray, J. A. (1992a). Abolition of latent inhibition by a single 5 mg dose of d-amphetamine in man. *Psychopharmacology, 107,* 425–430.

Gray, N. S., Hemsley, D. R., & Gray, J. A. (1992b). Abolition of latent inhibition in acute, but not chronic, schizophrenics. *Neurology, Psychiatry and Brain Research, 1,* 83–89.

Gray, N.S., Pilowsky, L.S., Gray, J.A., & Kerwin, R.W. (1995). Latent inhibition in drug naive schizophrenics: relationship to duration of illness and dopamine D2 binding using SPET. *Schizophrenia Research, 17,* 95–107.

Grossberg, S. (1975). A neural model of attention, reinforcement and discrimination learning. *International Review of Neurobiology, 18,* 263–327.

Haracz, J. L. (1982). The dopamine hypothesis: an overview of studies with schizophrenic patients. *Schizophrenia Bulletin, 8,* 438–469.

Hemsley, D. R. (1987). An experimental psychological model for schizophrenia. In H. Hafner, W.F. Gattaz & W. Janzavik (Eds), *Search for the causes of schizophrenia*. Stuttgart: Springer-Verlag.

Hemsley, D. R. (1993). A simple (or simplistic?) cognitive model for schizophrenia. *Behaviour Research and Therapy, 31,* 633–645.

Hemsley, D. R. (1994). Cognitive disturbance as the link between schizophrenic symptoms and their biological bases. *Neurology, Psychiatry and Brain Research, 2,* 163–170.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton Century Crofts.

Jones, E. G. (1975). Some aspects of the organisation of the thalamic reticular complex. *Journal of Comparative Neurobiology, 162,* 285–308.

Jones, S. H., Gray, J. A., & Hemsley, D. R. (1992). Loss of the Kamin blocking effect in acute but not chronic schizophrenics. *Biological Psychiatry, 32,* 739–755.

Joseph, M. H., Peters, S. L., & Gray, J. A. (1993). Nicotine blocks latent inhibition in rats: evidence for a critical role of increased functional activity of dopamine in the mesolimbic system at conditioning rather than preexposure. *Psychopharmacology, 110,* 187–192.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B.A. Campbell & R.M. Church (Eds), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton Century Crofts.

Killcross, A. S., Dickinson, A., & Robbins, T. W. (1994). Amphetamine-induced disruptions of latent inhibition are reinforcer mediated: implications for animal models of schizophrenic attentional dysfunction. *Psychopharmacology, 115,* 185–195.

Killcross, A. S., & Robbins, T. W. (1993). Differential effects of intra-accumbens and systemic amphetamine on latent inhibition using an on-baseline, within-subject conditioned suppression paradigm. *Psychopharmacology, 110,* 479–489.

Kohonen, T. (1977). *Associative memory*. Berlin: Springer-Verlag.

Lavin, A., & Grace, A. A. (1994). Modulation of dorsal thalamic cell activity by the ventral pallidum: its role in the regulation of thalamocortical activity by the basal ganglia. *Synapse, 18,* 104–127.

Lipp, O. V., & Vaitl, D. (1992). Latent inhibition in human Pavlovian differential conditioning: effect of additional stimulation after pre-exposure and relation to schizoptypal traits. *Personality and Individual Differences, 13,* 1003–1012.

Lubow, R. E. (1989). *Latent inhibition and conditioned attention theory*. Cambridge: Cambridge University Press.

Lubow, R. E., Ingberg-Sachs, Y., Zalstein-Orda, N., & Gewirtz, J. C. (1992). Latent inhibition in high and low psychotic prone subjects. *Personality and Individual Differences, 13,* 563–572.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of non-reinforced pre-exposure to the conditioned stimulus. *Journal of Comparative and Physiological Psychology, 52,* 415–419.

Meltzer, H. Y., & Stahl, S. M. (1976). The dopamine hypothesis of schizophrenia: a review. *Schizophrenia Bulletin, 2,* 19–76.

Moran, P. M., Gray, J. A. & Joseph, M. H. (1996). Disruption of latent inhibition in the rat by a single administration of amphetamine. *Journal of Psychopharmacology, 10,* A53.

Nagel, T. (1993). What is the mind–body problem? In G. R. Bock & J. Marsh (Eds), *Experimental and theoretical studies in consciousness* (pp. 1–7). Chichester: John Wiley and Sons.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review, 87,* 532–552.

Peters, S. L., & Joseph, M. H. (1993). Haloperidol potentiation of latent inhibition in rats: evidence for a critical role at conditioning rather than preexposure. *Behavioural Pharmacology, 4,* 183–186.

Rescorla, R. A. (1967). Inhibition of delay in Pavlovian fear conditioning. *Journal of Comparative and Physiological Psychology, 64,* 114–120.

Rescorla, R. A., & LoLordo, V. M. (1965). Inhibition of avoidance behavior. *Journal of Comparative and Physiological Psychology, 59,* 406–412.

Robertson, G. S., & Jian, M. (1995). D1 and D2 dopamine receptors differentially increase fos-like immunoreactivity in accumbal projections to the ventral pallidum and midbrain. *Neuroscience, 64,* 1019–1034.

Schmajuk, N. A. (1997). *Animal learning and cognition*. Cambridge: Cambridge University Press.

Schmajuk, N. A., & Axelrad, E. (1995). Communication and consciousness: a neural network conjecture. *Behavioral and Brain Sciences, 18,* 695–696.

Schmajuk, N. A., Lam, P., & Christiansen, B. A. (1994). Hippocampectomy disrupts latent inhibition of the rat eyeblink conditioning. *Physiology and Behavior, 55,* 597–601.

Schmajuk, N., Lam, Y.-W. & Gray, J. A. (1996) Latent inhibition: a neural network approach. *Journal of Experimental Psychology (Animal Behaviour Processes), 22,* 1029–1041.

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

Sokolov, E. N. (1960). Neuronal models and the orienting reflex. In M. A. B. Brazier (Ed.), *The central nervous system and behaviour*. New York: Josiah Macey Junior Foundation.

Solomon, P. R., & Moore, J. W. (1975). Latent inhibition and stimulus generalisation of the classically conditioned nictating membrane response in rabbits following dorsal hippocampal ablation. *Journal of Comparative and Physiological Psychology, 89,* 1192–1203.

Solomon, P. R., & Staton, D. M. (1982). Differential effects of microinjections of d-amphetamine into the nucleus accumbens or the caudate-putamen on the rat's ability to ignore an irrelevant stimulus. *Biological Psychiatry, 16,* 519–537.

Solomon, P. R., Crider, A., Winkelman, J. W., Turi, A., Kamer, R. M., & Kaplan, L. J. (1981). Disrupted latent inhibition in the rat with chronic amphetamine or haloperidol-induced supersensitivity: relationship to schizophrenic attention disorder. *Biological Psychiatry, 16,* 519–537.

Swerdlow, N., & Koob, G. F. (1987). Dopamine, schizophrenia, mania and depression: toward a unified hypothesis of cortico-striato-pallidothalamic function. *Behavioral and Brain Sciences, 10,* 197–245.

Taylor, J. G., & Alavi, F. N. (1992). Mathematical analysis of a competitive network for attention. In Taylor, J.G. (Ed.), *Mathematical approaches to neural networks* (pp. 341–382). Amsterdam: Elsevier Science Publishers.

Thornton, J. C., Dawe, S., Lee, C., Capstick, C., Corr, P. J., Cotter, P., Frangou, S., Gray, N. S., Russell, M. A. H., & Gray, J. A. (1996). Effects of nicotine and amphetamine on latent inhibition in human subjects. *Psychopharmacology, 127,* 164–173.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York: Irvington Publishers.

Weiner, I. (1990). Neural substrates of latent inhibition: the switching model. *Psychological Bulletin, 108,* 442–461.

Weiner, I., Feldon, J., & Katz, Y. (1987). Facilitation of the expression but not the acquisition of latent inhibition by haloperidol in rats. *Pharmacology, Biochemistry and Behavior, 26,* 241–246.

Weiner, I., Lubow, R. E., & Feldon, J. (1981). Chronic amphetamine and latent inhibition. *Behavioural Brain Research, 2,* 285–286.

Weiner, I., Lubow, R. E., & Feldon, J. (1984). Abolition of the expression but not the acquisition of latent inhibition by chronic amphetamine in rats. *Psychopharmacology, 83,* 194–199.

Weiner, I., Lubow, R. E., & Feldon, J. (1988). Disruption of latent inhibiton by acute administration of low doses of amphetamine. *Pharmacology, Biochemistry and Behaviour, 30,* 871–878.

Weisman, R. G., & Litner, J. S. (1969). Positive conditioned reinforcement of Sidman avoidance in rats. *Journal of Comparative and Physiological Psychology, 68,* 597–603.

Yee, B., Feldon, J., & Rawlins, J. N. P. (1995). Latent inhibition in rats is abolished by NMDA-induced dopamine loss in the retrohippocampal region but this lesion effect can be prevented by systemic haloperidol treatment. *Behavioral Neuroscience, 109,* 227–240.

Young, A. M. J., Joseph, M. H., & Gray, J. A. (1993). Latent inhibition of conditioned dopamine release in the nucleus accumbens. *Neuroscience, 54,* 5–9

APPENDIX A:

A.1. A Formal Description of the SLG Model

This section formally describes the SLG model as depicted in Figure 2.

A.1.1. Short-Term Memory Trace of the CS

CS_i generate an STM trace, τ_i, according to

$$d(\tau_i)/dt = K_1(CS_i - \tau_i) \tag{1}$$

where K_1 represents the rate of increase and decay of τ_i, and the initial value of τ_i is zero.

A.1.2. Internal Representation of the CS

The combination of τ_i with the prediction of CS_i, B_i, can be modified by the attentional value z_i, to yield the internal representation of CS_i, X_i, according to

$$X_i = K_2(\tau_i + K_3 B_i)(K_4 + z_i) \tag{2}$$

where K_3 represents a reinjection coefficient for B_i, K_4 represents an unmodifiable connection between input, $\tau_i + K_3 B_i$, and X_i, and z_i is the attentional memory defined by eqn (5). By eqn (2), X_i is active either when (a) CS_i is present and τ_i is greater than zero, or (b) when CS_i is predicted by other CSs and B_i is greater than zero. Increasing values of z_i increase the magnitude of X_i. Because increasing CS_i–CS_i associations increase B_i, they also increase the magnitude of X_i.

A.1.3. Aggregate Predictions

The *aggregate prediction* of event k (CS_k) by all CSs with representations active at a given time, B_k, is given by

$$B_k = \Sigma_i V_{i,k} X_i \tag{3}$$

where $V_{i,k}$ represents the association of X_i with CS_k.

The aggregate prediction of the US by all CSs with representations active at a give time, B_{US}, is given by

$$B_{US} = \Sigma_i V_{i,US} X_i \tag{3'}$$

where $V_{i,US}$ represents the association of X_i with the US.

A.1.4. Long-Term Memory CS–US Associations

Changes in the association between X_i and the US, $V_{i,US}$, are given by

$$d(V_{i,US})/dt = K_5 X_i(\lambda_{US} - B_{US})(1 - |V_{i,US}|) \tag{4}$$

where X_i represents the internal representation of CS_i, λ_{US} the intensity of the US, B_{US} the aggregate prediction of the US by all X's active at a given time. The term $(1 - |V_{i,US}|)$ bounds $V_{i,US}$ between 1 and -1. By eqn (4), $V_{i,US}$ increases whenever X_i is active and $\lambda_{US} > B_{US}$ and decreases when $\lambda_{US} < B_{US}$. In order to prevent the extinction of conditioned inhibition or the generation of an excitatory CS by presenting a neutral CS with an inhibitory CS, we assume that $B_{US} = 0$ when $B_{US} < 0$.

A.1.5. Long-Term Memory CS–CS Associations

Changes in the association between X_i and CS_j, $V_{i,j}$, are given by

$$d(V_{i,j})/dt = K_5 X_i(\lambda_j - B_j)(1 - |V_{i,j}|) \tag{4'}$$

where X_i represents the internal representation of CS_i, λ_j represents the intensity of CS_j, B_j represents the aggregate prediction of event j by all X's active at a given time. The term $(1 - |V_{i,j}|)$ bounds $V_{i,j}$ between 1 and -1. By eqn (4′), $V_{i,j}$ increases whenever X_i is active and $\lambda_j > B_j$ and decreases when $\lambda_j < B_j$. When $B_j < 0$ then $B_j = 0$. $V_{i,i}$ represents the association of CS_i with itself.

A.1.6. Attentional Memory

Changes in the association between $(\tau_i + K_3 B_i)$ and Novelty, z_i, are given by

$$d(z_i)/dt = (\tau_i + K_3 B_i)(K_6 \text{Novelty}(1 - z_i) - K_7(1 + z_i)) \tag{5}$$

where K_6 represents the rate of increase, K_7 the rate of decay of z_i, and Novelty is given by eqn (8). When Novelty is zero, z_i decreases until $z_i = -1$.

We assume that when $z_i \leq 0$, then $X_i = K_2(\tau_i + K_3 B_i)K_4$ (see eqn (2)). This means that when z_i becomes negative, input $(\tau_i + K_3 B_i)$ activates X_i only through the unmodifiable connnection K_4. By eqn (6), whereas a CS_i with negative z_i has a relatively small X_i, and therefore, its associations change slowly, a CS_i with positive z_i has a relatively large X_i, and therefore, its associations change rapidly. LI is the consequence of the negative value acquired by z_i during CS preexposure (as Novelty decreases) and, consequently, of the time needed to restore z_i to its positive value (as Novelty increases) during conditioning. Interestingly, in line with Lubow *et al.*'s (1981) suggestion, positive values of z_i can be interpreted as a measure of the attention directed to CS_i, whereas negative values of z_i can be interpreted as a measure of the *inattention* to CS_i.

A.1.7. Total Novelty

The average observed value of event k is given by

$$\bar{\lambda}_k = (1 - \bar{\lambda}_k)\lambda_k + K_8 \bar{\lambda}_k \tag{6}$$

where K_8 represents the rate of decay of λ_k.

The average prediction of event k is given by

$$\bar{B}_k = (1 - B_k)\bar{B}_k + K_8 \bar{B}_k \tag{7}$$

where K_8 represents the rate of decay of λ_k. Novelty is given by

$$\text{Novelty} = \Sigma_k |\bar{\lambda}_k - \bar{B}_k| \tag{8}$$

where k includes all CSs and the US.

A.1.8. Orienting Response

The magnitude of the OR is given by

$$OR = \text{Novelty}^2/(K_9^2 + \text{Novelty}^2) \tag{9}$$

A.1.9. Performance Rules

The amplitude of the CR is given by

$$CR = B'_{US}(1 - K_{10}\mathrm{OR}) \tag{10}$$

where B'_{US} is given by $B'_{US} = B^2_{US}/(K^2_{11} + B^2_{US})$. According to eqn (9), the magnitude of the CR increases with increasing predictions of the US, and decreases with increasing ORs elicited by novelty.

A.1.10. Effects of Amphetamine and Haloperidol Administration

It is assumed that the administration of amphetamine or nicotine increases, and administration of haloperidol decreases, the action of the VTA dopaminergic projection to the n. accumbens coding Novelty. Novelty is given by

$$\text{Novelty} = K_{12}\Sigma_k|\lambda_k - B_k| \tag{8'}$$

where $K_{12} > 1$ for DA agonist and $K_{12} < 1$ for DA antagonist administration.

A.2. Simulation Parameters

In our computer simulations, CS-preexposure trials, context-preexposure trials, and conditioning trials last 600 time units. A CS of amplitude 1.0 is presented between time units 55 and 80, a US of intensity 1.8 is applied between time units 75 and 80. Context amplitude is 0.1. Parameter values used in all simulations are $K_1 = 0.2$, $K_2 = 2$, $K_3 = 0.4$, $K_4 = 0.1$, $K_5 = 0.005$, $K_6 = 0.02$, $K_7 = 0.005$, $K_8 = 0.005$, $K_9 = 0.75$, $K_{10} = 0.7$, $K_{11} = 0.15$, and $K_{12} = 1$. Administration of amphetamine was simulated by assuming $K_{12} = 2.0$, and administration of haloperidol by assuming $K_{12} = 0.65$.

A DOS version of the program used in the present paper is available on request.

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1269–1288, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00052-X

1997 SPECIAL ISSUE

Plato's Theory of Ideas Revisited

JAVIER ROPERO PELAEZ

Department of Mathematical Engineering, Faculty of Engineering, The University of Tokyo, 7-3-1 Bunkyo-ku, 113,

(*Received* 15 *July* 1996; *accepted* 15 *April* 1997)

Abstract—*According to Plato's theory, every pattern perceived by the mind is a composition of raw concepts (ideals). In this paper it is argued that the thalamus can decompose input patterns into these ideals and, during development, generate them from incoming patterns. As the neuronal counterpart of a certain ideal is proposed to be the activation of a reticular neuron inside a competitive layer in the thalamus, these ideals are not coincident over time. In order to recruit the ideals of a certain pattern, ideals are delayed in the long dendrites of pyramidal neurons. When several parts of a pattern are to be related, these are bound through the association of their equal constituent ideals. Coincidences of the same ideals are stored through synaptic reinforcement in nearby spines of pyramidal neurons. After binding, if a partial pattern is input, it will be reconstructed by the recall of its ideals over the reinforced synapses. © 1997 Elsevier Science Ltd.*

Keywords—Thalamus, Cortex 4th layer, Pyramidal neurons, Spine, Temporal coding, Burst firing, Orthogonalization, Binding.

1. INTRODUCTION

The 'Myth of the Cave', found in Plato's dialogue, *The Republic*, illustrates how, according to the philosopher, objects perceived by the senses are nothing but a combination of pure entities called 'ideals'. It can be imagined that some people are trapped in an underground cave in such a way that they can only see the cave's inner wall and cannot see the outside world. They think that the shadows projected onto the wall constitute the real world, and do not realize that these shadows are only poor reflections of radiant and perfect figures that exist outside the cave. In this way, every object of their world is a mixture of the projections of these figures onto the cave wall. These figures are a metaphor for the pure 'ideals' that, when combined, give rise to an entire world represented by the shadows on the cave wall. For example, according to Plato, the idea of a real tree would be a combination of the ideas of an ideal tree, the idea of greenness, the idea of height, etc. However, these ideals (the basic and distinctive characteristics that combine to give rise to the various phenomena we come across in nature) can be more abstract. For example, in the brain there appears to exist a kind of process for the decomposition of incoming visual patterns in such a way that, at a cortical level, different groups of neurons become specialized in detecting individual lines that compose the pattern.

The dwellers of the cave have no consciousness of this ideal world because they are unable to see beyond the images on the cave wall. Indeed, we are incapable of perceiving the abstract seeds of knowledge that, as a whole, produce our consciousness. In the above example we do not appreciate a visual pattern as a group of distinctive lines even though we appreciate it as a whole. Plato considered that the philosopher's (or perhaps nowadays, the neuroscientist's) task was to present their findings to those people who seemed to be trapped at the entrance of the cave.

The author writer believes that the Platonic world of ideas exists within us. It is neither eternal nor immutable, as Plato thought, but rather is created through learning and experience. According to the model of the thalamus and cortex introduced in this paper, one of the functions of the thalamus is to break down arriving information into simple and relevant components similar to Platonic ideals. Through a binding process that will be explained later, these ideals are recomposed to generate similar or other related patterns. When a new pattern arrives, there

Acknowledgements: The author expresses his gratitude to Professor Shun-ichi Amari, director of the Brain Information Processing Group in the RIKEN Institute, and Professor Kazu Aihara of the University of Tokyo for their valuable criticisms and suggestions. The author is also grateful to Shirley F. Taniguchi for reviewing this contribution.

Requests for reprints should be sent to Javier Ropero Peláez, Department of Mathematical Engineering, Faculty of Engineering, The University of Tokyo, 7-3-1 Bunkyo-ku, Tokyo 113. Tel./fax 03-5689-5752; e-mail: javier@sat.t.u-tokyo.ac.jp

would be some plausible 'candidate ideals' which are used in combination by the brain to match the incoming patterns. During this process a combination of candidate ideals would win the competition for matching the incoming pattern. If there is not a perfect match, new ideals will be created and stored for later use in dealing with upcoming patterns.

Although input patterns are perceived as simultaneous, the thalamus will split them into a temporal sequence of ideals. These ideals will be synchronized again through the delays in the long dendrites of cortex pyramidal neurons. If two or more patterns are presented simultaneously, their equal ideals will be associated through synaptic reinforcement in pyramidal neurons. For example, continuing with the Platonic metaphor, if a bush and a tree are intended to be memorized as simultaneous, the binding of both concepts must be done at the level of the characteristics or ideals they share, such as the idea of greenness, height, etc. We are not aware of this binding, that is performed in different temporal layers (Thomson & Deuchars, 1994). Therefore, this information must be synchronized in order for us to be conscious of it. This synchronization is hypothesized to be performed at a cortical level where asynchronous, non-overlapping information coming from the thalamus is delayed. The result of this process would be returned as an input to the thalamus at the same level of other input patterns (Contreras et al., 1995).

Up to this point two processes have been discussed: the generation of ideals and the binding process. As will be revealed, these two processes are closely related to the two modes of thalamic neuron firing: the tonic and burst modes. An attempt will be made to explain the overall behaviour of the thalamus and cortex in order to develop a new paradigm of neural network (Ropero, Pelaez, 1996a, b). This will include an examination of the rules of interconnectivity of the thalamus and cortex, the synchronizing role of the nonspecific thalamic nucleus, the modelization of synapses, the interplay between the tonic and bursting mode in thalamic neurons, the specific computation in every kind of neuron with a deep analysis of pyramidal processing, the mathematical algorithms that describe the 'ideals' extraction and the binding problem and, finally, the explanation of the neural counterpart that supports these algorithms.

2. NEUROLOGICAL FOUNDATIONS

2.1. Rules of Interconnectivity in the Thalamus and Cortex

The majority of the information that arrives at the neocortex is relayed by the thalamus, an ovoid-shaped structure in the middle of the inverted bowl formed by the cerebral hemispheres. Although most of the afferences from the senses to the cortex are relayed through the thalamus, the thalamus receives ten times more information back from the cortex than from the senses. Figure 1 schematically depicts the synaptic organization of the thalamus (see Steriade & Llinas, 1988).

Neurons that arrive at the thalamus make contact with two different kinds of neuron: thalamocortical and local circuit neurons. Local circuit neurons exert inhibition over each other and over thalamocortical neurons. The two types of thalamocortical neuron projections which go to cortical and to reticular neurons are excitatory. Each of the thalamic projections to the cortex is accompanied by a reverse projection from cortex pyramidal neurons to the thalamus, with matching conduction velocities. The spatial organization of the reverse projections is in a fairly precise reciprocal relationship with the forward spatial organization (see article by Crick & Asanuma in McClelland et al., 1986).

Reticular neurons in the reticular nucleus of the thalamus (RNT) possess dendrodendritic synapses and exert inhibition over each other, and over thalamocortical neurons. The inhibition over thalamocortical neurons is mediated in higher mammals' LGN parvocellular pathway by inhibitory local circuit neurons whose hypothetical role in learning will be explained later.

The fourth layer of the cortex receives most of the afferences from the thalamus. 40 Hz sparsely inhibitory spinous interneurons inhibit nearby pyramidal neurons and seem to be restricted to this area. Chandelier neurons inhibit pyramidal neurons in their axon hillock which is the strategic point in which axon potentials are generated. Basket neurons seem to regulate in a control loop manner the threshold of surrounding pyramidal neurons.

2.2. Synchronization Role of Nonspecific Thalamic Nucleus

Intralaminar and paralaminar thalamic nuclei are called nonspecific because their projections are not restricted to a unique sensorial modality. Intralaminar thalamocortical neurons exhibit a firing frequency of 40 Hz (Steriade et al., 1993). Paralaminar neurones receive inhibition from reticular neurons and project to the thalamus and to cortical layers I and IV (Llinas et al., 1994). First layer projections are performed mainly in the tuft of pyramidal neurons. Since branches in this tuft converge over the pyramidal dendritic shaft, excitations over the tuft are integrated and generate a large response that will travel through the shaft and will reinforce other stimuli over a 40 Hz basis.

2.3. Modelling Synaptic Weights

In the next example, A and B are the spikes in the presynaptic and postsynaptic space respectively. The conditional probability of B given A, $P(B/A)$, could well describe the weight which is the value of the reinforcement or weakening in the synapse. The above

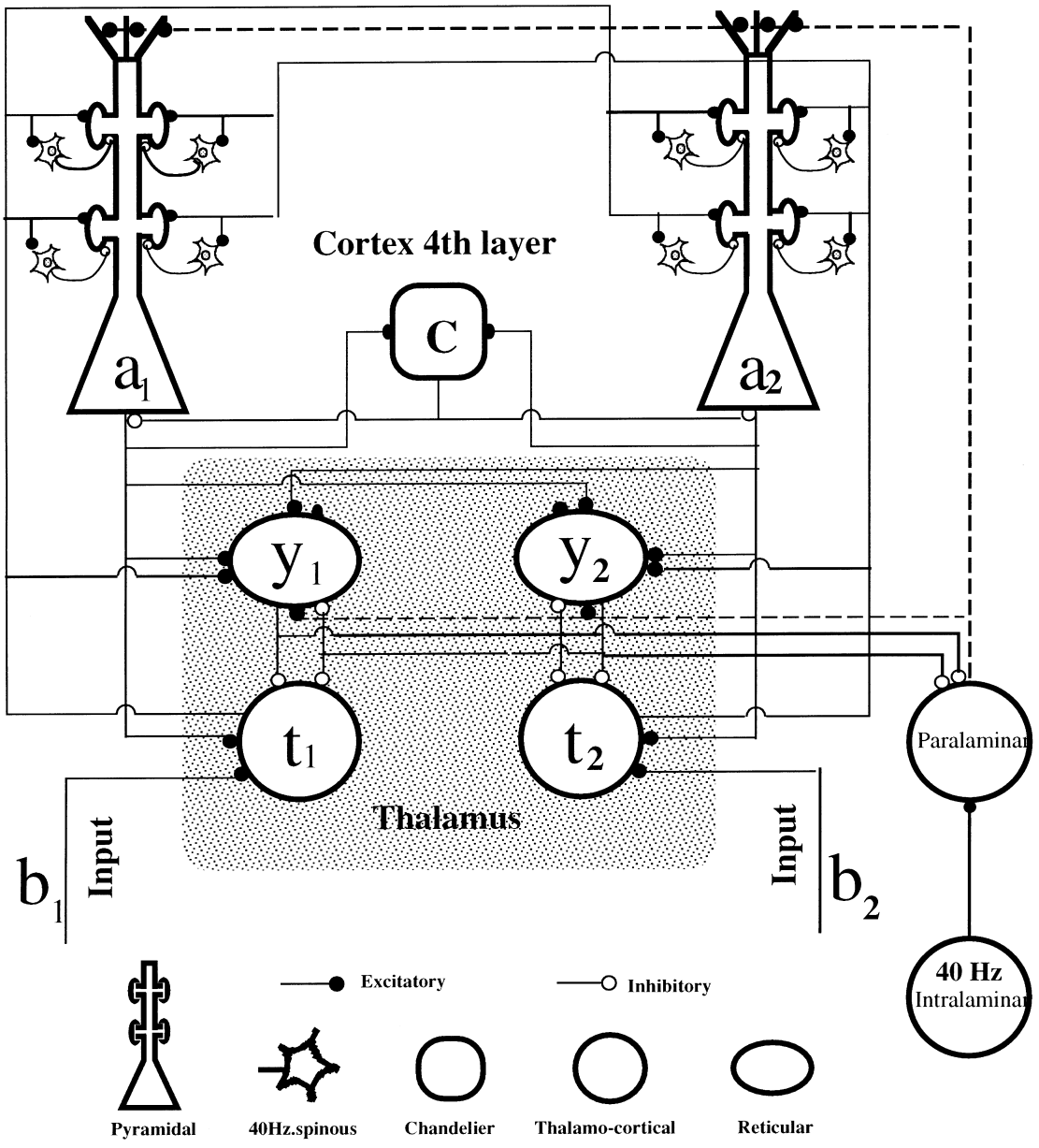


FIGURE 1. Connections between thalamus and cortex 4th layer.

expression is consistent with the latest biological discoveries (Artola et al., 1990; Huntley et al., 1994). According to these studies, reinforcement in the synapse only occurs when there is a postsynaptic spike *B* over a specific threshold, after or together with the presynaptic spike *A* (glutamate must be bound at an NMDA gate, due to event *A*, while magnesium in this gate is drawn away due to event *B*). In this case $P(B/A)$ increases. In the opposite situation, when given *A* there is no *B*, the synapse is weakened (Desmond & Levy, 1983) and $P(B/A)$ decreases. It is possible to calculate, for example, the $P(B)$ (probability of a shot in the postsynaptic space over a certain threshold) when the shot in the presynaptic space has a probability $P(A)$. The method for obtaining $P(B)$ is to multiply $P(A)$ by the weight $P(B/A)$. In this way $P(B) = P(B/A)P(A)$.

In those synapses in which the presynaptic neuron is inhibitory, there are no modifiable weights because NMDA channels are not present there. Despite this, there exists a possibility of computing synaptic weights in the synapses of reticular thalamic neurons through a structure called the glomerulus. A glomerulus is a complex synaptic structure in which contacts between thalamocortical and local circuit neurons occur in a synaptic island encapsulated by glial cells. The most frequent case is depicted in Figure 2. In this case an axon terminal of an input neuron *b* excites the dendrites of local-circuit neurons l_1 and l_2 and thalamocortical neuron *t*. Local-circuit neuron l_2 inhibits l_1 and this latter exerts inhibition over *t*. Corticothalamic and reticular neurons produce, respectively, excitation and inhibition over local circuit neurons in the extraglomerular

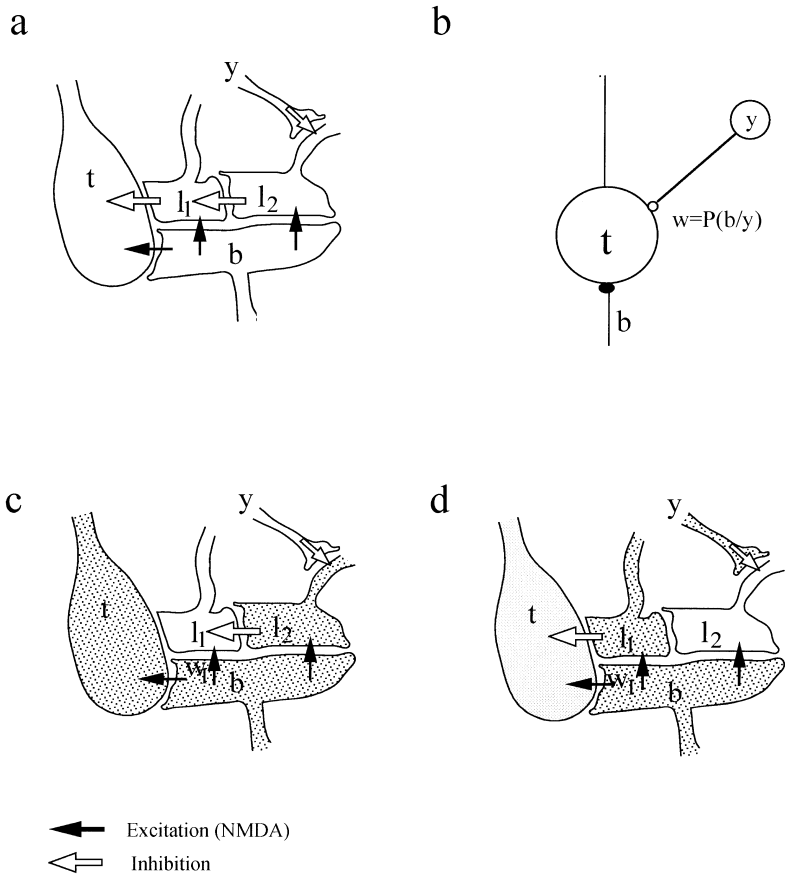


FIGURE 2. Equivalence between the hypothetical process carried out in thalamic glomeruli in which inhibitory synapses are not modifiable (a) and in an artificial model (b) in which these synapses have a weight defined by $P(b/y)$: t = thalamocortical dendrite; l_1 and l_2 = local-circuit neurons' dendrites; b = input terminal; y = reticular neuron. (c) and (d) On 'b' being activated, l_1 is active only when there is inhibition from 'y'. This allows us to calculate weight w_1 as $P(b/y)$, the influence from 'b' being $P(b)P(b/y)$ which can be rewritten as $P(y)P(y/b)$. This expression is equivalent to an inhibition from 'y' through a synaptic weight $P(y/b)$, as depicted in (b).

neuropil (in Figure 2, inhibition from reticular neuron y is represented). The following explanation shows that the structure in Figure 2a is equivalent to that in Figure 2b in which the weight in the inhibitory synapse from the reticular neuron is modelled as $P(b/y)$. The net input to the thalamic neuron can be measured as the probability of the excitation from b minus the probability of the inhibition from y: $P(b) - P(b/y)P(y)$.

In Figure 2c, dotted areas represent depolarized structures. In this instance input b depolarizes l_1 and l_2, but l_2 inhibition over l_1 neutralizes l_1 depolarization. No inhibition from l_1 over t is produced.

Figure 2d shows what happens when there is a reticular inhibition by y. In this instance, l_2 depolarization is neutralized so that it cannot inhibit l_1. Because of this, an input from b excites l_1, which inhibits the thalamocortical t. This can lead us to a rule of thumb in which, when there is reticular inhibition from y, which is made effective by input b, then l_1 is active as well. In this way, activity of l_1 is highly correlated to activity in y. Because of this, it is possible to model the weight w_1 between l_1 and b as $P(y/b)$. Thus the excitation from b over l_1 is $P(y/b)P(b)$, which can be rewritten as $P(b/y)P(y)$. The net input probability over t is $P(b) - P(b/y)P(y)$, as in Figure 2b.

2.4. Types of Neuronal Firing

2.4.1. *Thalamocortical Neurons.* Thalamocortical (thal-cor) neurons have two ways of firing (Llinas & Jahnsen, 1982). They may fire tonically when they are excited above a certain threshold, as happens in the rest of the neurons (Figure 3a). The frequency of firing of thalamocortical neurons during their tonic mode depends on the summation of their inputs. In Figure 3a, the thalamocortical neuron would fire in this mode with probability $P(b) - P(y)$, which would be the sum of the excitatory and inhibitory input probabilities.

The other mode of firing is a high frequency (> 250 Hz) burst firing that occurs after the accomplishment of a precise sequence of facts. First, a small depolarization from rest is required. Afterwards, this depolarization must be reduced by an active hyperpolarization or from a reduction of a depolarizing input. When the neuron is maintained in this state for at least 50–100 ms, the bursting mode is produced by a

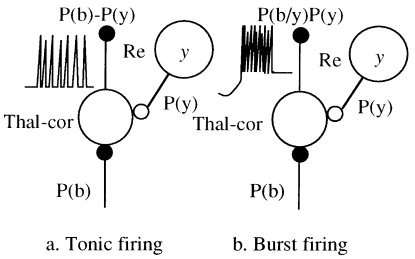


FIGURE 3. Two types of firing in thalamic neurons. (a) The probability of firing in tonic mode is the sum of the probabilities of the inputs. (b) In bursting mode, a necessary condition for bursting is the previous inhibition by neuron '*y*'. Therefore the probability of bursts in the thalamocortical neuron is modelled by the mathematical expression of the conditioned probability of '*b*' due to '*y*' multiplied by the probability of bursts in *y*.

small depolarization, The probability of a high frequency burst has been modelized in Figure 3b as $P(b/y)P(y)$, The expression of the conditional probability of b given y is used here because the condition for b to allow thalamocortical neuron burst firing is the previous inhibition from the reticular neuron 'y', which has a probability $P(y)$ (Figure 2b).

Reticular neurons exhibit not only the same kind of firing as thalamocortical neurons (Llinas, 1988) but also a unique pacemaker-like rhythmicity (Pinault & Deschenes, 1992). Their firing is also characterized by a second rebound in the burst mode whose importance will be remarked upon during the course of the paper. Threshold accomodation is another specific characteristic of reticular neurons that is necessary to produce alternation of reticular neuron activations in the NRT, as will be suggested.

2.4.2. *40 Hz Inhibitory Stellate Spinous Neurons.* 40 Hz inhibitory stellate spinous neurons are present mainly in the fourth layer of the cortex, which is the region that receives most of the afferences from the thalamus. Their inputs come from the thalamus and their efferences go to the dendritic spines of pyramidal neurons. In vitro intracellular recordings demonstrate a 40 Hz oscillatory response after depolarization. Another characteristic that is relevant, due to its computational possibilities, is that the response outlasts the first stimulus but comes to an abrupt cessation in the middle of a second one (Llinas et al., 1991) in a 'flip-flop' manner.

2.4.3. *Pyramidal Neurons and Sequential Processes.* Pyramidal cells represent the most abundant and characteristic neuronal type of the cerebral cortex (DeFelipe & Fariñas, 1992). They have a long apical dendrite with bulbous-like synapses called dendritic spines connected to the dendrite by a 'neck'. One or two excitatory inputs could arrive at a dendritic spine mainly in the spine head and, in many cases, there is also an afference from an inhibitory input neuron near or on the spine neck. Inhibition in this specific place could cut off most of a previous

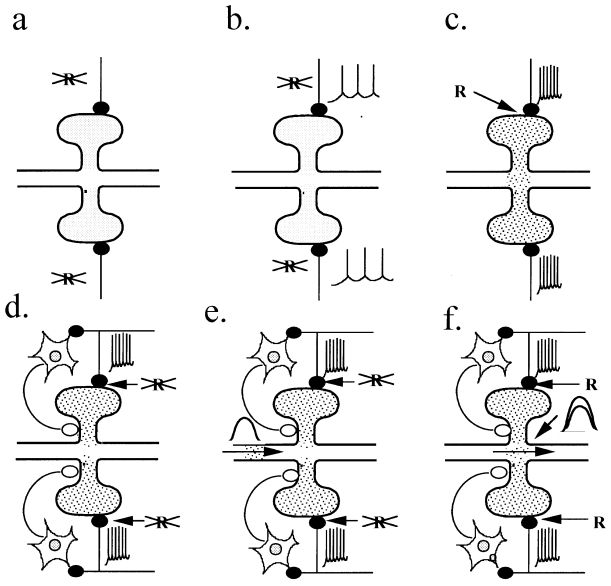


FIGURE 4. Different theoretical situations in spine signal processing. (a) No excitation over the spines; calcium ions remain inside. (b) Tonic excitation arrives at both spines. Although a little calcium leaves the spines, the frequency of the inputs is not high enough to allow the communication of the two spines through a calcium ion's path, nor to allow properly the synchronicity of the spikes in both spines. There is not synaptic reinforcement *R*. (c) Every spine receives a burst of spikes. The high frequency of bursting allows the release of calcium to the dendrite and the reinforcement of both synapses. (d) A burst excitation is applied over the spines and over inhibitory neurons. The consequence is a depolarization that tries to release calcium. However, when this ion tries to go to the dendrite the hyperpolarization due to the inhibitory neuron brings back the calcium to the spine. Neither calcium path nor reinforcement is produced. (e) The same case as in (d), but there is a bearer depolarization that comes through the dendrite. As it does not arrive at the inter-spine gap at the time that the spines receive bursting excitation, the result is as in (d). (f) In this case the bearer wave is synchronized with the spikes' burst so that it arrives at the inter-spine gap at the same time as the burst firing. LTP is produced and the bearer wave is boosted due to the added calcium ion potential.

excitation. In the cortex 4th layer, efferences from the thalamus are used to make contact over 40 Hz inhibitory neurons and over pyramidal neurons. Looking carefully at the disposition of the synapses over a spine in the cortex 4th layer (Figure 1) it is possible to hypothesize that these inhibitory neurons are able to cut off the excitation from the thalamus. Spines have the property of accumulating calcium. Although this fact has an important biological meaning (calcium leads to cell death, so spines can be considered as isolated compartments where high concentrations of Ca^{2+} are stored without damage to the cell function), calcium would be stored as a reservoir of ions for boosting travelling potentials through the dendrite. It also could provide an effective path for signal transmission and synaptic reinforcement between two nearby spines (Yuste & Denk, 1995; Yuste & Tank, 1996).

Taking into account these properties of dendritic spines, several theoretical situations of spine computation have been depicted in Figure 4. These situations will allow us to understand temporal computation of inputs over dendrites. The conclusion is that high frequency bursts of spikes could promote reinforcement in nearby spines (Figure 4c) while this situation is exceptional when low frequency spikes (Figure 4b) arrive at the spine. When an input volley arrives simultaneously at the spine and at a 4th layer inhibitory spinous neuron, the inhibitory neuron could hypothetically block both the signal and calcium diffusion into the dendrite, making difficult the reinforcement of synapses of nearby spines, Figure 4d. However, when, synchronously with excitation of the spine, a big depolarization passes through the dendritic shaft invading the neck of these spines, excitation in a spine would be transmitted faithfully to the other spines and reinforcement would be accomplished, Figure 4f. Furthermore, a travelling depolarization would be boosted. If the depolarization is enhanced along its course by different active spines in the dendritic shaft, this depolarization can be seen as a bearer wave that carries the information of the sum of different inputs. If every input is required to be summed once, and all the inputs are to be considered in the sum, a top-down summation must be accomplished, from the most distal input in the upper layer to the lowest one near the soma. In order to obtain this top-down summation, the inputs must be ordered according to time; in other words, the first input must be in the upper part of the dendrite and the last input in the lower part. If it is supposed that every event inside consciousness not only generates a specific spatial pattern but also a sequentially ordered pattern, it can be seen that every input to the dendrite might be ordered according to time in a top-down manner. According to the hypothesis stated below, this

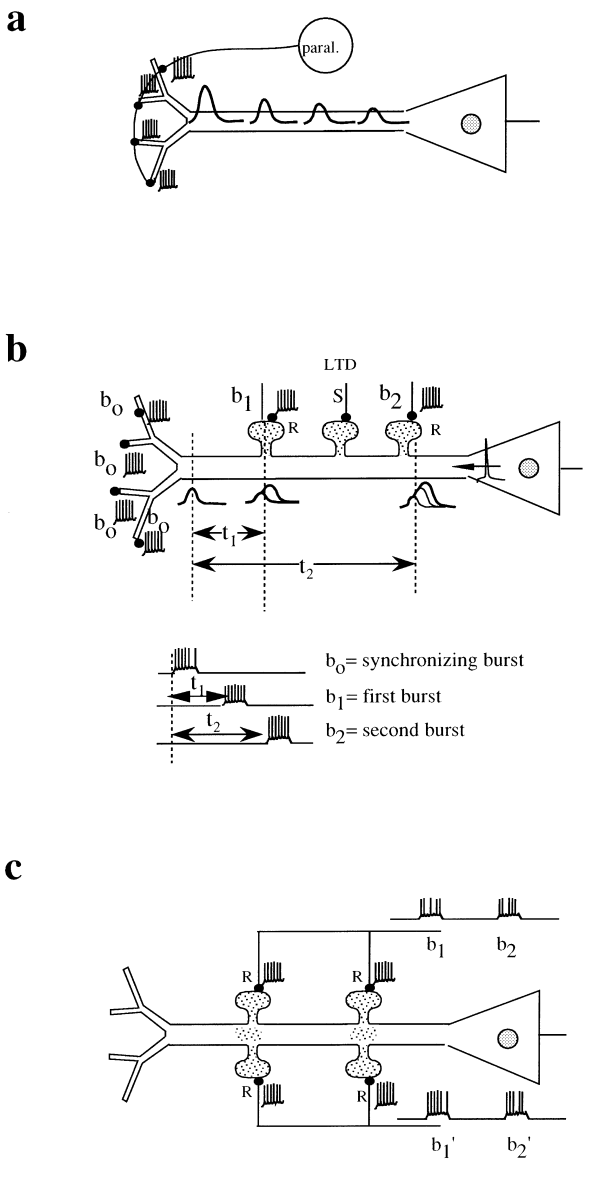


FIGURE 5. Retrograde action potentials can theoretically give rise to a hebbian learning that reinforces (*R*) not only the appropriate synapses but also a temporal order of the synaptic stimulus. Paralaminar thalamic burst depolarizes the branches of the pyramidal tuft, and this depolarization is integrated to form a big 'bearer wave' that can be boosted if the spine's synaptic depolarization occurs when this wave passes near the spine. The consequence of this boosting might be the reinforcement of the active synapses through a retrograde action potential. (a) Various responses due to a paralaminar thalamic neuron in the tuft of an apical dendrite are summed and generate a large wave that travels through the dendritic shaft. We have considered that, because of the shaft cable properties, the amplitude will diminish and will not be large enough to generate an action potential after reaching the soma. (b) Let us suppose that, as a consequence of the paralaminar burst over cortical neurons, some of the subsequent bursts will arrive at the neuron under study with a fixed delay after the intralaminar burst. Let us say that the synchronizing burst arrives at time $t = 0$ while the others will arrive with delays t_1 and t_2. In the figure b_1 and b_2 are ordered according to time, from the upper part of the dendritic shaft to the lower part. In this way the first burst will take place nearer the tuft and the last one nearer the soma. The synaptic locations of b_1 and b_2 are purposely placed at a proper distance so that the 'bearer wave', the one originating at the tuft, could reach the first synapse at time t_1 and the second at time t_2. In this case, the bearer wave is boosted twice and will become big enough to promote a retrograde potential that will reinforce the previously excited synapses (because glutamate appears to be bound at the NMDA gate, at least until the retrograde spike depolarizes the postsynaptic membrane above a certain threshold). Although a stimulation in spine *S* may not coincide with the passing of the bearer wave, it is also reinforced by means of the retrograde action potential. (c) This case could be helpful for understanding those in Figure 7. In this case we have burst b_1 and b_2 going through the same axon. They are produced at the same time as bursts $b_{1'}$ and $b_{2'}$ that are travelling through the axon below. Since this case is the same as that of Figure 4c, reinforcement is produced. In this particular case every synapse is reinforced twice, first by the pair b_1 and $b_{1'}$ and secondly by b_2 and $b_{2'}$. As has been explained in Figure 4c, the 'bearer wave' is not necessary here for promoting reinforcement.

organization is achieved through a mechanism that involves the reinforcement of the synapses that accomplish this time-based top-down organization, as explained in Figure 5.

In Figure 5a is represented the course of a big 'bearer wave' generated by the convergence of the 40 Hz paralaminar spikes over all the dendrites of the tuft of a certain pyramidal neuron. This response is weakened gradually and if, when reaching the soma, it is not high enough it will not generate an action potential. This bearer wave can be amplified during its course if it is coincident with other incoming inputs. In Figure 5b, the paralaminar burst b_0 is supposed to produce excitation, not only in the tuft of the neuron but also in the other two neurons that, due to burst b_0, will subsequently fire with bursts b_1 and b_2 over the neurons depicted in the figure having a t_1 and t_2 delay time respectively from burst b_0. If the synapses are placed in a position that allows the coincidence of the bearer wave produced by b_0 and the excitations from b_1 and b_2, the bearer wave will be boosted twice and probably will be big enough to produce an action potential in the soma. Together with this action potential, a retrograde action potential (Stuart & Sakmann, 1994) will return to the apical dendrite, reinforcing the synapses that were previously active. This reinforcement is produced because the neurotransmitter glutamate that was previously released in the synaptic gap is still bound to the NMDA gate when the retrograde action potential reaches the synapse (Lester et al., 1990). In this moment the strong potential repels the magnesium that was blocking the channel, and calcium is able to enter the spine, promoting the synaptic reinforcement. If, in the case of a retrograde potential, there were synapses that previously were not coincident with the bearer wave, they would be potentiated too. In short, potentiation will be produced in all the synapses that receive inputs if there is a retrograde action potential due to a suprathreshold activation of the soma produced by a minimal group of active inputs that boosted the bearer wave.

Another way of producing synaptic reinforcement is by means of the coincidence of impulses in nearby neurons, Figure 5c. The situation in this figure anticipates some concepts that will be seen later: two trains of highly correlated bursts will reinforce all the synapses in the figure as many times as the number of bursts in each train, without any consideration of the order of bursts nor of the exact point where spines are placed.

Several other questions must be addressed. There is one way to avoid the reinforcement of the synapses in this last case of coincidence of pulses, or even in the case where the proper order of inputs is accomplished. Since the action potential generated in the axon hillock is the one that, retrogradely, allows the reinforcement of the synapses, if an inhibition is applied in the hillock, no action potential is generated and no reinforcement occurs. This could give some insight into the role of certain inhibitory neurons, such as chandelier cells that strongly inhibit the cell at that precise point.

The top-down organization of synapses inside a pyramidal neuron could give rise to an organization of sequential firing in different pyramidal neurons. As a general idea, the order of the input neurons in the apical dendrite will define the order in the sequence. In Figure 6, a simple case of sequentiality with three pyramidal neurons is shown with examples of clockwise, counterclockwise, cluster and reverberant states. It tries to draw the reader's attention towards the importance of inhibition in sequential processing in the cortex, and towards the possible role of inhibitory chandelier and basket neurons. According to this hypothesis, chandelier cells avoid coincident activation over pyramidal neurons, allowing sequentiality in their inputs. It is remarkable that chandelier cells inhibit the coincidence of the activation of pyramidal neurons over other pyramidal neurons, but are not capable of avoiding thalamocortical coincident activation because afferences from the thalamus over chandelier cells are unusual. In this way, pyramidal neurons seem to promote sequentiality through temporal integration, mainly at a cortical level, as well as detection of thalamocortical coincidences. This hibrid mode hypothesis integrates former ones (König et al., 1996).

Inhibition from basket neurons regulates the firing threshold of pyramidal neurons. Consequently, if there is little inhibition, neurons fire when receiving the sequential activation of very few neurons. The range of input neurons that, depending on the threshold, produce the pyramidal activation define which neurons take part in a certain sequence. If the inhibition is high, a lot of input neurons are necessary for reaching the firing threshold. In this case the sequence comprises more neurons. If the inhibition is low, very few inputs produce the pyramidal firing, and the sequences comprise only these few input neurons. In this case, different clusters of neurons can produce different sequential loops at the same time.

In the cortex 4th layer, the present work suggests that not spines but rather clusters of spines (see Yuste & Tank, 1996) are also ordered through time in the way explained above. In this case, the reinforcement is not due to a retrograde action potential but, instead, is due to the simultaneous coincidence of the bearer wave and the activations of the inputs over the same spine as in Figure 4f. While in the previous case there was reinforcement in active spines whose inputs were not coincident with the bearer wave, in the present case of the 4th layer, reinforcement does not occur if the activation of the inputs is not coincident with the synchronizing wave (Figure 7a). If the same thalamocortical axons make contact over different spines on the section of a dendritic shaft in the 4th layer, the spatial organization of the spines will mimic the temporal organization of the bursts in thalamocortical axons, Figure 7b and 7c.

The hypothesis that thalamocortical axons convey precise sequences of bursts will be explained in this work.

Sequentiality, as will be shown, is promoted not only at a cortical level; the thalamus could also promote sequentiality due to the possibility that it could decompose input patterns into a sequence of raw 'ideals' that are produced in order through an iterative sequential method that will be explained.

These two reasons, sequentiality promoted by

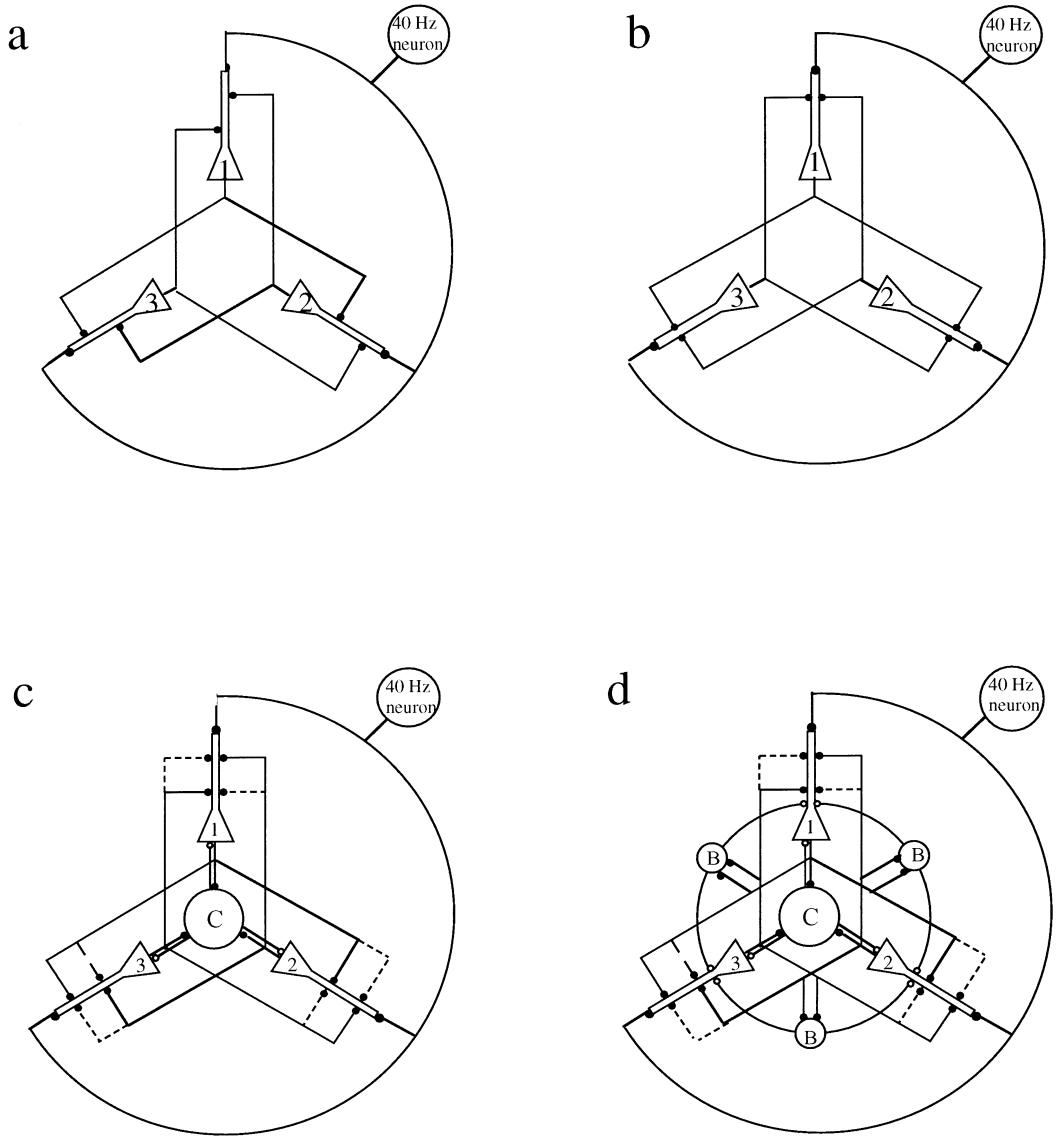


FIGURE 6. A temporal organization of synapses in the apical dendrite of pyramidal neurons could cause a group of these neurons to fire sequentially. (a) In this case a neuron will fire if the 40 Hz neuron fires first and the other two inputs receive excitation in the appropriate order (threshold = 2; 2 inputs). In this group of neurons the sequence of firing is clockwise, thus: 1 2 3 1 2 3 1 2 .... The firing rate of each neuron is 40/3 Hz. (b) In this case the three neurons fire simultaneously (in an 'epileptic' manner) at 40 Hz because of the coincident activation of two inputs in all neurons. (c) In a combination of the above cases, there are two more connections in each pyramidal neuron. There is also a chandelier cell (with an arbitrary threshold = 2 input neurons). When it receives two or more excitations it inhibits all pyramidal neurons, avoiding coincident activation over them. Here there are two possible sequences of firing: clockwise as in (a), or counterclockwise 3 2 1 3 2 1 3 2 ... (which is performed through broken lines). The system is hooked in one of the sequences depending on which organization synchronizes first with the 40 Hz synchronizing input. However, it could be forced into one sequence by means of an external activation. (d) Three basket neurons are added. Basket neurons only regulate the threshold of pyramidal neurons, instead of cutting off their firing. They are inhibiting according to the surrounding level of pyramidal activation in a control loop manner. Subsequently the pyramidal neurons' threshold increases when it is more inhibited. Depending on the pyramidal neurons' threshold, the sequences of firing vary from an 'epileptic' behaviour (threshold = 0 when at the very beginning the network was in a steady state) to a wide sequence of three neurons as in cases (a) or (c) (threshold = 2 at the end) through three oscillatory possibilities 1 2 1 2 1 ..., 1 3 1 3 1 ... or 2 3 2 3 ... (when threshold = 2 in a transitory state). In these cases the threshold is proportional to the number of active basket neurons.

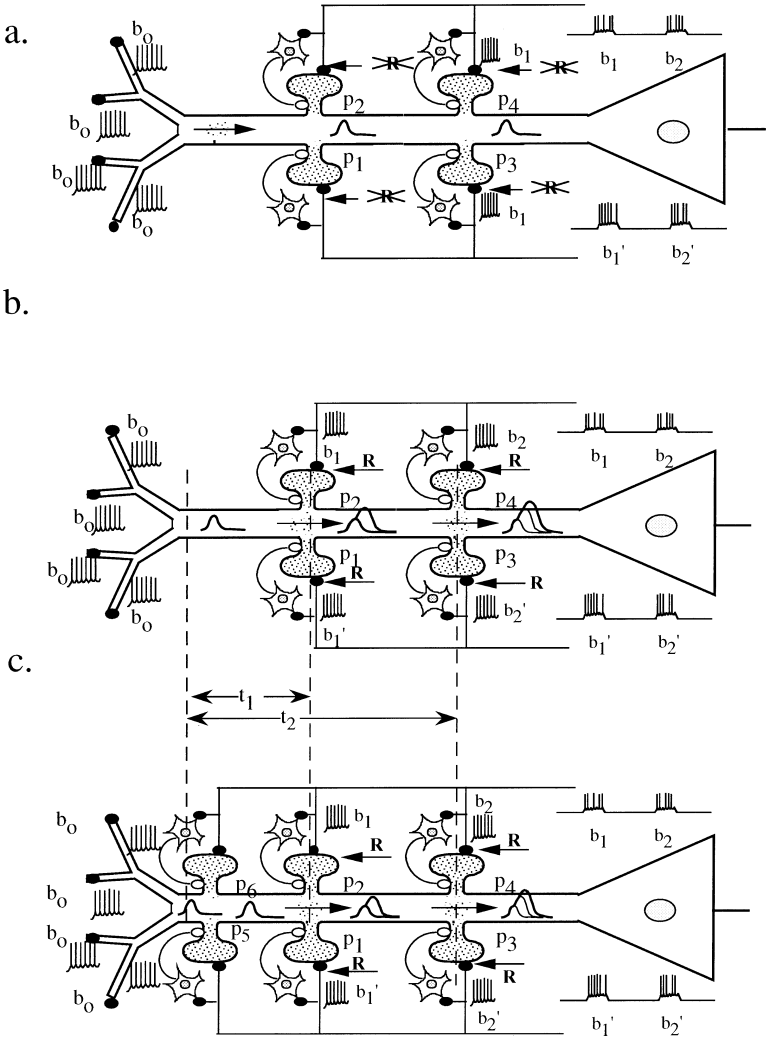


FIGURE 7. Three different theoretical situations in layer IV signal processing. (a) Two trains of bursts coming through two different axons are represented at both sides of the neuron. Bursts b_1 and $b_{1'}$ are produced at the same time as b_2 and $b_{2'}$. Since the first couple of bursts, b_1 and $b_{1'}$, arrive synchronously at the spines p_4 and p_3 but not at the same instant as the 'bearer wave' generated by b_O, synapses in these spines are not reinforced (see also Figure 4e). Synapses in p_1 and p_2 are not reinforced because they have not already been stimulated. (b) At time t_1, $b_1$and $b_{1'}$ arrive at the synapses in p_2 and p_1 respectively synchronously with the bearer wave produced by b_0. At t_2 the same occurs by means of bursts b_2 and $b_{2'}$ over synapses in p_4 and p_3. The result is a reinforcement R of all these synapses (see also Figure 4f). (c) This is the same case as (b) but with an extra pair of spines. Since these spines p_5 and p_6 are not properly placed, the bearer wave is not coincident with the bursts and reinforcement is not produced. Synaptic depression in the spines of both pyramidal and sparsely spinous interneurons will hypothetically cause these spines to be disconnected. In this way spatial and temporal synchronization can be produced during development. It should be noted that in the above cases, in which spines act as coincident detectors, a retrograde potential is not necessary for reinforcing an appropiate spatial order for placing incoming inputs over the dendritic shaft, as seen in Figure 5.

pyramidal inputs and a sequential decomposition of input patterns in the thalamus, are why thalamocortical inputs to the cortex might be ordered according to time in dendrites.

3. IDEALS AS AXES IN PATTERN VECTOR CODING

In the Myth of the Cave Plato describes the 'world of ideas' in which eternal and immutable patterns give rise to various phenomena in nature. Similarly, there exist in the brain specific locations that store very abstract codifications of the world, as in the retina which decomposes the colour into the three basic types, or in the tonotopic and visuotopic map in the cerebral cortex. According to several authors (DeValois & DeValois, 1988), these visuotopic maps, in which cortical areas respond differently to different spatial frequencies of patterns, are the result of Fourier decomposition, which is a kind of mathematical process known as orthogonalization. One way to think of this is that all the information entering the cortex must previously undergo some kind of orthogonalization.

Figure 8 is an example of orthogonalization

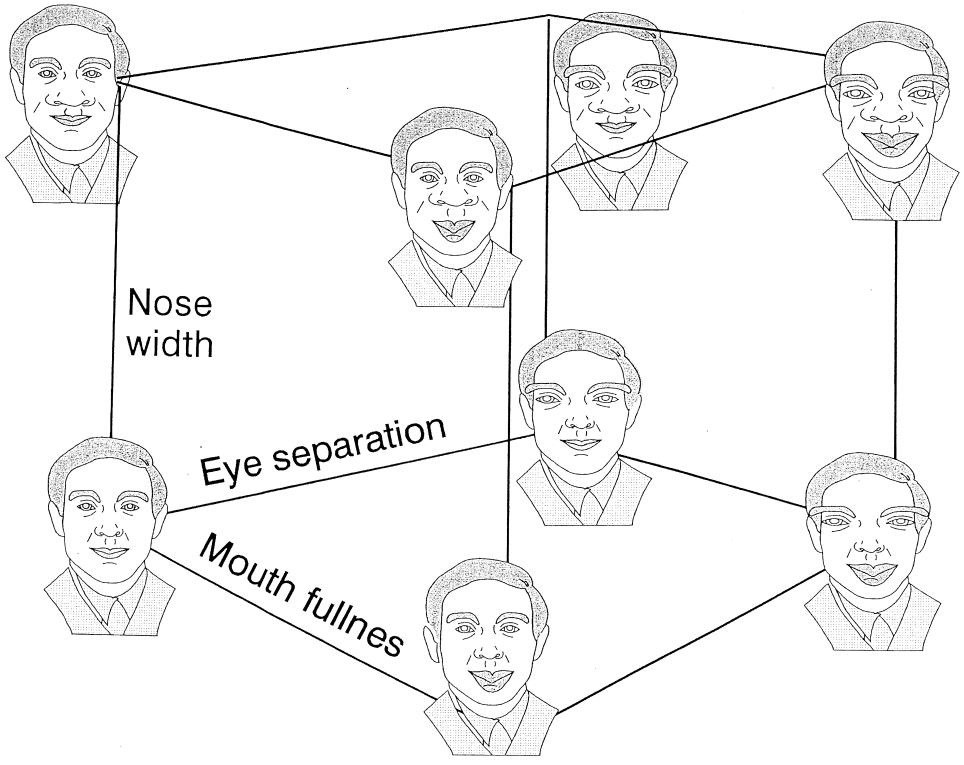


FIGURE 8. Vector coding. Three axes define an infinite number of faces.

and vectorial coding adapted from Churchland and Sejnowsky, 1992. It shows that an infinite number of faces can be defined by merely varying the intensity of the three basic characteristics, or ideals, that are depicted in the figure. To represent *n* ideals, an *n*-dimensional space would be required.

Another fact that leads us to believe that there is some kind of orthogonalization process taking place in the brain is that the orthogonalization of input information is a necessary condition to achieve a *hebbian* learning in pattern associator models (McClelland et al., 1986). It seems that the brain utilizes this kind of hebbian learning instead of complicated neural network algorithms.

Events in the brain occur by means of impulse rates, making the use of statistics in the present research seem convenient. But if orthogonalization is considered inside statistics it is necessary to redefine some important statistical notions, so that statistics can be seen as a branch of euclidean algebra. This topic is treated at the end of the paper to provide the reader with a general background without diving into abstract mathematics. Several results from that section will be utilized in the course of the paper.

On the basis of that section, nervous information will be treated as statistical events. For example, if the input layer is a matrix of *m* neurons, any of the 2^m subsets that could be accomplished with the *m* neurons will be considered an event. Every input would be related to the others in an auto-associative manner, as if all the inputs constituted a unique pattern that could be completed coherently if some of the inputs were missing.

Every input is defined as a vector of *n* components, *n* being the number of orthogonal neurons in the reticular neuron layer. Orthogonality can be considered as follows.

Two patterns, *A* and *B*, are orthogonal when the scalar product, defined as the number of coincidences of the two of them, is zero (see Section 9):

$$n(\vec{A} \cap \vec{B}) = 0$$

In other words, two patterns are orthogonal if the probability of their coincidence is zero. All the neurons y_i in the reticular layer could be considered orthogonal or mutually exclusive because there exists a competitive inhibition between them (when one neuron fires the others do not). Orthogonality in reticular neurons interactions allows these neurons to be considered as axes in a coordinate system.

According to these ideas, input *b* is defined as:

$$\vec{b} = k_1\vec{y}_1 + k_2\vec{y}_2 + \ldots + k_n\vec{y}_n$$

In this pattern $\vec{b}$, all the components are mutually exclusive and define a partition of the whole space. Every factor k_i is a flag that exhibits the partial contribution of the reticular activation $\vec{y}_i$, to the pattern $\vec{b}$; in other words, $k_1 = P(\vec{b}/\vec{y}_i)$. It is also possible to compute the probability of pattern $\vec{b}$ performing its scalar product

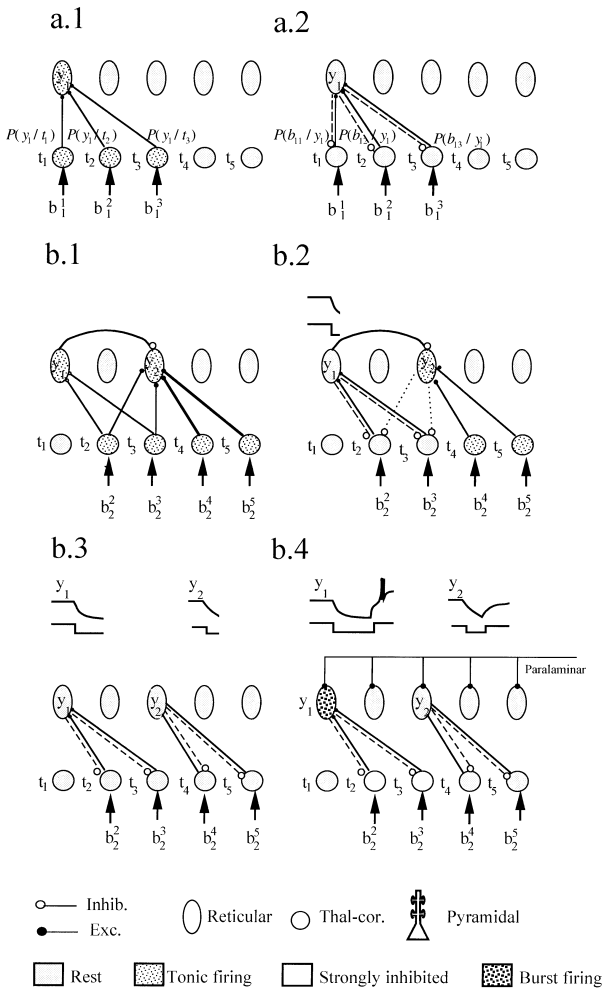


FIGURE 9. This figure illustrates the Gram–Schmidt (G–S) process of axis generation. Only active connections are shown. (a.1) Pattern b_1 is presented. Reticular neuron y_1 is selected through a competitive process of inhibitory neurons. Connections between thalamocortical (*t*) and reticular neurons (*y*) are reinforced. The first iteration of G–S is computed: $P(y_1) = P(b_1)$. (a.2) Feedback connections are trained using a process that is similar to a delta rule. (b.1) The second pattern b_2 is input. Candidate reticular neurons are selected. (b.2) Reinforced feedback connections of y_1 inhibit inputs B_2^2 and b_2^3. The remaining inputs activate y_2. The probability of y_2 is calculated as follows: $P(y_2) = P(b_2) - P(b_2/y_1)P(y_1)$, which is the second iteration of G–S. The activity of y_1 decreases because the activity of inputs also diminishes. This situation prepares y_1 to start firing in a burst mode if an excitation arrives from the paralaminar nucleus. (b.3) A little later, y_2 also diminishes its activity. (b.4) y_1 and y_2 commence firing in a burst mode due to an excitation from the paralaminar nucleus.

by the normalized diagonal vector $\vec{I}$ (see Section 9):

$$P(\vec{B}) = n(\vec{B} \cap \vec{I})$$

through which the usual expression of a vector in terms of its components is obtained:

$$P(\vec{b}) = P(\vec{b}/\vec{y}_1)P(\vec{y}_1) + P(\vec{b}/\vec{y}_2)P(\vec{y}_2) + \ldots + P(\vec{b}_n/\vec{y}_n)P(\vec{y}_n)$$

4. HOW IDEALS ARE OBTAINED FROM INPUT PATTERNS

According to the author's hypothesis, every new pattern that arrives at the thalamus determines, in part, the processing of subsequent patterns, and, conversely, is processed according to the organization defined by previous patterns. Similarly, in the Gram–Schmidt orthogonalization process, the first pattern will define the first axis of a coordinate system, with successive patterns helping to find the orthogonal basis over which all the vectors of a certain euclidean space can be defined.

If the generic scalar product in the Gram–Schmidt equation is substituted by the previously defined scalar product, then the next version of Gram–Schmidt is obtained (see Section 9):

$$\vec{y}_1 = \vec{b}_1$$

$$\vec{y}_{r+1} = \vec{b}_{r+1} - \sum_{i=1}^{r} P(\vec{b}_{r+1}/\vec{y}_i)\vec{y}_i$$

Calculating the probabilities of the obtained axes, the result is:

$$P(\vec{y}_1) = P(\vec{b}_1)$$

$$P(\vec{y}_{r+1}) = P(\vec{b}_{r+1}) - \sum_{i=1}^{r} P(\vec{b}_{r+1}/\vec{y}_i)P(\vec{y}_i)$$

Orthogonalization allows all the input patterns to be represented in a coordinate system. During the orthogonalization process thal-cor neurons and RE neurons hypothetically discharge in the tonic mode. This is corroborated by the fact that the presentation of new information is normally done in the awakened state in which the tonic mode is also more frequently detected.

In the orthogonalization process, thalamocortical neurons serve as inputs and reticular neurons serve as the axes over which the information is projected. For the sake of simplicity, Figure 9 depicts only the connection between these two types of neurons.

In this very simple example, an analysis is performed of the computation produced when the first two patterns are input. In the first pattern b_1 (Figure 9a.1 and 9a.2), the first three thalamic neurons t_1, t_2 and t_3 are activated with a certain probability that is considered to be the same in all of them. The second pattern presentation (Figure 9b.1) corresponds to the set b_2 that, in this example, is characterized by the activation of thalamic neurons t_2, t_3, t_4 and t_5. Inhibitory connections from reticular neurons are not modifiable but behave as if they were, according to the description of the glomerulus given in Section 2.3. The weight in the glomerulus is calculated as the probability of an input neuron *b* being

conditioned by a reticular neuron y, $P(b/y)$, so that the inhibition by y over the thalamic neurons is $P(b/y)P(y)$, as was discussed in Section 2.3.

When the first pattern $\vec{b}_1$ is presented, Figure 9a.1, a certain reticular neuron y_1 in the competitive reticular neuron layer wins the competition to represent the input. During presentation, the conditional probabilities $P(y_1/t_j)$ for each of the relay neurons t_{1j} are reinforced. After reinforcement, every time $\vec{b}_1$ is presented it will generate the activation of y_1, such that $P(y_1) = P(b_1)$, which is the first iteration in the Gram–Schmidt algorithm. The weights of the inhibitory backward projections $P(\vec{b}_j^{\,u}/\vec{y}_i)$ are also reinforced (u being each of the input units of the input pattern $\vec{b}_j$) (Figure a.2), and finally the thalamic neurons are silenced.

When the second input pattern $\vec{b}_2$, Figure 9b.1, is presented, thalamic neurons t_2 and t_3, that were active during $\vec{b}_1$ presentation, project towards y_1 through the previously reinforced synapses. Subsequently, the reinforced backward inhibition will cause t_2 and t_3 to rest, Figure 9b.2.

During the presentation of the second pattern $\vec{b}_2$ another reticular neuron y_2 wins the competition inside the RNT. It could be argued that the previously selected reticular neuron, y_1, should win the competition to represent the second input pattern because of the previously reinforced forward connections. However, reticular neurons possess threshold accommodation, a unique property among thalamic neurons (Mulle et al., 1986, p. 2140). In that paper it is shown that *tonic firing is always associated with a marked increase of the spike threshold*, as is demonstrated in an experiment in which the first spike was fired at -50 mV while the tenth spike reached threshold at -38 mV. This means that a continuously increasing input voltage is required for maintaining the spiking rate of reticular neurons. In the present case (during the presentation of the second input pattern b_2), y_1 diminishes its firing rate because inputs t_2 and t_3 go rapidly to rest. If the slope of the net input activation over y_2 is bigger than that over y_1, y_2 wins the competition in the reticular layer. The probability of reticular neuron y_2 being activated, which corresponds to the activation of inputs $\vec{b}_2^{\,4}$ and $\vec{b}_2^{\,5}$, is calculated as the probability of pattern $\vec{b}_2^{\,3}$ after the inhibition of inputs $\vec{b}_2^{\,2}$ and $\vec{b}_2^{\,3}$ from y_1:

$$P(\vec{y}_2) = P(\vec{b}_2) - P(\vec{b}_2/\vec{y}_1)P(\vec{y}_1)$$

which is the second iteration of the Gram–Schmidt equation. Inhibitory connections $P(\vec{b}_2^{\,2}/\vec{y}_2)$ and $P(\vec{b}_2^{\,3}/\vec{y}_2)$ from y_2 over t_2 and t_3 would also be reinforced if there were no inhibition by y_1 (activated by b_{22} and b_{23}) over y_2, which counterbalances this tendency.

y_2 now starts firing, and, as a result, the backward inhibitory connections to the thalamus are reinforced. This inhibition causes t_4 and t_5 to be at rest again; consequently, having no active inputs, y_2 decreases its activation, as occurred previously in the case of y_1 (see the upper part of Figure 9b.2). The sudden drop in input signals, to y_1 and subsequently to y_2, inactivates them gradually (Figure 9b.3), and the reticular neuron prepares to start firing in a bursting mode after a slight depolarization arriving from the paralaminar nucleus (the paralaminar neurons seem to burst when, being inhibited by reticular neurons, they receive excitation from 40 Hz intralaminar neurons, filtering in this manner the activity of the latter according to the state inside the thalamus). Although thalamocortical neurons should also be in the burst mode, brain stem reticular formation, through its cholinergic inputs, prevents thalamocortical neurons from bursting (McCormick & Prince, 1987a, b). On the contrary, these same cholinergic inputs facilitate the bursting mode in reticular neurons. After depolarization, the bursting mode commences earlier and with a stronger frequency in those neurons that were inactivated earlier or which had been further inactivated. In this case, in which y_1 and y_2 suffer a similar inactivation, y_1 changes into burst mode first because it was the first to be inactivated, Figure 9b.4. This bursting mode in reticular neurons represents the beginning of a new process for binding the different inputs and features that characterize a certain input pattern. This binding ensures that a certain pattern will be remembered as a whole, with inner coherence between the different constituent ideals and between the inputs that constitute every ideal. This topic will be discussed in Section 5. However, it is still necessary to remark further on the process explained above.

Once the selection of axes is achieved in a competitive neural network-like manner, the process of correlating inputs and outputs is a typical example of a hebbian delta rule (a good review of this work can be found in McClelland & Rumelhart, 1988, pp. 84–95) in which the input patterns are coded in the reticular neuron layer and the target patterns are those input to the thalamus. According to the delta rule, the weights in the network are proportional to the correlation between the inputs and a certain error computed in the output neurons. This error is equal to the difference between the target pattern and the actual network output. In the thalamus, this error is computed by thalamocortical neurons. In them the pattern of inhibitions by the reticular neurons over thalamocortical neurons corresponds to the output in the delta learning rule, while inputs from senses to the thalamus correspond to the target pattern. According to the delta rule, when the input patterns used in training are independent, the correct association with the target patterns is obtained, thereby avoiding contamination between the desired output and the output to other learned input patterns. In the present case, during the presentation of pattern $\vec{b}_1$, vector (1,0) that corresponds to reticular neuron y_1 activation was obtained. During

the presentation of $\vec{b}_2$, reticular neurons y_1 and y_2 were sequentially activated, giving rise to vector (1,1). Both vectors (1,0) and (1,1) constitute an independent input set. During each learning pattern presentation a new axis is obtained, but former axes are also recalled, giving rise to a new independent vector. These vectors are associated with each target pattern $\vec{b}_j$ through the delta rule. Moreover, when a testing pattern in RNT is a linear combination of the independent vectors obtained during learning, the same linear combination (in this case between the outputs) explains the output to the testing pattern. A certain independent vector in the reticular nucleus of the thalamus codifies a linear combination of the different features of an incoming pattern. Each component of this combination is stored, as will be explained in Section 5, in the connections from pyramidal neurons to reticular neurons.

5. BINDING

Input patterns to the thalamus are split into characteristic features or ideals, each one associated through a delta rule with an independent vector in the RNT. One precondition to preserve the identity of each pattern is to bind the activation of different neurons inside a certain feature and to bind the different features inside a pattern. Pyramidal neurons in the inferior temporal cortex provide an example of this type of binding. Lesions in this region produce prosopagnosia, in which patients are able to identify the parts of a face and even specific emotions expressed on it, but are incapable of identifying a person's identity. Identity can be defined as the concurrent activation of certain inputs and ideals at a specific instant. The concurrent activation of two input neurons b_1 and b_2, or in general of two parts of an image, can be measured as $P(\vec{b}_2 \cap \vec{b}_1)$. In this section it will be suggested that a pyramidal neuron can produce this computation, firing according to a certain combination of ideals which constitute the identity of a pattern.

Binding any two inputs b_1 and b_2 also means that the conditional probabilities $P(b_1/b_2)$ or $P(b_2/b_1)$ could be computed and stored in memory. In general, every input can share different ideals or, in other words, different ideals can be partially represented by a certain input. The only way by which the above computation can be accomplished is by projecting every pattern over the orthogonal axes previously found. The general expression for a conditional probability is:

$$P(\vec{b}_2/\vec{b}_1) = \frac{P(\vec{b}_2 \cap \vec{b}_1)}{P(\vec{b}_1)}$$

And, expressing every projection of b_j over the y_i axes $P(\vec{b}_j/\vec{y}_i)\vec{y}_i$ in the more compact $\vec{b}_{ji}$, the conditional probability expression of the two events b_2 and b_1 in terms of their n orthogonal components is derived (see Section 9):

$$P(\vec{b}_2/\vec{b}_1) = \frac{P(\vec{b}_1 \cap \vec{b}_2)}{P(\vec{b}_1)} = \frac{\sum_{i=1}^{n} P(\vec{b}_{1i} \cap \vec{b}_{2i})}{P(\vec{b}_1)}$$

The computation of this algorithm inside the thalamus will now be described using Figure 10. For this purpose, the explanation will first describe the learning process (having all the inputs) followed by the test process in which omitted inputs will be obtained through the presentation of the remaining inputs.

Figure 10 does not aim to be completely exhaustive but rather to facilitate the reader's understanding. Near some thalamic synapses the value of their weight is written, while the equation near the spine corresponds to the value of the projection of an incoming input over a certain axis.

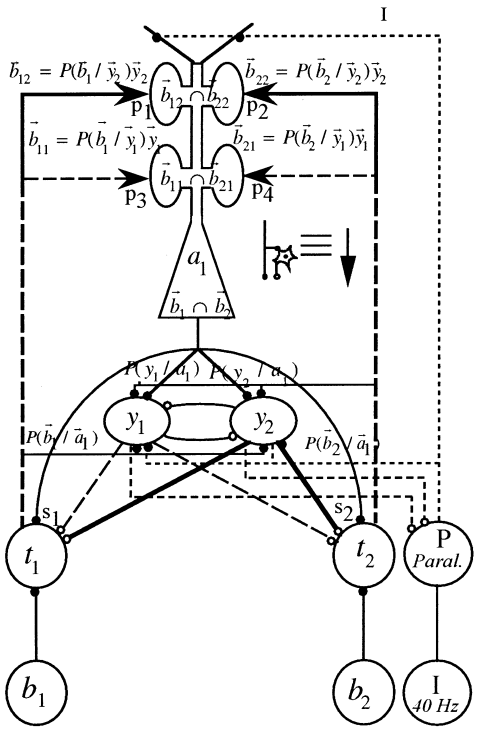
5.1. Learning Process

Incoming information promotes the activation of certain reticular neurons that constitute the axes over which the information is going to be projected. During activation, these axes inhibit thalamocortical neurons, completely counterbalancing their excitation through a sculpturing process explained in Section 4. For example, the probability of the excitation over t_1 from input b_1 in Figure 10 is exactly equal to the probability of the inhibition from y_1 and y_2:

$$P(\vec{b}_1) = P(\vec{b}_1/\vec{y}_1)P(\vec{y}_1) + P(\vec{b}_1/\vec{y}_2)P(\vec{y}_2)$$

which is the theorem of total probability. After this inhibition, thalamocortical neurons return to the resting level.

Reticular neurons are awakened through the Gram–Schmidt orthogonalization process. When all the selected reticular neurons simultaneously receive excitation from paralaminar neurons, the first inactivated is the first to change its tonic firing mode into burst firing (Llinas & Jahnsen, 1982). The remaining reticular neurons will produce burst firing in the order in which they were inactivated. In the present case of two reticular neurons, y_2 fires first (instant 4 in Figure 10), inhibiting deeply and abruptly all the thalamocortical neurons that, as we concluded, were at a resting level. This deep inhibition is necessary to counterbalance projections from the brain-stem reticular formation that try to keep thalamocortical neurons in the tonic mode. In this situation of deep hyperpolarization, if an input volley arrives to t_1 and t_2, they start bursting (instant 5). The bursts in t_1 and t_2 are then conditioned by the burst in y_2 (see also Section 2.4.1 and Figure 3b). Bursts in t_1 and t_2 due to y_2 correspond, respectively, to the projections of b_1 and b_2 over y_2, b_{12} and b_{22} (Figure 10, top). Therefore burst b_{12} (which occurs with probability $P(b_{12})$) and burst b_{22} pass beside spines p_3 and p_4, respectively (instant 6), without



LEARNING PROCESS b1 and b2 as inputs	axes selection y1 and y2				projection over y2				projection over y1				Binding	
At instant:	0	1	2	3	4	5	6	7 p_2 p_1 tuft	8	9	10 p_4 p_3	11	12	
the neuron	t_1 t_2	t_1 f.e.	t_1 f.e.	p	y_2	t_1 t_2	a_1	a_1	y_1	t_1 t_2	a_1	a_1	t_1 t_2	
receives excitation/ inhibition from:	b_1 b_2	y_2	y_1	y_1 y_2 I	p	y_2 b_2	$t_1 t_2$ p	t_1 t_2	p	y_1 b_1	t_1 t_2	—	a_1	
through its synapses	—	—	—	—	—	—	$p_3 p_4$ tuft resp.	p_1 p_2	—	—	p_3 p_4	—	$s_1 s_2$ res.	
with synaptic input value	—	V_3	V_2	—	—	—	V_3 V_5 res.	V_3 V_5 res.	—	—	V_2 V_4 res.	—	—	
and a net input	$V_0 V_1$ resp.	V_3	V_6	—	—	—	No sum	V_9	—	—	V_8	V_{10}	$V_0 V_1$ res.	
according to figure:	3.a	3.a	3.a	—	—	3.b	7.a	7.a	—	3.b	7.b	—	—	
and it will produce excitation/inhibition over :	y_1 y_2 a_1	y_1 y_2 a_1	y_1 y_2 a_1	y_1 y_2	t_1 t_2	a_1	Not yet	Not yet	t_1 t_2	a_1	No yet	y_1 y_2 $t_1 t_2$	—	
in tonic/bursting mode	T	T	T	B	B	B	—	—	B	B	—	B	—	
(this mode is due to a previous excitation/inhib.	E	E	E	—	I	I	—	—	I	I	—	—	—	
with a value:)	—	—	—	—	—	V_6	—	—	—	V_6	—	—	—	
Comments:	1.	—	2.	3	—	—	4.	—	—		—	—	—	

TESTING PROCESS only b1 as input	axes selection y1 and y2				projection over y2				projection over y1				Binding	
At instant:	0	1	2	3	4	5	6	7 p_2 p_1 tuft	8	9	10 p_4 p_3	11	12	
the neuron	t_1	t_1 f.e.	t_1 f.e.	p	y_2	t_1 t_2	a_1	a_1	y_1	t_1 t_2	a_1	a_1	t_1 t_2	
receives excitation/ inhibition from:	b_1	y_2	y_1	y_1 I y_2	p	y_2 b_1	t_1 p	t_1	p	y_1 b_1	t_1	—	a_1	
through its synapses	—	—	—	—	—	—	p_3 tuft resp.	p_1 p_2	—	—	p_3	—	$s_1 s_2$ res.	
with synaptic input value	—	V_3	V_2	—	—	—	V_3 res.	V_3 res.	—	—	V_2	—	—	
and a net input	$V_0 V_1$ resp.	V_3	V_6	—	—	—	No sum	V_9	—	—	V_8	V_{10}	$V_0 V_1$ res.	
according to figure:	3.a	3.a	3.a	—	—	3.b	7.a	7.a	—	3.b	7.b	—	—	
and it will produce excitation/inhibition over :	y_1 y_2 a_1	y_1 y_2 a_1	y_1 y_2 a_1	y_1 y_2	t_1 t_2	a_1	Not yet	No yet	t_1	a_1	No yet		—	
in tonic/bursting mode	T	T	T	B	B	B	—	—	B	B	—	B	—	
(this mode is due to a previous excitation/inhibition	E	E	E	—	I	I	—	—	I	I	—	—	—	
with a value:)	—	—	—	—	—	V_6	—	—	—	V_6	—	—	—	
Comments:	1.	—	2.	3.	—	—	4.	5.	—		6.	—		

$V_0 : P(\vec{b}_1)$ $V_1 : P(\vec{b}_2)$

$V_2 : P(\vec{b}_1 / \vec{y}_1)P(\vec{y}_1)$ $V_3 : P(\vec{b}_1 / \vec{y}_2)P(\vec{y}_2)$

$V_4 : P(\vec{b}_2 / \vec{y}_1)P(\vec{y}_1)$ $V_5 : P(\vec{b}_2 / \vec{y}_2)P(\vec{y}_2)$

$V_6 : P(\vec{b}_1 / \vec{y}_1)P(\vec{y}_1) + P(\vec{b}_1 / \vec{y}_2)P(\vec{y}_2) = P(\vec{b}_1)$

$V_7 : P(\vec{b}_2 / \vec{y}_1)P(\vec{y}_1) + P(\vec{b}_2 / \vec{y}_2)P(\vec{y}_2) = P(\vec{b}_2)$

$V_8 : P(\vec{b}_{11} \cap \vec{b}_{21})$ $V_9 : P(\vec{b}_{12} \cap \vec{b}_{22})$

$V_{10} : P(\vec{b}_1 \cap \vec{b}_2) = P[(\vec{b}_{11} \cap \vec{b}_{21}) + (\vec{b}_{12} \cap \vec{b}_{22})]$

Comments:

1/ Although thalamic neurons are firing over pyramidal neurons, we assume that tonic firing frequency is not strong enough to excite them.

2/ In this moment the probability of the inhibition over t_1 due to y_1 and y_2, equates the probability of the excitation from b1.

3/ Although pararalaminar thalamic neurons fire over y_1 and y_2 the neuron that has been inhibited more will produce a burst firing in first place. (Jahnsen,1984)

4/ The course of the bearer wave produced by the 40Hz neuron in the tuft is depicted above.

5/ Although there is no input from b_2, the net value is as if it were, because the weight in p1 is : $P(\vec{b}_{11} \cap \vec{b}_{21}) / P(\vec{b}_{11})$ and multiplying the weight by the input the result is $P(\vec{b}_{11} \cap \vec{b}_{21})$

6/ Analogous reason as 6.

FIGURE 10. This figure explains the process of selection of 'ideals' (axes) and the binding process of two inputs. These tables are intended to be read column by column from instant 1 to 12. Every column is designed to form a sentence: the first column will say "At instant 0 the neurons t_1 and t_2 that receive excitation/inhibition [in this case excitation] from b_1 and b_2 and a net input V_0, V_1 respectively...". Dashes must be skipped when reading. Other abbreviations: f.e. = for example; res. = respectively. Data that is squared in both tables must be compared carefully to understand the difference between the learning and the testing process.

producing any effect, as explained in Figure 7a and Figure 4e; but, when arriving at spines p_1 and p_2 (instant 7), their simultaneous coincidence (their intersection) boosts the 40 Hz response that has travelled up to this point through the dendritic shaft (see also Figure 7b). At this moment bursts b_{12} and b_{22} also reinforce the synapses in the spine (the tonic mode firing of the thalamocortical neuron was not strong enough to allow reinforcement during the previous process of axis selection as seen in Figure 4b). While the consequences of the burst of y_2 reticular neuron evolve, the burst mode of y_1 produces its first barrage, inhibiting strongly the same thalamocortical neurons t_1 and t_2 (instant 8) and thereby producing the projections of b_1 and b_2 over axis y_1 (that are led through the path depicted as broken lines in the figure). These projections correspond to the second pair of travelling bursts b_{11} and b_{21}. The intersection of these two bursts, when arriving at spines p_3 and p_4, boosts the same bearer wave that was previously boosted by the intersection of b_{12} and b_{22} (see also Figure 7c). The bearer wave then conveys the information of the sum of the intersection of projections over y_1 and y_2. This sum defines the intersections of the inputs b_1 and b_2:

$$P(\vec{a}_1) = P(\vec{b}_1 \cap \vec{b}_2) = P[(\vec{b}_{11} \cap \vec{b}_{21}) + (\vec{b}_{12} \cap \vec{b}_{22})]$$

The concurrent activation of both inputs is what defines the identity of the input pattern as a whole. This means that a pyramidal neuron only responds to a certain combination of inputs, firing concurrently each of them with a specific probability. The activity of each pyramidal neuron expresses the inner coherence that defines a certain input pattern, as would a signature. The activation of the pyramidal neuron represents the input pattern itself, or the coherent characteristics that define it, without considering other circumstantial details. For example, in the identification of a face neither the length of the hair nor the type of ear rings or the luminosity will be taken into consideration. In this sense, the output of the pyramidal neuron can be understood as a transformation of the input pattern through a coherence function over pattern b that will be in the notation b^*:

$$a = b^*$$

5.2. Testing Process

During the testing process only b_1 is presented. As b_1 is considered to share axes y_1 and y_2, both are alternately activated. When y_1 and y_2 burst over thalamocortical t_2, burst firing does not result since t_2 is without the excitatory input that is necessary to deactivate t_2 and start bursting. Only the projections of the unique input b_1 over y_2 and y_1, b_{11} and b_{12} respectively, approach the spines sequentially. When arriving at the spines, and multiplied by the weights, they produce the same result as if b_{22} and b_{21} had also arrived at the spines. In the case of spine p_3, if the probability of the input to the spine

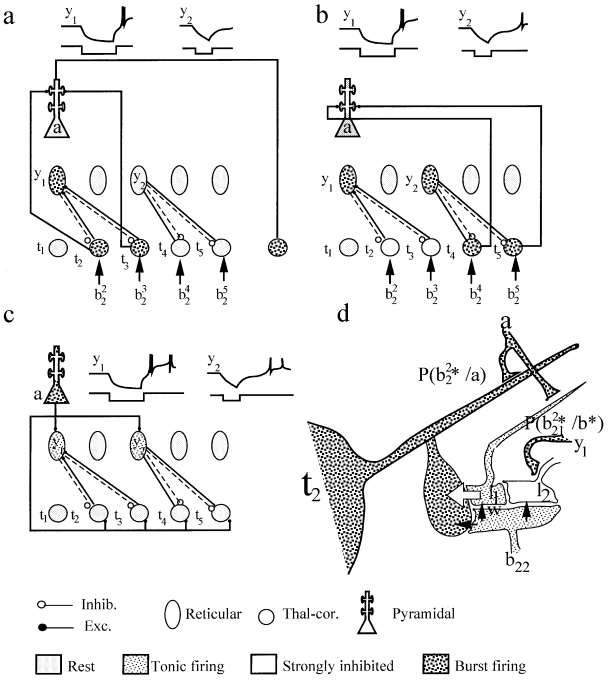


FIGURE 11. The second barrage of the burst in reticular cells seems to play a specific role in the reinforcement of backward connections from the cortex. (a) Bursts of y_1 hyperpolarize t_2 and t_3, creating in them the conditions for burst firing that commences after an excitation from the paralaminar nucleus. (b) y_2 also facilitates the burst mode in t_4 and t_5. (c) Corticothalamic connections are reinforced, silencing incoming inputs. (d) Weights from pyramidal to thal-cor are calculated as $P(b_2^{2*}/a)$ and from pyramidal to local circuit cell I_1 as $P(b_{21}^{2*}/b^*)$.

$P(\vec{b}_{11}) = P(\vec{b}_1/\vec{y}_1)P(\vec{y}_1)$ is multiplied by the weight $P(\vec{b}_{11} \cap \vec{b}_{21})/P(\vec{b}_{11})$ (that was computed in the spine during the learning process) the result is $P(\vec{b}_{11} \cap \vec{b}_{21})$, as if the other input b_{21} had just arrived at spine p_4. In the case of spine p_1, the result is $P(\vec{b}_{12} \cap \vec{b}_{22})$, as if there were a stimulus in spine p_2. These two fictitious intersections produce, however, the firing of the pyramidal neuron that encodes the intersection of inputs b_1 and b_2:

$$P(\vec{a}_1) = P(\vec{b}_1 \cap \vec{b}_2) = P[(\vec{b}_{11} \cap \vec{b}_{21}) + (\vec{b}_{12} \cap \vec{b}_{22})]$$

This result is the same as that of applying a certain conditional probability $P(\vec{b}_2/\vec{b}_1)$ over the input pattern $\vec{b}_1$ which occurs with probability $P(\vec{b}_1)$) so that

$$P(\vec{b}_1)P(\vec{b}_2/\vec{b}_1) = \sum_{i=1}^{n} P(\vec{b}_{2i} \cap \vec{b}_{1i}) = P(\vec{b}_2 \cap \vec{b}_1)$$

In this way the expression of a conditional probability is identified in the computation between thalamus and cortex, which was the purpose of this section.

When the probability of pyramidal neuron a_1 is multiplied by the weight between thalamocortical neuron t_2 and pyramidal neuron a_1, which is $P(\vec{b}_2/\vec{a}_1) = 1$, the result is the excitation over t_2:

$$P(\vec{a}_1) \cdot P(\vec{b}_2/\vec{a}_1) = P(\vec{a}_1) = P(\vec{b}_1 \cap \vec{b}_2)$$

This is the same result as that of multiplying $\vec{b}_1$ by a conditioned probability $P(\vec{b}_2/\vec{b}_1)$ stored in the network. In this case, the pattern $\vec{b}_2$ is reconstructed, but only to the extent of spiking with a probability $P(\vec{b}_1 \cap \vec{b}_2)$ which is not the original probability $P(\vec{b}_2)$. However, this probability defines the contribution of t_2 to the identification of the input pattern.

Pyramidal neuron a_1 also fires in a burst mode over y_1 and y_2. When doing this, the conditional probabilities $P(y_1/a_1)$ and $P(y_2/a_1)$ are stored in its synapses. These probabilities represent the extent to which a certain feature or ideal contributes to the identification of an input pattern. As mentioned at the end of Section 4, one of the characteristics of a neural network trained through a delta rule is that it is possible to obtain an output pattern that is a blend of several features, each of which is associated with an orthogonal input, if this same blend of orthogonal inputs is introduced to the network. In this case the orthogonal inputs y_i are combined as follows:

$$\vec{b}^{*} = P(\vec{b}^{*}/\vec{y}_1)\vec{y}_1 + P(\vec{b}^{*}/\vec{y}_2)\vec{y}_2$$

This combination of activations of reticular neurons produces an output pattern of inhibitions over thalamocortical neurons (through the weights stored during the Gram–Schmidt process, $P(b_j/y_i)$) that reproduces pattern $\vec{b}^{*}$. Therefore the sculpturing inhibition of reticular neurons over thalamocortical ones is accomplished through an ideal image of the input pattern $\vec{b}$.

Some remarks on the adjustment of backward connections from pyramidal neurons to thalamic neurons are still necessary. For this purpose the example given in Section 4 will be continued. In the upper part of Figure 11c, the typical response of a reticular neuron to an input consisting of a depolarization, a hyperpolarization and another depolarization is shown. The response to a slight excitation after hyperpolarization is composed of a burst with a rebound around 40 ms later. While the first reticular burst produces the hyperpolarization of thalamic neurons so that the projections of the input patterns over the reticular axes y_1 and y_2 can be transferred to the cortex (Figure 11a and 11b), the second burst seems to play a special role over the reinforcement of the connections from the pyramidal neuron to the reticular and thalamocortical neurons (Figure 11c). It seems as if the rebound waited until the pyramidal neuron sent its message back to the thalamus so that both stimuli could be related through hebbian reinforcement. The process of this reinforcement is the following: the bursting reticular neuron y_1 inhibiting cell l_2 (Figure 11d) allows for l_1 to inhibit thalamocortical cell t with a probability $P(\vec{b}_2^{2}/\vec{y}_1)\cdot P(\vec{y}_1)$ (see also Figure 2). If the influence of all the inhibitions over the thalamocortical neuron (in the example from y_1 and y_2) is computed, then the probability of the whole inhibition

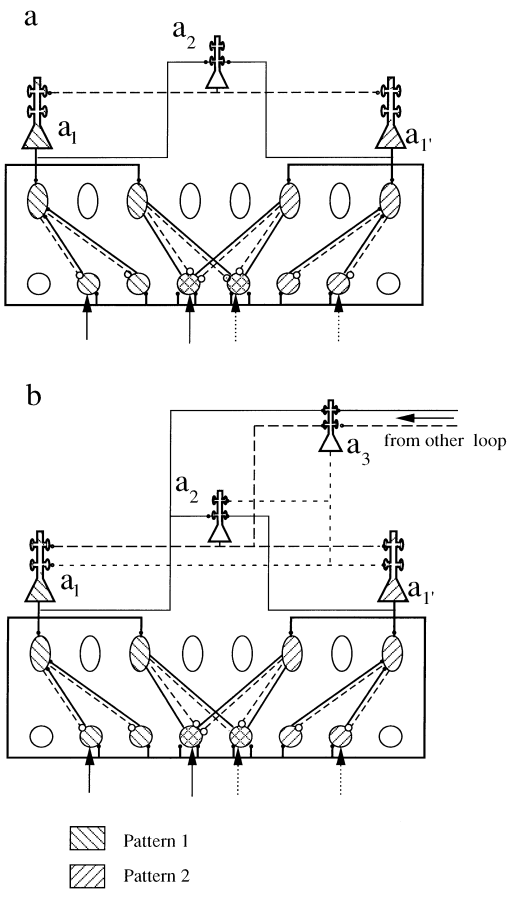


FIGURE 12. (a) Two pyramidal neurons a_1 and $a_{1'}$, each one representing a certain coherent input pattern, are related by means of a third pyramidal neuron a_2. The reverberant activity of the two pyramidal neuron layers in the figure ensures the maintenance of corticothalamic projections that complete (lined areas) partial input patterns (arrows). These loops of reverberant activity tend to vanish because pyramidal neurons recall surrounding basket neurons that inhibit them. This fading is counterbalanced if pyramidal neurons of higher layers are recruited (b) so that these loops become wider. Connections between layers are the same as between the neurons depicted in Figure 6a. their activity being regulated by the same rules of sequentiality. Higher layer neurons are able to participate in several reverberant loops (comprising lower layer neurons), correlating them. The highest layer neurons cannot be correlated by upper neurons but form a large loop that is hypothesized to be the hippocampal formation.

equals that from the stimulus $\vec{b}_2^{2*}$:

$$P(\vec{b}_2^{2*}) = P(\vec{b}_2^{2}/\vec{y}_1)P(\vec{y}_1) + P(\vec{b}_2^{2}/\vec{y}_2)P(\vec{y}_2)$$

Depending on the strength of this inhibition over t_2, a proportional bursting frequency is produced whithin it. For example, the postsynaptic burst in t_2 with probability $P(b^{*}_{22})$, combined with a presynaptic burst from pyramidal neuron 'a' with frequency $P(b^{*})$, produces the synaptic weight $P(\vec{b}_2^{2*}/\vec{b}^{*})$ in their synapse. The synaptic weight between a and l_1 is $P(\vec{b}_{21}^{2*}/\vec{b}^{*})$. This second type of connection ensures that when there is no inhibition over t_2 from reticular neurons, the same

inhibition is produced over t_2 to allow it to enter into the bursting mode, in this case promoted by a pyramidal neuron (the reticular inhibition is absent over t_2 if $\vec{b}_2^2$ is not present). Another interesting effect is that when the thalamocortical loop is reinforced, pyramidal neurons mute sensorial inputs hypothetically through l_1 (Sillito et al., 1993); see Figure 11c, in which the input arrows are omitted.

6. TRANSIENT MEMORIES MAINTAINED THROUGH THE ACTIVATION OF CORTICAL LOOPS

In Section 5 was proposed a mechanism by which pyramidal neurons that represent a certain coherent pattern can reconstruct a partial one by means of the feedback stimulation of thalamocortical neurons. It was also explained that pyramidal neurons can mute sensorial inputs so that, transiently, the input to the thalamus would be *b** (from cortex) instead of *b* (from senses).

Figure 12a shows that two pyramidal neurons a_1 and $a_{1'}$, each of which represents a certain unitary content or identity, can also be related through a third pyramidal neuron a_2 if both contents are simultaneous. In the example, pyramidal neurons a_1 and $a_{1'}$ were recalled, respectively, by a fragmented presentation of two patterns depicted as continuous and dotted arrows respectively. The activation of pyramidal neurons a_1 and $a_{1'}$ complete the patterns 1 and 2 over thalamocortical neurons, muting actual inputs. At the same time, the concurrent activation of a_1 and $a_{1'}$ over a_2 reinforces synapses in a_2, as explained in Section 2.4.3 and Figure 4c. Reciprocally pyramidal neuron a_2, firing over the previously activated a_1 and $a_{1'}$, activates the backward path. If only the inputs represented by the continuous arrows were present, both input patterns would be reconstructed, because only one input to a_2 produces the same response as if both inputs were present (see Section 5.2). These reconstructed patterns would be only evanescent images if the inputs completely disappeared. However, these memories are maintained active in the reinforced loop of pyramidal neurons when a_1 and $a_{1'}$ produce the shot of a_2 and this neuron fires again over a_1 and $a_{1'}$. Despite this, these reverberant loops tend to desappear because of basket cell inhibition unless pyramidal neurons recruit higher level neurons (Figure 12b).

In this latter case neuron a_3 is recruited and is involved in a circuit in which the three layers of neurons fire in a cycle. This cycle is the same as that previously shown in Figure 6a. The activity of pyramidal neuron a_3 not only contributes to the maintainance of this loop but also to the engagement of other loops. The pyramidal neurons of higher layers engage wider loops. It is assumed here that the hippocampus is hypothetically capable of coordinating inside other loop the activity of neurons in the highest layer.

7. EXPERIMENTAL RESULTS

A simplified version of the model described above has been implemented in a computer program. It comprises thirty modules with a thalamocortical, a reticular, a pyramidal and a basket neuron in each module. A two hemispheres version with callosal neurons connecting both of them was also accomplished. The following conclusions should be noted.

1. The model learns non-orthogonal patterns.
2. Eliminating or modifying connections confirms the previously described properties of different kind of neurons.
3. After learning, when inputs are omitted and the neurons' threshold is decreased, the model continues its inner dynamic mixing of previously learned patterns in a continuous way.

8. DISCUSSION

The hypothesis of thalamus and cortex computation just presented allows incoming patterns to be considered as vectors that are unequivocally characterized by their projections over a coordinate system. The axes in this basis (the 'ideals' in Plato's theory) are represented by the firing in each reticular thalamic neuron. These ideals are extracted from input patterns through a Gram–Schmidt orthogonalization process that is accomplished in the thalamus. The association between an active reticular neuron and the feature or ideal that it defines is achieved through a delta-rule-like process in which dendritic glomeruli have a special role. Binding the different ideals that are present in a certain input pattern is achieved in pyramidal neurons. Since Gram–Schmidt selection of axes is an iterative process accomplished over time, the exact temporal sequence of axes or ideals for a determined pattern is crucial, and this sequence is stored like a certain sequence of inputs in layer IV pyramidal apical dendrites.

This process also allows the binding of the input neurons that give rise to a certain ideal, and of the different parts of an input pattern. When a certain pyramidal neuron shoots, it produces an ideal image of the input pattern over thalamocortical neurons. If only a part of the input pattern is presented, it is fully reconstructed through the projections from pyramidal to thalamic neurons. The firing of pyramidal neurons also represents a measure of the coherence of a certain pattern. Several coherent patterns represented by different pyramidal neurons are related through the activation of pyramidal neurons of higher layers. Cycling loops of neurons in different cortical layers maintains the integrity of memories, and higher layer neurons engage lower layer cycling loops. Hypothetically, the hippocampus would engage the activity of pyramidal neurons in the highest layers, giving a unified representation of current sensorial events.

9. MATHEMATICAL FOUNDATIONS

This section is devoted to reviewing the mathematical background of the model. Each of the 2^m subsets that can be formed with the m input neurons defines a certain pattern. Patterns can be represented as elements of a linear n-dimensional space that will be referred to as Y, with addition and multiplication defined in the usual way. In this linear space of patterns Y, a factor k multiplied by a certain pattern y indicates the extent to which this pattern collaborates in the identification of another pattern b ($b = ky + ...$). In Y, a negative pattern is that which, when added to its opposite, annihilates it (excitation and inhibition are an example in neural systems). Taking this into account, Y accomplishes all the properties that define a linear or euclidean space.

In Y an inner product $(\vec{A},\ \vec{B})$ can be defined as:

$$n(\vec{A} \cap \vec{B})$$

which represents how many times the coincidence of two patterns takes place. This inner product shares the properties of every inner product:

1. Commutative: $n(\vec{A} \cap \vec{B}) = n(\vec{B} \cap \vec{A})$;
2. Distributive: $n(\vec{A} \cap (\vec{B} + \vec{C})) = n(\vec{A} \cap \vec{B}) + n(\vec{A} \cap \vec{C})$;
3. Associative: $n(\vec{A} \cap k\vec{B}) = k \cdot n(\vec{A} \cap \vec{B})$ (for $k < 0$ this scalar product is a measure of opposition rather than coincidence);
4. Positivity: $n(\vec{A} \cap \vec{A}) \geq 0$.

This scalar product makes it possible to define orthogonality between two patterns as:

$$n(\vec{A} \cap \vec{B}) = 0$$

Y becomes a *euclidean space* because it possesses the above inner product. In a generic euclidean space, a vector B is written in terms of its projections over an orthogonal basis as follows:

$$\vec{B} = \sum_{i=1}^{n} \frac{(\vec{B},\ \vec{y}_i)}{(\vec{y}_i,\ \vec{y}_i)} \vec{y}_i$$

Substituting the new scalar product $n(\vec{A} \cap \vec{B})$ in the above expression yields

$$\vec{B} = \sum_{i=1}^{n} \left[\frac{n\left(\vec{B} \cap \vec{y}_i\right)}{n\left(\vec{y}_i \cap \vec{y}_i\right)} \right] \cdot \vec{y}_i$$

As all axes y_i are orthogonal they can be considered as mutally exclusive events from a statistical point of view because $n(\vec{y}_i \cap \vec{y}_j) = 0 \Rightarrow P(\vec{y}_i \cap \vec{y}_j) = 0,\ \ \forall i \neq j$.

It can also be considered that the sum of the probabilities of all the axes in Y is equal to 1, thereby defining a partition of the sample space.

Taking this into account, the above expression of a pattern B in terms of its components can be rewritten as shown below, taking into account that the part within brackets is equivalent to a conditional probability

$$\vec{B} = \sum_{i=1}^{n} \left[P(\vec{B}/\vec{y}_i) \right] \vec{y}$$

Let us define a diagonal vector $\vec{i} = (\vec{y}_1,\ \vec{y}_2,\ ...\ \vec{y}_n)$. Dividing $\vec{i}$ by the *sum norm*, $\vec{i}^2 = n(\vec{i} \cap \vec{i}) = n(\vec{y}_1) + n(\vec{y}_2) + ,\ ...\ n(\vec{y}_n)$, a normalised vector $\vec{I}$ is obtained:

$$\vec{I} = \frac{\vec{i}}{\vec{i}^2}$$

The probability of event $\vec{B}$ in terms of its components can be defined as the scalar product $P(\vec{B}) = n(\vec{B} \cap \vec{I})$:

$$\begin{aligned} P(\vec{B}) = n(\vec{B} \cap \vec{I}) &= \sum_{i=1}^{n} \left[P(\vec{B}/\vec{y}_i) \cdot \frac{n(\vec{y}_i \cap \vec{y}_i)}{i^2} \right] \\ &= \sum_{i=1}^{n} \left[P(\vec{B}/\vec{y}_i) \cdot \frac{n(\vec{y}_i)}{n(\vec{y}_1) + n(\vec{y}_2) + ,\ ...\ n(\vec{y}_n)} \right] \\ &= \sum_{i=1}^{n} \left[P(\vec{B}/\vec{y}_i) P(\vec{y}_i) \right] \end{aligned}$$

making it possible to apply the probability as if it were a linear operator over pattern $\vec{B}$ and its $\vec{y}_i$ components:

$$P(\vec{B}) = P\left(\sum_{i=1}^{n} P\left(\vec{B}/\vec{y}_i\right) \cdot (\vec{y}_i) \right) = \sum_{i=1}^{n} P\left(\vec{B}/\vec{y}_i\right) \cdot P(\vec{y}_i)$$

The above expression helps to clarify the meaning of the components in the euclidean space Y. Every factor can be understood as a conditional probability: it shows the probability of '$\vec{B}$' having the additional information of '$\vec{y}_i$'. This is consistent with our previous interpretation of a factor that is multiplied by a pattern as a measure of the extent to which a certain pattern, $\vec{y}_i$, collaborates or prejudices (if the factor were negative) in the identification of another one (in this case $\vec{B}$).

It was suggested that the thalamic model orthogonalizes the afferent information. A basis of orthogonal patterns $\vec{y}_i$ can be obtained from other non-orthogonal ones by substituting in the Gram–Schmidt equation the previously defined inner product for the generic inner products:

$$\vec{y}_1 = \vec{b}_1$$

$$\vec{y}_{r+1} = \vec{b}_{r+1} - \sum_{i=1}^{r} \frac{n(\vec{b}_{r+1} \cap \vec{y}_i)}{n(\vec{y}_i \cap \vec{y}_i)} \vec{y}_i$$

The probability of each axis is obtained by applying the above new definition of probability over each member of the Gram–Schmidt equations:

$$P(\vec{y}_1) = P(\vec{b}_1)$$

$$P(\vec{y}_{r+1}) = P(\vec{b}_{r+1}) - \sum_{i=1}^{r} P(\vec{b}_{r+1}/\vec{y}_i) P(\vec{y}_i)$$

It is also possible to calculate the intersection of two

patterns $\vec{b}_1$ and $\vec{b}_2$ by substituting for each of them the sum of their projections over the previously obtained orthogonal axes:

$$\vec{b}_1 \cap \vec{b}_2 = \sum_{i=1}^{n} \left[P(\vec{b}_1/\vec{y}_i)\cdot\vec{y}_i\right] \cap \sum_{j=1}^{n} \left[P(\vec{b}_2/\vec{y}_j)\right]\cdot\vec{y}_j$$

When $\vec{b}_1$ and $\vec{b}_2$ are projected over a generic axis $\vec{y}_i$ the resulting projections $\vec{b}_{1i}$ and $\vec{b}_{2i}$ can be expressed as:

$$\vec{b}_{1i} = P(\vec{b}_1/\vec{y}_i)\cdot\vec{y}_i$$

$$\vec{b}_{2i} = P(\vec{b}_2/\vec{y}_i)\cdot\vec{y}_i$$

Thus the above scalar product can be written as:

$$\vec{b}_1 \cap \vec{b}_2 = \sum_{i=1}^{n} \vec{b}_{1i} \cap \sum_{j=1}^{n} \vec{b}_{2j}$$

but as there is no intersection, $\forall i \neq j$

$$\vec{b}_1 \cap \vec{b}_2 = \sum_{i=1}^{n} (\vec{b}_{1i} \cap \vec{b}_{2i})$$

The probability of the intersection of $\vec{b}_1$ and $\vec{b}_2$ is calculated through the scalar product of this intersection and vector $\vec{I}$:

$$\begin{aligned}
P(\vec{b}_1 \cap \vec{b}_2) &= n\left[(\vec{b}_1 \cap \vec{b}_2) \cap \vec{I}\right] = n\left[\sum_{i=1}^{n} (\vec{b}_{1i} \cap \vec{b}_{2i}) \cap \vec{I}\right] \\
&= \vec{n}\left[(\vec{b}_{11} \cap \vec{b}_{21}) \cap \frac{\vec{y}_1}{i^2}\right] + n\left[(\vec{b}_{12} \cap \vec{b}_{22}) \cap \frac{\vec{y}_2}{i^2}\right] \\
&\quad + \ \ldots \ + n\left[(\vec{b}_{1n} \cap \vec{b}_{2n}) \cap \frac{\vec{y}_n}{i^2}\right] \\
&= n\left[\left(P(\vec{b}_1/\vec{y}_1)\vec{y}_1 \cap P(\vec{b}_2/\vec{y}_1)\right) \cap \frac{\vec{y}_1}{n(\vec{i})}\right] \\
&\quad + \ldots + n\left[\left(P(\vec{b}_1/\vec{y}_n)\vec{y}_n \cap P(\vec{b}_2/\vec{y}_n)\vec{y}_n\right) \cap \frac{\vec{y}_n}{n(\vec{i})}\right] \\
&= \sum_{i=1}^{n} P(\vec{b}_{1i} \cap \vec{b}_{2i})
\end{aligned}$$

This expression yields either

$$P(\vec{b}_1 \cap \vec{b}_2) = \sum_{i=1}^{n} P(\vec{b}_{1i} \cap \vec{b}_{2i})$$

or

$$\begin{aligned}
P(\vec{b}_1 \cap \vec{b}_2) &= \sum_{i=1}^{n} P\left((\vec{b}_1 \cap \vec{b}_2)/\vec{y}_i\right)\cdot P(\vec{y}_i) \\
&= \sum_{i=1}^{n} P(\vec{b}_1/\vec{y}_i)\cdot P(\vec{b}_2/y_i)\cdot P(\vec{y}_i)
\end{aligned}$$

so that

$$P\left((\vec{b}_1 \cap \vec{b}_2)/\vec{y}_i\right) = P(\vec{b}_1/\vec{y}_i)\cdot P(\vec{b}_2/\vec{y}_i)$$

A useful expression of the conditional probability of a certain pattern $\vec{b}_2$ due to another one $\vec{b}_1$ is also derived from those above:

$$P(\vec{b}_2/\vec{b}_1) = \frac{P(\vec{b}_1 \cap \vec{b}_2)}{P(\vec{b}_1)} = \frac{\sum_{i=1}^{n} P(\vec{b}_{1i} \cap \vec{b}_{2i})}{P(\vec{b}_1)}$$

REFERENCES

Artola, A., Brocher, S., & Singer, W. (1990). Different voltage-dependent threshold for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature, 347,* 69–72.

Churchland, P. S. & Sejnowsky, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Contreras, D., Destexhe, A., Sejnowski, T.J., & Steriade, M. (1995). Synchronization of thalamic spindle oscillation is enhanced by cortical inputs. *Society for Neuroscience Abstracts, 21,* 1187.

DeFelipe, J., & Fariñas, I. (1992). The pyramidal neuron of the cerebral cortex. Morphological and chemical characteristics of the synaptic inputs. *Progress in Neurobiology, 39,* 563–607.

Desmond, N.L., & Levy, W.B. (1983). Synaptic correlates of associative potentiation/depression: an ultrastructural study in the hippocampus. *Brain Research, 265,* 21–30.

DeValois, L. R. & DeValois, K. K. (1988). *Spatial vision.* Oxford University Press.

Huntley, G.W., Vickers, J.C., & Morrison, J.H. (1994). Cellular and synaptic localizaton of NMDA and non-NMDA receptor subunits in neocortex: organizational features related to cortical circuitry, function and disease. *Trends in Neuroscience, 17*(12), 536–542.

König, P., Engel, A.K., & Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revised. *Trends in Neuroscience, 19*(4), 130–137.

Lester, R.A.J., Clements, J.D., Westbrook, G.L., & Jahr, C.E. (1990). Channel kinetics determine the time course of NMDA receptor-mediated synaptic currents. *Nature, 346,* 565–566.

Llinas, R.R. (1988). The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science, 242,* 1654–1664.

Llinas, R., & Jahnsen, H. (1982). Electrophysiology of mammalian thalamic neurons in vitro. *Nature, 297,* 406–408.

Llinas, R.R., Grace, A.A., & Yarom, Y. (1991). In vitro neurons in mammalian cortical layer 4 exhibit intrinsic oscillatory activity in the 10- to 50-Hz frequency range. *Proceedings of the National Academy of Sciences, 88,* 897–901.

Llinas, R., Ribary, U., Joliot, M., & Wang, X. J. (1994). Content and context in temporal thalamocortical binding. In G. Buzsaki et al. (Eds.), *Temporal coding in the brain* (pp. 151–172). Berlin: Springer-Verlag.

McClelland, J. L. & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing.* Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E. and PDP Research Group (1986). *Parallel distributed processing: Exploration in the microstructure of cognition*. Cambridge, MA: MIT Press.

McCormick, D. A., & Prince, D. A. (1987a). Actions of acetylcholine in the guinea-pig and cat medial and lateral geniculate nuclei, in vitro. *Journal of Physiology (London),* **392**, 147–165.

McCormick D. A., & Prince, D. A. (1987b). Acetylcholine causes rapid nicotinic excitation in the medial habenular nucleus of guinea-pig, in vitro. *Journal of Neurosciences,* **7**, 742–752.

Mulle, C., Madariaga, A.M., & Deschenes, M. (1986). Morphology and electrophysiological properties of reticularis thalami neurons in cat:

in vivo study of a thalamic pacemaker. *Journal of Neurosciences*, *6*(8), 2134–2145.

Pinault, D., & Deschenes, M. (1992). Voltage dependent 40-Hz oscillations in rat reticular thalamic neurons in vitro. *Neuroscience*, *51*, 245–258.

Ropero Pelaez, J. (1996a). A formal representation of thalamus and cortex computation. In R. Moreno-Diaz & J. Mira-Mira (Eds.), *Brain processes, theories, and models: an international conference in honor of W. S. McCulloch 25 years after his death* (pp. 154–163). Cambridge, MA: MIT Press.

Ropero Pelaez, J. (1996b). A model of information processing and learning through thalamus and cortex neural connectivity and their neuron specific properties. In Institute of Electrical Engineers of Japan and Muroran Institute of Technology (Eds.), *Proceedings of International Workshop on Soft Computing in Industry '96* (pp. 139–147). Japan: Muroran.

Sillito, A., Cudeiro, J., & Murphy, P. (1993). Orientation sensitive elements in the corticofugal system from visual cortex to lateral geniculate nucleus in the cat (with an appendix on geniculo-cortical monosynaptic connections). *Experimental Brain Research*, *93*, 6–16.

Steriade, M., & Llinas, R. R. (1988). The functional state of the thalamus and the associated neuronal interplay. *Physiology Review*, *68*(3), 649–739.

Steriade, M., Curró Dossy, R., & Contreras, F. (1993). Electrophysiological properties of intralaminar thalamocortical cells discharging rhythmic (40 Hz) spike-bursts at 1000 Hz during waking and rapid eye movement sleep. *Neuroscience*, *59*, 21–31.

Stuart, G. J., & Sakmann, B. (1994). Active propagation of somatic action potential into neocortical pyramidal cells dendrite. *Nature*, *367*, 69–72.

Thomson, A. M., & Deuchars, J. (1994). Temporal and spatial properties of local circuits in neocortex. *Trends in Neuroscience*, *17*(3), 119–126.

Yuste, R., & Denk, W. (1995). Dendritic spines as basic functional units of neuronal integration. *Nature*, *375*, 682–684.

Yuste, R., & Tank, D. W. (1996). Dendritic integration in mammalian neurons, a century after Cajal. *Neuron*, *16*, 701–716

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1289–1302, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00061-0

1997 SPECIAL ISSUE

Metastable Associative Network Models of Dream Sleep

M. NAKAO,[1] I. HONDA,[1] M. MUSILA[2] AND M. YAMAMOTO[1]

[1]Tohoku University and [2]Third Medical School of Charles University

(*Received* 10 *July* 1996; *accepted* 13 *May* 1997)

Abstract—*Up to the present day, simulations using a neural network model have been carried out under the global inhibition and the uncorrelated noise in order to simulate the dynamics transition of neuronal activities from the flat (slow wave sleep) to the 1/f (dream sleep) power spectral density profile during the sleep cycle in cats. In particular, the metastability of the network attractor is shown to be essential in generating the 1/f fluctuations. Here, the dynamics of neuronal and network activities are analyzed under the correlated noises mimicking a cholinergic drive. Regardless of the network structure, symmetry and asymmetry, the behavior of network activity and the escape time distributions show that the long-lasting autocorrelation of the noise prolongs its prescence in the metastable states. These results and the newly estimated network attractor show that the interplay between the metastability of the network attractor and the noise statistics determines the dynamics of network activity. Our results may be able to provide the novel framework to investigate the function of dreaming in the higher-order brain function. © 1997 Elsevier Science Ltd.*

Keywords—Correlated noise, Dream sleep, Dynamics transition, Escape time, 1/*f* Fluctuations, Metastability, Network attractor, Slow wave sleep.

1. INTRODUCTION

Sleep state is one of the substantial aspects of consciousness. Concerning the function of sleep in memory and learning, much physiological and psychological data have been accumulated (e.g. Smith, 1995; Antrobus, 1997). Many ideas have been proposed to elucidate the mechanisms underlying them (e.g. Crick and Mitchison, 1983). However, none of these ideas has been established. It is essential to construct a model which can provide an insight into the mechanisms of sleep and enables its computational interpretation. We have made efforts following this line from the physiological and model-based points of view.

In a cat's central nervous system, we have found the following phenomena concerning dynamics of single neuronal activities during sleep cycle. (i) During rapid eye movement sleep (REM or dream sleep), neuronal activities showed slow fluctuations, and their power spectral densities (PSD) were approximately inversely proportional to frequency in the frequency range of 0.01–1.0 Hz (simply abbreviated as ''1/*f*''). (ii) During the steady state of slow wave sleep (SWS), neurons demonstrated the almost flat spectral profiles in the same frequency range. These phenomena have been found in various regions of the cat's brain, such as the mesencephalic reticular formation (Yamamoto et al., 1986), the hippocampus, the thalamus, and the cortex (Mushiake et al., 1988; Kodama et al., 1989; Yamamoto et al., 1995).

Based on neurophysiological knowledge, we successfully simulated the dynamics transition using a symmetry neural network model including a globally applied inhibitory input and a random noise. That is, the neuronal dynamics during SWS and REM was reproduced by the network model under the strong and the weak inhibitory inputs, respectively. We suggested the monotonous structure of the network attractor, where the ''0'' state is highly attractive under the strong inhibition, and the metastability of the attractor is dominant under the weak. Thus, the structural change in the network attractor associated with an increase in the global inhibitory input could underlie the neuronal dynamics transition. We also showed that the variance of the noise could

Acknowledgements: This work was partly supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture of Japan, no. 07459003. M. Musila wished to thank the Japan Society for Promotion of Science, which sponsored his postdoctoral research fellowship in Japan. He passed away last year. The authors would like to thank Professor Risaburo Sato and Hiroaki Sato for their simulations and useful comments.

Requests for reprints should be sent to M. Nakao, Graduate School of Information Sciences, Tohoku University, Sendai 980-77, Japan; Tel and Fax: +81-22-217-7178; E-mail: nakao@ecei.tohoku.ac.jp.

differentiate the network state behavior in the metastable attractor. We assumed the serotonergic system as the physiological substrate providing the global inhibition from our pharmacological experimental results (Mushiake et al., 1988; Kodama et al., 1989). More generally, from the results of recent studies aminergic systems such as the noradrenergic should be included in the substrate (McCarley et al., 1995). On the other hand, the noise might consist of a membrane noise of a neuron, external inputs from other neural networks, etc. Our experimental results also suggest the necessity of cholinergic drive for generating $1/f$ fluctuations (Mushiake et al., 1988). In addition, this is shown to be possibly mediated by the muscarinic pathway rather than the nicotinic (Yamamoto et al., 1993). Furthermore, acetylcholine is released in high concentration during REM (Kodama et al., 1990). Therefore, the cholinergic system which originates from the brainstem and projects directly/indirectly its fibers to the whole brain area could be a possible candidate of a noise source. Actually, our previous results showed how the dynamics of the network state evolution depends on the magnitude of the inhibition and the variance of a Gaussian noise (Nakao et al., 1992). However, it is not yet been known whether or not those results could be applied for dynamics under more generalized conditions. Furthermore, a possible rationale and functional interpretations of the $1/f$ fluctuation during dream sleep remain unsolved.

In this paper, the dynamics of neuronal activities and network state evolution are investigated under a correlated noise which is a first-order autoregressive process. In addition, escape time distributions in metastable states are derived for varied correlations of the noise in order to understand the interaction between the geometrical structure of network attractor and the correlated noise. Through these studies, the temporal correlation of noise is shown to distinctly change the network dynamics cooperating with the metastable structure of network attractor. So far, we have been regarding the global inhibition and the noise variance as control variables to induce the neuronal dynamics transition. However, they may not be the only possibilities. Here, by showing that diluting connections between neurons in the network could play a similar role to them, we propose a general possible mechanism for generating the $1/f$ fluctuations during dream sleep and the white PSD during SWS. We will discuss the neurophysiological mechanisms underlying the dynamics transition as well as the relationship between the dynamics of neuronal activities and dreaming.

2. NEURAL NETWORK MODEL

Here, the model structure is reviewed briefly (Nakao et al., 1990). The neural network model we used consists of fully interconnected neuron-like elements (abbreviated as ''neuron''). For the ith neuron, the state evolution rule is defined as follows:

$$u_i(t+1) = \sum_{j=1}^{N} w_{ij} x_j(t) - h + \xi_i(t+1) \tag{1}$$

$$x_i(t+1) = g(u_i(t+1)) \tag{2}$$

$$g(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3}$$

$$\xi_i(t+1) = \alpha \xi_i(t) + \varepsilon_i(t), \ i = 1, 2, \ldots, N \tag{4}$$

where N denotes the number of neurons contained in the network, and t is a discrete time. $\varepsilon_i(t)$ denotes random noises which are assumed as mutually independent, zero-mean white gaussian noises with an identical variance $\sigma^2(1-\alpha^2)$, in which a variance of autoregressive process is kept constant regardless of α. $h(>0)$ is the inhibitory input which is fixed here independent of neurons for the sake of simplicity, and α is an autoregressive parameter. In this case, the autocorrelation function $r_\xi(k)$ is given by

$$r_\xi(k) = \sigma^2 \alpha^{|k|}, \ k = \ldots, -2, -1, 0, 1, 2, \ldots \tag{5}$$

As described in Section 1, an acetylcholine-containing neuron, which is called a cholinergic neuron, could be one of possible sources for this noise. Although their state-dependent activities are heterogeneous, most cholinergic neurons raise their activities during REM and reduce them during SWS (Kodama et al., 1990; Koyama and Kayama, 1995). Furthermore, a muscarinic receptor in the cat brain is known to respond as fast as on the order of several seconds (McCormick and Price, 1987), and facilitatory effects are exerted on most pyramidal neurons in the hippocampus (Stewart and Fox, 1990) and cortex (Sato et al., 1987; McKenna et al., 1988), and a thalamic relay neuron (McCormick and Price, 1987). Considering the responsiveness of a muscarinic receptor and temporal variabilities in the discharge pattern of cholinergic neurons during REM (Sei et al., 1993; Koyama and Kayama, 1995), the cholinergic input is supposed to be a band-limited noise with a d.c. component in the frequency range 0.01–1.0 Hz, which we are interested in. Taking account of supposition, the correlated noise is assumed to come from various sources, such as the cholinergic/noncholinergic input from external environment, and an intrinsic membrane noise.

Synaptic weights w_{ij} are defined following the case of an associative memory:

$$w_{ij} = \begin{cases} \dfrac{1}{N} \displaystyle\sum_{m=1}^{M} (2x_i^{(m)} - 1)(2x_j^{(m)} - 1) & i \neq j \\ 0 & i = j \end{cases} \tag{6}$$

where $x_i^{(m)}$ indicates the ith neuron's state of the mth memorized pattern, and M is the number of memorized patterns. This possibly enables the parametric control of the fundamental structure of the network attractor. In this case, the symmetry condition

$$w_{ij} = w_{ji} \tag{7}$$

is satisfied.

In addition, the asymmetry network is subject to simulations, which can be regarded as more physiological than the symmetry. In the symmetry network, each neuron is allowed to exert its inhibitory and excitatory synaptic effects on target neurons simultaneously. On the other hand, in the asymmetry neural network, a permitted sign of synapse depends on the attribute of a neuron, i.e. excitatory or inhibitory (Dale's rule: Gray, 1987). Here, synapses of the symmetry network which do not obey Dale's rule are cut in order to obtain the asymmetry neural network (Shinomoto, 1987). The attribute of a neuron is assigned at random on each neuron with proportions ρ (excitatory) and $1 - \rho$ (inhibitory). With synchronous state evolution, the constructed asymmetry network was shown to preserve the same retrieving property of the embedded memory patterns as the original symmetry network except for the limit cycle, which some initial states induce (Shinomoto, 1987). Our simulations with asynchronous state evolution find no limit cycles and no chaotic behavior.

For both types of networks, state evolution in networks is performed in the asynchronous (cyclic) manner (Hopfield, 1982). The memorized patterns and the initial states are given as equiprobable binary random sequences. Unless otherwise stated, simulations are carried out for 11 000 Monte Carlo steps (MCS), and the initial 1000 MCS is not analyzed to exclude the state sequence dependent on the initial pattern. Since, in our case, a PSD of a state sequence is almost invariant against the temporal translation of the sequence, the starting time of the analysis scarcely affects the resulted PSD.

The data length, 10 000 MCS, is selected to estimate PSD in the frequency band width of three decades with sufficient statistical reliability. The PSDs of actual neuronal activities referred here were given in the similar frequency band width (Yamamoto et al., 1986). Furthermore, the data length of the neuronal spike train analyzed was at most several hundred seconds. Comparing this actual data length with 10 000 MCS, 1MCS could correspond to several tens of milliseconds. This could be regarded as a time unit during which a neuron keeps its state active (1) or inactive (0). The neuronal state may be determined to be responsible for the number of spikes during this time unit. This time resolution is presumably sufficient, considering that the firing rates of actual neurons under study were at most 30–40 spikes s^{-1}, and the concerned frequency range is lower than 1 Hz (Yamamoto et al., 1986).

3. SPECTRAL ANALYSIS OF NEURONAL ACTIVITIES IN NEURAL NETWORK MODEL

Typical PSD profiles of single neuronal activities in the network model are shown in Figure 1 for varied inhibitions and α values. Unless otherwise stated, the number of neurons $N = 100$ and the number of memorized patterns $M = 30$. In this figure, activity of a single neuron is picked up from 100 neurons included in the network. The raster plot of $x_i(t)$ is shown together with the corresponding PSD, where a dot indicates $x_i(t) = 1$. As one can see for the case of $\alpha = 0$ (i.e. white noise), PSD changes its profile from the $1/f$ to the flat as the inhibitory input increases. Here, the parameter values h and σ are selected regardless of the connection type, so that most of neurons in the network show the $1/f$ PSD profiles under the weak inhibition and the flat PSD profiles under the strong inhibition. The time series $x_i(t)$ responsible for the $1/f$ PSD shows larger and slower variations than that for the flat PSD. Naturally, the activity is reduced as the inhibitory input increases. As described previously (Nakao et al., 1990), the strong and the weak inhibitory inputs are responsible for SWS and REM in our framework, respectively. Qualitatively, the PSD profiles and the temporal characteristics of activities are well reproduced in our simulations. Through simulations with $\alpha = 0$ presented here, an inhibitory input is manipulated so that an activity under the strong inhibition is reduced to about 10% of that under the weak inhibition. This proportion is selected as the limit of inhibition, because under the stronger inhibition the number of excitatory states could not be sufficient to statistically validate the estimated PSD. Regardless of inhibition level, finely fragmented activities tend to be suppressed as α increases, which is more obvious in the strong inhibition case. Slopes of PSDs commonly become steeper, and are associated with an increase in α. Since an increase in α results in longer lasting autocorrelation, the above results can be attributed to a change in the correlation structure of the noise. The neuronal activities appear to more closely follow the dynamics of the noise as α increases.

The simulation results for the asymmetry network ($M = 30$, $\rho = 0.5$) are presented in Figure 2. In contrast with the symmetry, clusters of activities still survive even under the strong inhibition. However, the above results summarized for the symmetry network qualitatively apply to the asymmetry network.

4. DYNAMICS OF NEURAL NETWORK IN STATE SPACE

Up to now, we have been studying the dynamics of neural network models which receive the global inhibition and the uncorrelated random noise (Nakao et al., 1990), where the structural change of the network attractor associated with the inhibition has been investigated

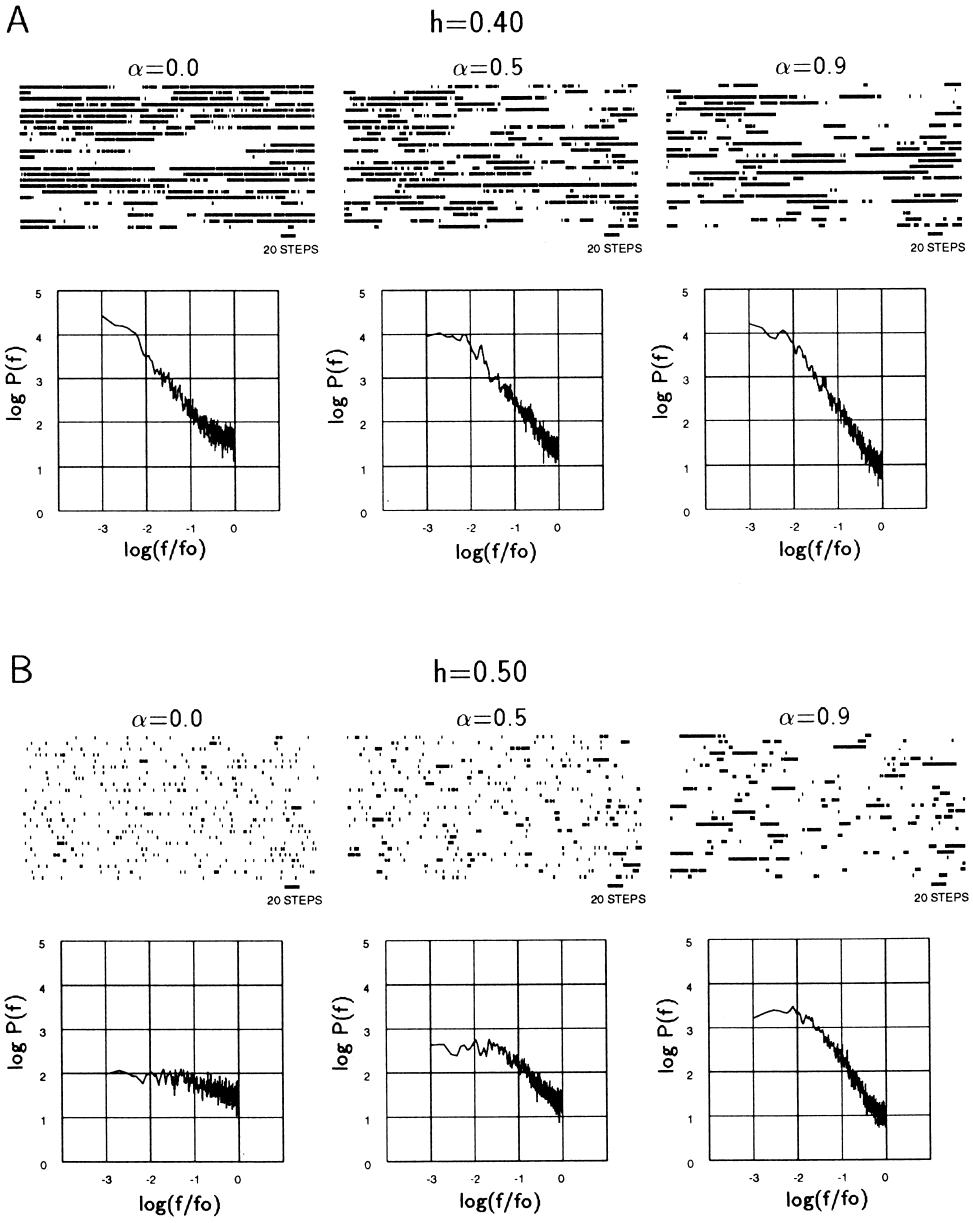


FIGURE 1. Simulation results on the dynamics transition of neuronal activities in the symmetry neural network for varied α under weak and strong inhibitions, h, where $M = 30$ and $\sigma = 0.26$. Raster plots of single activities and the corresponding PSDs for the picked-up neuron are shown together. The upper panels (A) show the results for $h = 0.40$ and the lower (B) for $h = 0.50$. In PSD, the frequency axis is normalized by f_0 which denotes an arbitrary standard frequency.

through the state space analysis. We suggested that this structural change could underlie the dynamics transition of neuronal activities during sleep cycle. Therefore, the metastable properties of the attractor could be a key issue in understanding the physiological mechanism which controls the dynamics of neuronal activities during the sleep cycle. Here, we analyze how the correlation of the random noise modifies the network dynamics in the state space.

For the same symmetry network in Figure 1, activities of all neurons (network activity) are briefly presented in Figure 3(A) under the weak inhibition, and (B) under the strong inhibition. In each figure, the autoregressive parameter α differentiates the pattern of network activities.

As shown in Figure 3(A), under the weak inhibition, the network activity explicitly indicates that the regular and irregular patterns appear alternatively with varied durations. In the regular states, several different stripe patterns can clearly be seen. In contrast, only the irregular state becomes dominant under the strong inhibition. It can be shown that these stripe patterns correspond to the vicinities of equilibrium states under this condition, while the irregular pattern corresponds to the vicinity of the ‘‘0’’ state where all neurons are silent. Naturally, these equilibrium states except for the ‘‘0’’ state are

different from those memorized by eqn (6) due to the global inhibition. The closest reference equilibrium state to the current network state is determined every MCS in terms of a direction cosine (DC). DC here represents the ''agreement'' between a current network state $\mathbf{x}(t)$ and a certain reference state $\mathbf{x}^*$, which is defined by

$$DC = \frac{(2\mathbf{x}^* - \mathbf{1})'(2\mathbf{x}(t) - \mathbf{1})}{|2\mathbf{x}^* - \mathbf{1}||2\mathbf{x}(t) - \mathbf{1}|},$$

$$x'(t) = [x_1(t),\ x_2(t), \ldots,\ x_N(t)],\ x'^* = [x_1^*,\ x_2^*, \ldots,\ x_N^*],$$

$$\mathbf{1}' = [1,\ 1, \ldots,\ 1], \tag{8}$$

where $\mathbf{x}'$ denotes a transpose of a vector $\mathbf{x}$, and $\mathbf{x}$ a length of $\mathbf{x}$. For the networks appearing in Figure 1, reference equilibrium states including the ''0'' state are reached from 4000 statistically independent initial states under no noise, i.e. $\sigma = 0$. Here, the ''0'' state is denoted by $\mathbf{x}_0 = [0,\ 0, \ldots,\ 0]$. Sixty-three and two equilibrium states are found for the networks in Figure 1(A) and (B), respectively; 115 and 18 for the networks in Figure 2(A) and (B), respectively. For all references, the closest reference to the current state is determined step by step by comparing the magnitude of the corresponding DCs.

Under the weak inhibition, the network state is drifting among the vicinities of the equilibrium states. Here, the

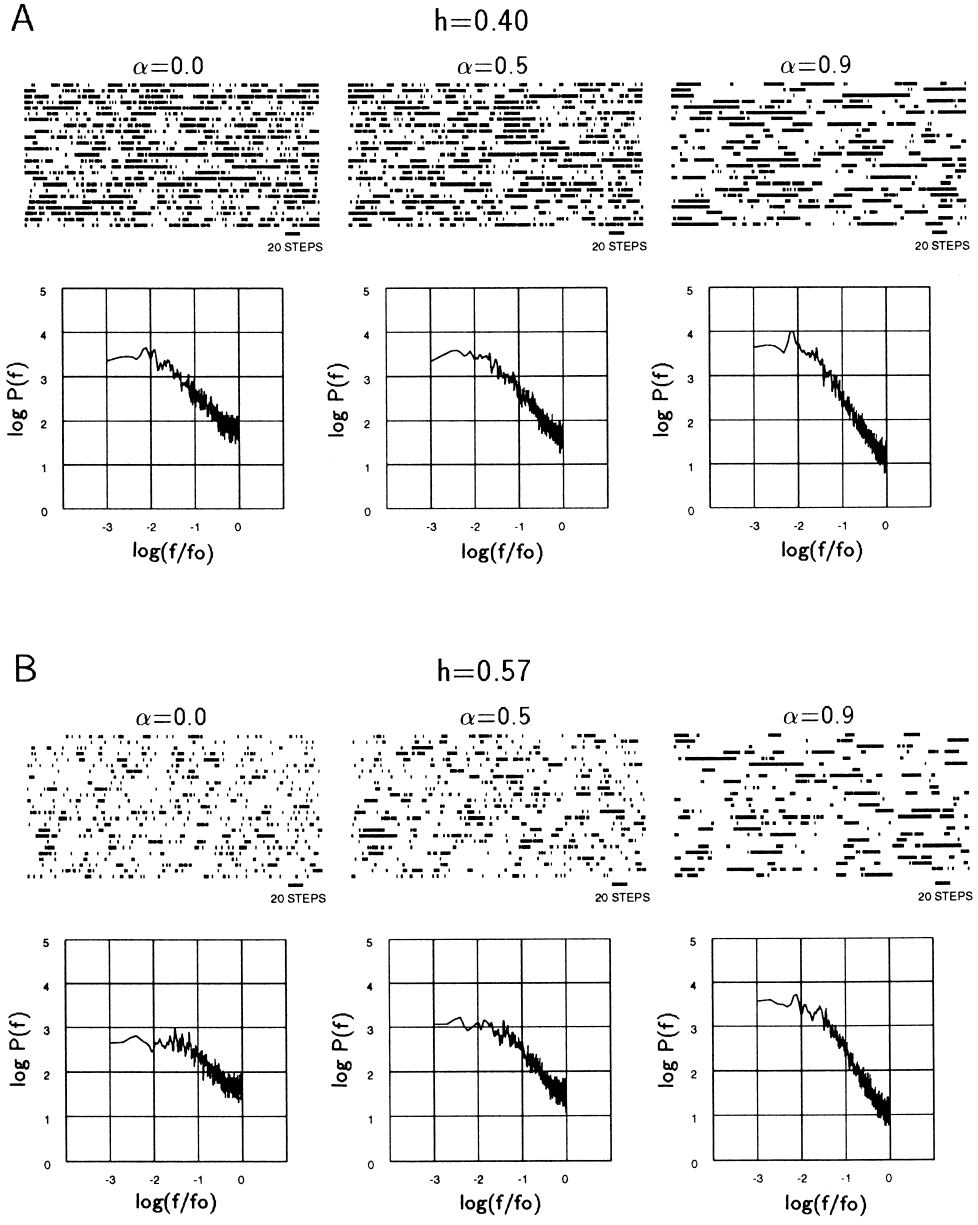


FIGURE 2. Simulation results on the dynamics transition of neuronal activities in the asymmetry neural network for varied α under weak and strong inhibitions, h, where $\rho = 0.5$, $M = 30$ and $\sigma = 0.21$. Raster plots of single activities and the corresponding PSDs for the picked-up neuron are shown together. The upper panels show the results for $h = 0.40$ and the lower for $h = 0.57$. In PSD, the frequency axis is normalized by f_0 which denotes an arbitrary standard frequency.

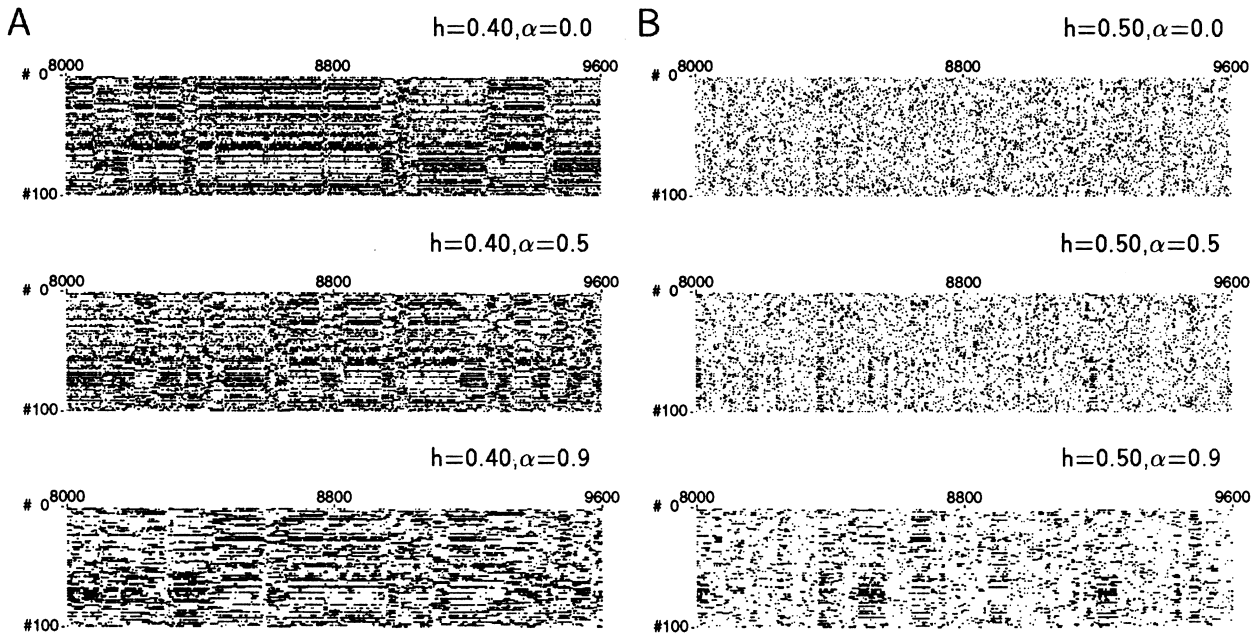


FIGURE 3. Dynamics of network state evolution for the network shown in Figure 1. This shows the brief sequences of all neuronal activities (network activity), where the numbers on the left end denote neuron numbers, and the number on the top indicates the number of steps from the beginning of evolution.

equilibrium states are not absolutely stable, because intermittent transitions among them driven by the noise are observed. In this sense, they are denoted here as ''metastable equilibrium states'' or simply metastable states following the terminology of statistical physics (Amit, 1989). In spite of a constant drive by the noise, the network state is trapped in metastable states for a certain period. In the irregular states, the network is possibly drifting around the vicinity of the ''0'' state. While in the vicinity of the ''0'' state, each neuronal state is expected to be determined by an instantaneous value of the noise rather than inputs from the other neurons. This is presumably the reason why the spatio-temporal activity pattern appears random. Here, the

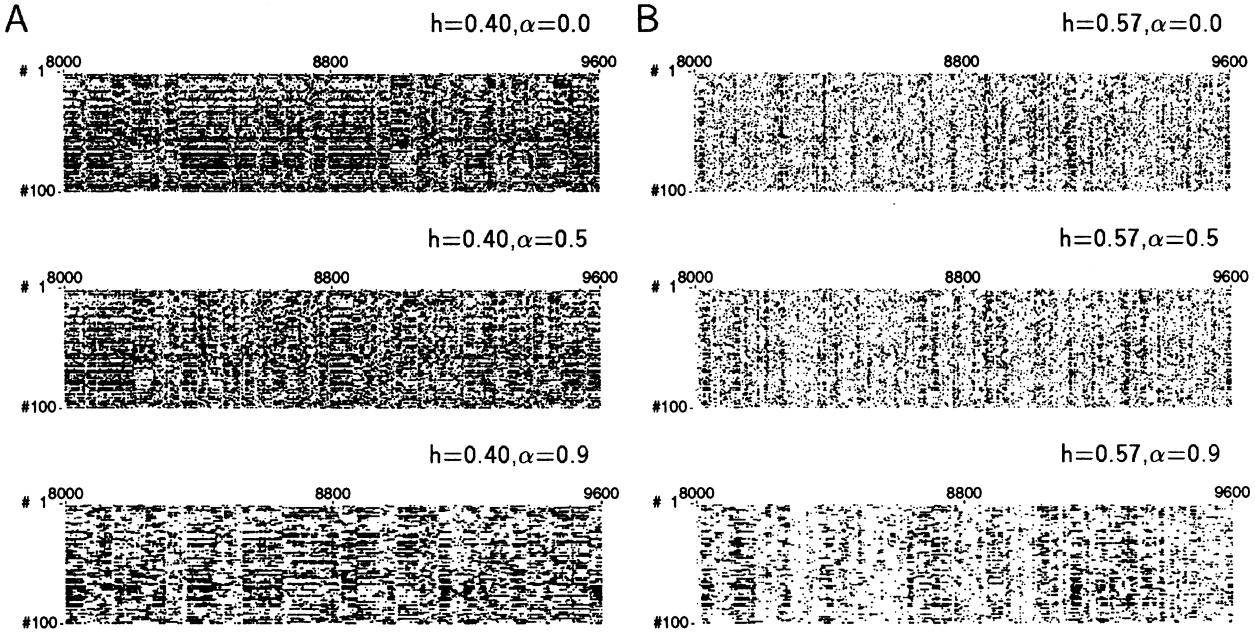


FIGURE 4. Dynamics of network state evolution for the network shown in Figure 2. This shows the brief sequences of all neuronal activities, where the numbers on the left end denote neuron numbers, and the number on the top indicates the number of steps from the beginning of evolution.

''metastability'' represents the structural properties of the network attractor in which such metastable equilibrium states dominantly exist. Therefore, the following description is possible concerning the dynamics transition of single neuronal activities. The globally applied inhibitory input modifies the structure of the network attractor. In the weakly inhibited case, the metastability of the network attractor becomes dominant so as to realize the $1/f$ fluctuations of single neuronal activities, and in the strongly inhibited case, the ''0'' state becomes the global attractor which underlies low and random activities. In other words, we suggest that these behavior reflect the geometrical structure of the network attractor: a ''bumpy'' structure becomes dominant but rather monotonous under the strong inhibition.

For $\alpha = 0.5$, one may not be able to recognize the difference between the behavior of the network state and the above results in the case of the white Gaussian noise, i.e. $\alpha = 0$. However, finely fragmented patterns such as snow noises become suppressed for $\alpha = 0.9$ in both the strongly and weakly inhibited cases. For the weakly inhibited case, more types of regular patterns could be recognized than the case with the smaller α. Similarly, in the strongly inhibited case, the distinct regular patterns are clearly raised as α increases.

For the same asymmetry network with $\rho = 0.5$ in Figure 2, the network state behavior almost resembles the dynamical features observed in the symmetry case (Figure 4). We suggested that several metastable states still survive even under the strong inhibition, which is distinct from the symmetry case (Nakao et al., 1992). This is reflected on the cluster activities of a single neuron in Figure 2(B). As seen in Figure 3(B), the large α distinctly raises the regular patterns under the stronger inhibition than the symmetry network.

5. METASTABLE NETWORK ATTRACTOR AND CORRELATED NOISE

In this section, we investigate a mechanism underlying the behavior of the network activity controlled by the correlation properties of the noise. As α increases, we have observed that the irregular patterns are suppressed for the weakly inhibited case, and the regular patterns are raised even under the strong inhibition. These observations suggest that the correlation properties of the noise could control an elapsed time during which the network state is trapped in a metastable state.

5.1. Escape Time Distribution in a Metastable Equilibrium State

In order to confirm the above implication more quantitatively, distribution of escape time in respective metastable state is obtained for varied α by a Monte-Carlo simulation with 10,000 trials. Here, the escape time is defined as the time required for the network state, which is initially located in a metastable

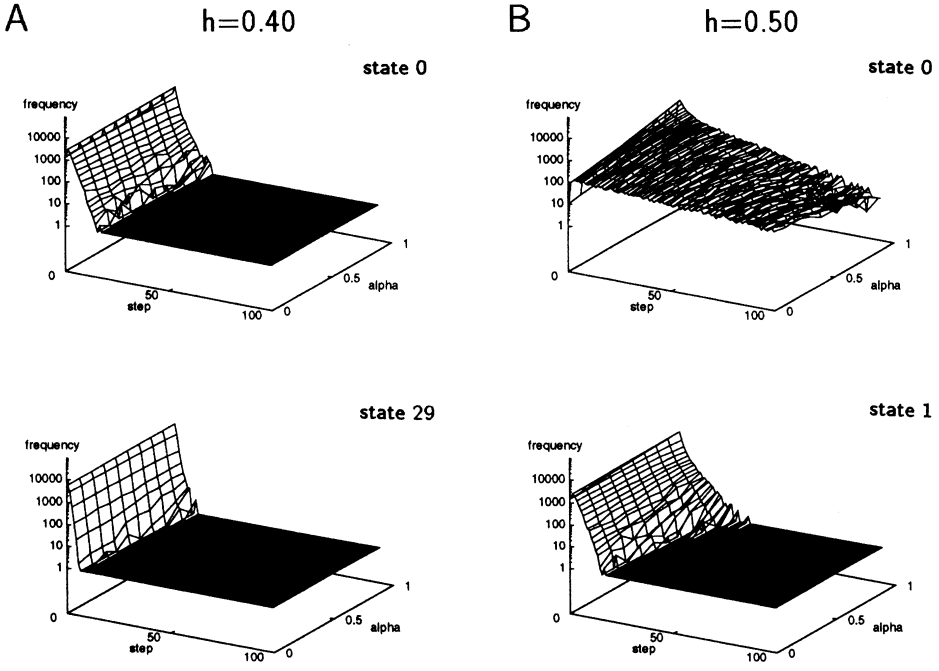


FIGURE 5. Escape time distributions in metastable states (on the definition of an escape time, see the text). Part (A) shows the semi-logarithmic plots of the escape time distribution in metastable states, state 0 and state 29, for the network shown in Figure 1(A), where $h = 0.43$ and $\sigma = 0.26$. Part (B) shows the semi-logarithmic plots of the escape time distribution in metastable states, state 0 and state 1 for the network shown in Figure 1(B), where $h = 0.50$ and $\sigma = 0.26$. Here, the abscissa denotes the escape time in Monte Carlo steps, the ordinate the value of α, and the vertical the frequency.

equilibrium state, to first cross a boundary distant from the corresponding equilibrium state by 0.2 in terms of the DC under the correlated noise with the parameter α. Figure 5 and Figure 6 show the escape time distributions for the symmetry and the asymmetry networks, respectively. For both types of networks, an escape time is distributed in a monophasic manner peaking in the short range. As shown in the distributions of states 29 and 1 in Figure 5 and those of states 8 and 1 in Figure 6, the escape time for a metastable state (except for the ''0'' state) usually tends to be prolonged and associated with an increase in α, while for the ''0'' state a consistent relationship between the distribution and the value of α could not be found; in some cases, it is rather shortened. This might be attributed to the peculiar landscape of potential energy around the ''0'' state described in Section 5.2. Nevertheless, since the prolongation of the escape time is commonly observed in all other metastable states, the above results support our implication.

5.2. Potential Walls Surrounding Metastable States

Stochastic properties of network activity is roughly characterized by staying probability in a metastable state as well as transition probability from one metastable state to the other, where a higher-order Markovian nature is assumed to be negligible. Both probabilities possibly reflect the height of the potential wall between metastable states. However, since the network potential function is a multi-dimensional function, it is difficult to estimate the height of the potential wall. Here, the following computational procedure is employed to estimate the height of the potential wall between metastable states. We now show how to estimate the height of the wall between equilibrium states $\mathbf{s}_i$ and $\mathbf{s}_j$. Firstly, the maximum network potential is obtained during the process approaching $\mathbf{s}_i$ to $\mathbf{s}_j$ by flipping the different bits (neuronal states) one by one. Then, the potential maxima are collected by repeating the same procedure, changing the flipping order 100 times. The minimum in the set of the maxima is selected as an potential wall height between $\mathbf{s}_i$ and $\mathbf{s}_j$. In addition, the transition probability from $\mathbf{s}_i$ to $\mathbf{s}_j$ is estimated by 10 runs of 10,000 MCS network state evolutions, where the transition probability for $i = j$ indicates the staying probability in state i. The wall height and the transition probability are presented for the symmetry network in Figure 7. Since a network potential function is not known for the asymmetry network studied here, the analysis is confined to the symmetry network. For the weakly inhibited case, the wall heights and the transition probabilities from the ''0'' state and state 29 to all other metastable states are presented. Note that the wall heights in the objective states are ''0''. Characteristically, there are high potential walls between the ''0'' state and any other metastable states. In contrast, there are several low walls around state 29, e.g. to states 4, 26 and 55. This structural property of the network potential around state 29 is

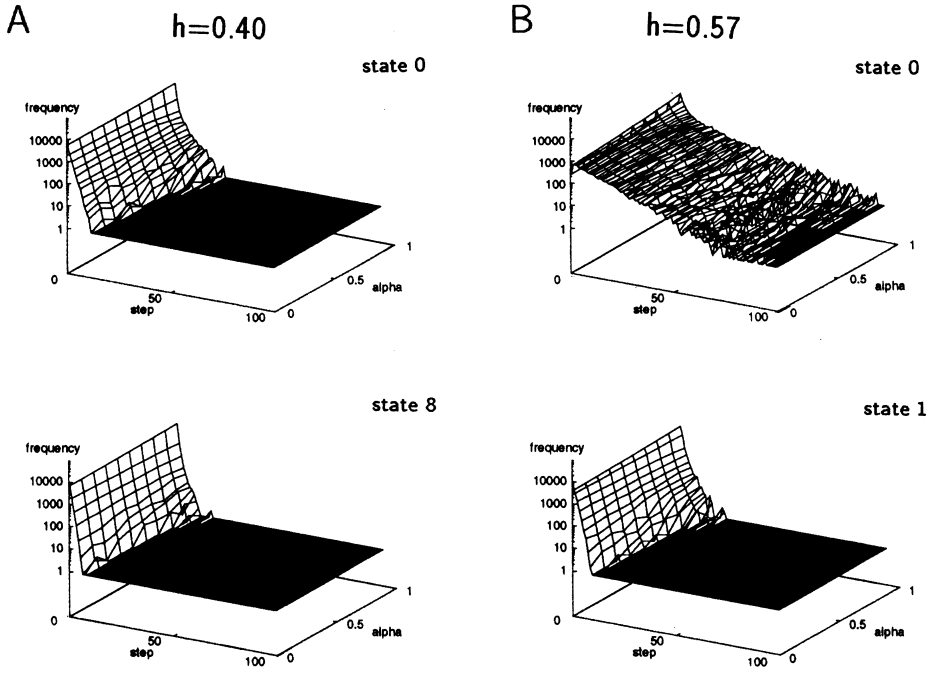


FIGURE 6. Escape-time distributions in metastable states. Part (A) shows the semi-logarithmic plots of the escape time distribution in metastable states, state 0 and state 8, for the network shown in Figure 2(A), where $h = 0.40$ and $\sigma = 0.21$. Part (B) shows the semi-logarithmic plots of the escape time distribution in metastable states, state 0 and state 1 for the network shown in Figure 2(B), where $h = 0.57$ and $\sigma = 0.21$. Here, the abscissa denotes the escape time in Monte Carlo steps, the ordinate the value of α, and the vertical the frequency.

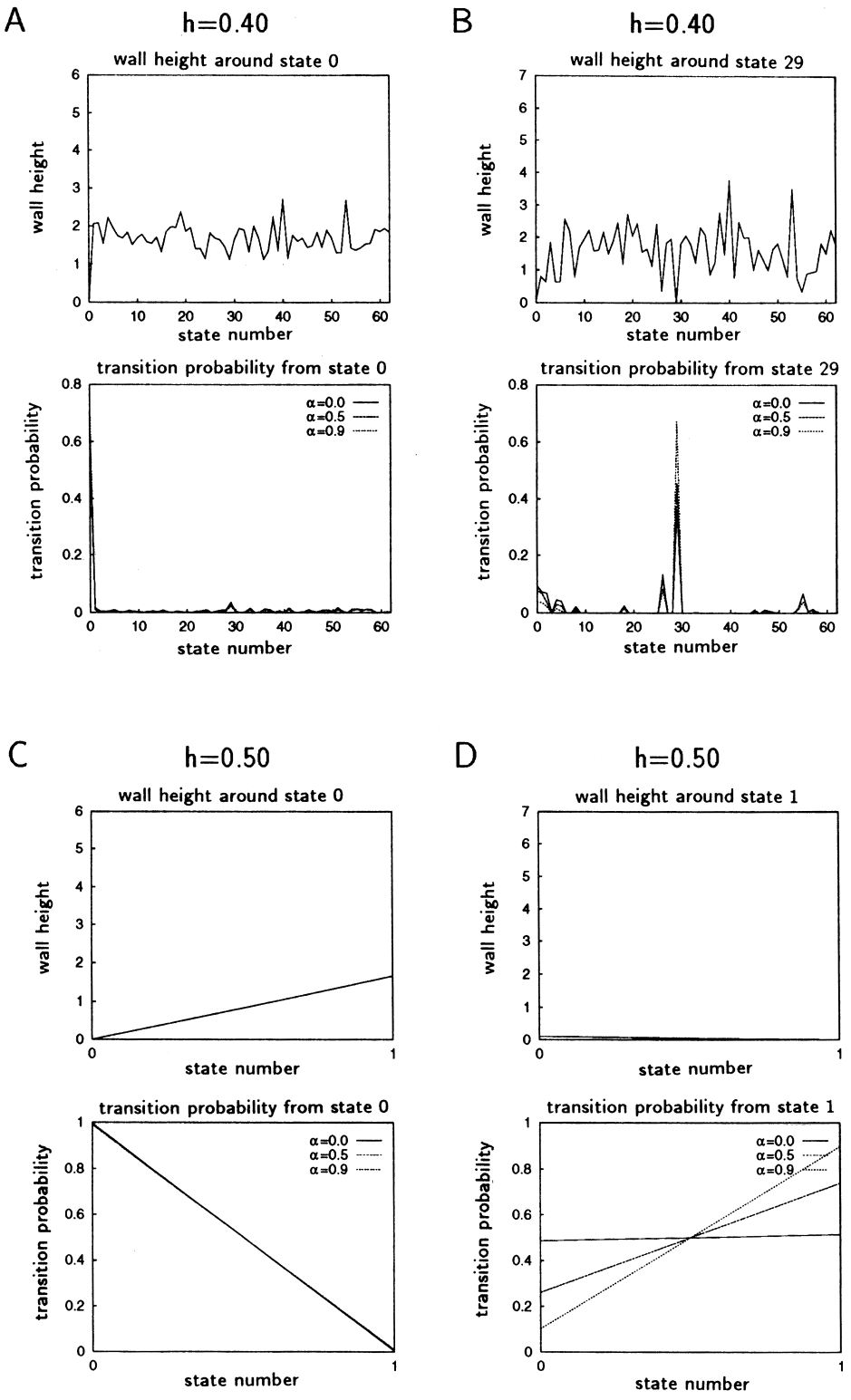


FIGURE 7. Estimated height of network potential walls between metastable states and the transition probability (for the procedure on estimating the potential wall height, see the text). (A) Height of network potential walls around the "0" state and the transition probability from the "0" state. (B) Height of network potential walls around state 29 and the transition probability from the state 29. Parts (A) and (B) for the symmetry network under the weak inhibition appeared in Figure 1(A). (C) Height of network potential walls around the "0" state and the transition probability from the "0" state. (D) Height of network potential walls around state 1 and the transition probability from the state 1. Parts (C) and (D) for the symmetry network under the strong inhibition appeared in Figure 1(B).

shared by the other metastable states except for the ''0'' state. In this sense, the potential landscape around the ''0'' state is special. Regarding the transition probability, transition from the ''0'' state to the other metastable states is rare in comparison with staying, which could be well understood from the above special potential landscape. On the other hand, the transition probability from state 29 is high in itself and the other metastable states with low potential walls. This agreement between the height of the potential wall and the transition probability demonstrates the validity of the procedure for deriving the wall height. A larger α is shown to make the transition to the other states less frequently and to increase a staying probability. Although, under the strongly inhibited condition, a few metastable states could only be analyzed, the results in Figure 7(C) and (D) show similar features to the weakly inhibited case in Figure 7(A) and (B). Under this condition, the potential wall from the ''0'' state to state 1 is much higher than from state 1 to the ''0'' state, which is thought to make the staying probability close to 1. For state 1, the staying probability increases and the transition decreases, as α moves closer to 1.

Escape time distribution is expected to depend on the local landscape of network potential around an equilibrium state. A symmetry neural network is known to be a multi-dimensional discrete gradient system (see for example Hopfield, 1982). However, there is no general theory describing the metastability of such a multi-dimensional system. On the other hand, for a one-dimensional continuous gradient system with a two-well potential, escape time under a small Gaussian noise obeys an exponential distribution whose parameter depends on the height of the potential wall between two wells (see Bulsara and Ross, 1991). That is, a staying probability in a shallow potential well has a faster decaying profile than in a deep well. Although under the correlated noise, the theoretical results are derived only under limited conditions even for one-dimensional potential case, from some numerical experiments the escape time distribution is expected to depend on the local geometry of the attractor as well as the correlation structure of the noise (Moss and McClintock, 1989). Our result obtained for the neural network qualitatively coincides with those of one-dimensional system.

These interpretations are based on the gradient property of the system. As far as the asymmetry network used here is concerned, it remains unclear whether this network is a gradient system or not, because its Lyapunov function appears difficult to obtain (Kobuchi, 1991). However, according to the basic property of the asymmetry network dynamics described in Section 2 and its α-dependency of the escape time distributions similar to the symmetry one, it could be regarded as a gradient system. Therefore, its behavior is expected to be understood, at least qualitatively, within the same framework as the symmetry.

In short, the behavior of network activity in the state space consists of stochastic transitions among metastable equilibrium states. The stochastic features of transitions are determined by the height of the potential walls around metastable states and the correlation structure of the noise. As far as the inhibition induced dynamics transition is concerned, the global inhibition reduces the height of the potential walls and the number of metastable states so that a PSD of a neuronal activity changes its profile from the $1/f$ to the white.

6. NEURONAL DYNAMICS TRANSITION INDUCED BY DILUTING CONNECTIONS

In the previous sections, the global inhibition is shown to change the structure of the network attractor so as to induce the neuronal dynamics transition. From another point of view, an increase in the global inhibition and/or in the noise variance implies that external inputs become dominant in comparison with the inputs from the other neurons interconnected (network input). Here, ''external'' means that neurons under consideration are not involved in its dynamics. According to this context, the previously proposed mechanism for the neuronal dynamics transition could be reinterpreted as follows. Under the weakly inhibited condition, interaction between the external inputs and the network inputs is prevailing, while under the strongly inhibited condition, the external inputs dominates the network inputs. That is, the balance between the external and the network inputs is supposed to play an essential role in inducing the neuronal dynamics transition. In order to realize the same situation in a different manner from the global inhibition, we investigate how randomly diluting connections between neurons affect the structure of the network attractor and the metastable behavior of the network.

Figure 8 shows briefly the network activities in the state space, as the connections are diluted with varied ratios, where the connection is symmetry, $N = 196$, $M = 59$, $\sigma = 0.28$ and $h = 0.53$. The noise is a white Gaussian, and the values of σ and h are set so that most of the neurons exhibit the $1/f$ fluctuations. The dilution is carried out at random and in a symmetrical manner. With no dilution, the metastable behavior is clearly shown to be similar to that in Figure 3(A) and Figure 4(A). As the dilution ratio increases, irregular patterns such as a snow noise become distinct in the network activities, where most of the neurons exhibit flat PSDs in the frequency range less than $f \sim 10^{-1}$. In order to understand the structural change of the network attractor associated with the dilution, the number of equilibrium states are derived by the same procedure described in Section 4. Figure 9 shows the number of metastable equilibrium states as a function of the dilution ratio together with that as a function of an inhibitory input in the fully interconnected network. Both curves

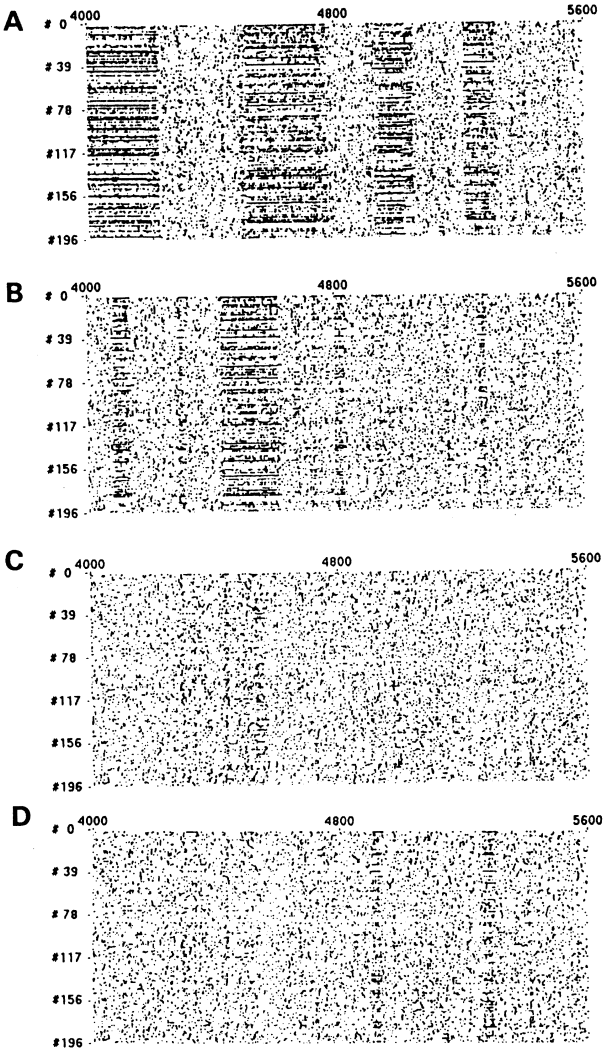


FIGURE 8. Dynamics of network state evolution for the network whose connections are partly diluted, where $N = 196$, $M = 59$, $h = 0.53$ and $\sigma = 0.28$. The dilution ratios are (A) 3.6%, (B) 5.5%, (C) 7.6% and (D) 9.4%. This shows the brief sequences of all neuronal activities (network activity), where the numbers on the left end denote neuron numbers, and the number on the top indicates the number of steps from the beginning of evolution.

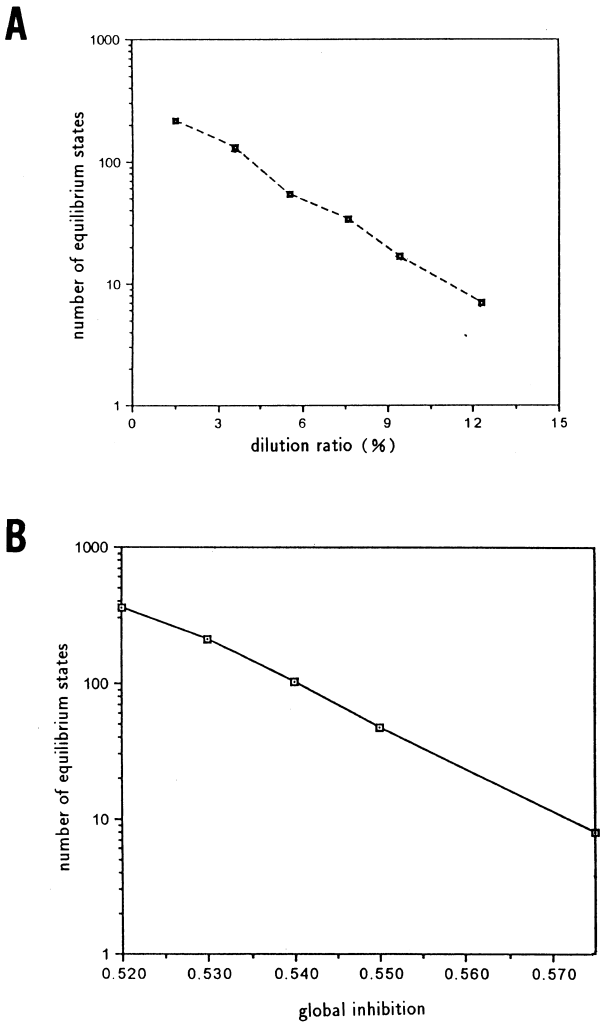


FIGURE 9. Number of metastable equilibrium states as functions of the dilution ratio (A) and the global inhibition (B). They are estimated for the network shown in Figure 8.

monotonically decline with an increase in the dilution ratio or the inhibitory input. Furthermore, a similar result is obtained for the other trials of random dilution. Naturally, full dilution isolates each neuron, which results in purely random neuronal activities elicited only by the noise. Within the numerical range of the inhibition and the dilution ratio used here, this result suggests that the similar structural change is caused by both of the dilution and the inhibitory input, also taking account of the network activities in Figure 8. Although, of course, there are many possible ways of reducing effective network inputs, the balance between the external inputs and the network inputs is suggested to be essential for inducing the neuronal dynamics transition.

7. DISCUSSION

In this paper, the neuronal activities and the network activities in the state space have been analyzed for the symmetry as well as the asymmetry neural network models under a drive by the correlated noise. This correlation represents the supposed dynamics of cholinergic inputs. The noise correlation was observed to suppress the finely fragmented activity so that the slope of the $1/f$ and the flat PSDs became steeper. This result indicates augmentation of frequency power in the lower frequency range. In order to demonstrate the relationship between the network attractor and the noise statistics, the escape time distributions for the metastable states have been derived for the varied correlations. Our main finding is that the escape time tends to be prolonged and associated with an increase in the correlation time of the noise regardless of the connection type, symmetry and asymmetry. Here

the correlation time indicates the time during which significant autocorrelation exists.

We have already suggested that the structural change of the network attractor associated with the global inhibitory input could be the underlying mechanism for the neuronal dynamics transition we observed physiologically (Nakao et al., 1993, 1994). In particular, the metastable properties of the network attractor are found to play a main role for generating the 1/*f* fluctuations of neuronal activities. We also found that the correlation properties of the noise could change the dynamics of the network activities, and the staying time in the metastable state was prolonged due to an increase in the correlation time of the noise (Nakao et al., 1994). Here, those results have been confirmed in terms of the escape time distribution in each metastable state, and the metastable structure of the network attractor has been roughly visualized by estimating the wall height of the network potential energy between metastable states. In addition, diluting the connections in the network has been shown to modify the structure of the network attractor so that the dynamics transition of neuronal activities took place, which was similar to that induced by the global inhibition. This result generalizes the conditions for generating the 1/*f* fluctuations and the dynamics transition of neuronal activities. According to this result, one of the essential factors for inducing the dynamics transition is suggested to be the balance between the external inputs such as the global inhibition and the noise, and the inputs from the other interconnected neurons, i.e. network inputs. In other words, when the network inputs and the external inputs are of a comparable order, the metastability of the network attractor distinctly appears: the 1/*f* fluctuation of neuronal activities are shown. In contrast, when the external inputs exceed the network input, the ''0'' state is highly attractive: the neuronal activities exhibit the flat PSD. According to our simulation results, mechanisms underlying the actual neuronal dynamics transition may be anticipated as follows. During SWS, neurons receive stronger inhibitory inputs and/or less input magnitude from interconnected neurons in comparison with REM. In contrast, during REM, neurons are released from inhibitory inputs and/or receives comparable input magnitude from interconnected neurons with the inhibition and the noise.

We suggest the serotonergic system as a possible candidate responsible for the globally working inhibitory system. This is based on our neuropharmacological results (Mushiake et al., 1988; Kodama et al., 1989). However, the noradrenergic system, which is known to have a similar state-dependent activity to the serotonergic system (Jacobs, 1985), is qualified as another substrate exerting a biasing effect. In addition, the cholinergic input, mediated mainly by the muscarinic pathway, possibly has a d.c. component in addition to the temporally variable component as a source of the noise. Furthermore, there seems to exist a negative feedback loop between serotonergic (noradrenergic) and cholinergic neurons (McCarley et al., 1995). Therefore, to be physiologically precise, the common inhibitory postsynaptic input, *h*, represents overall influences of the ''modulatory systems'' (Maeda, 1995)" such as the serotonergic, noradrenergic and cholinergic systems. So, from the physiological point of view, our results are meaningful with respect to the possible function of the modulatory systems being deeply involved in controlling and maintaining consciousness. In order to confirm our hypothesis and to clarify the contribution of respective systems, we are planning to set up an experiment where the relationship between extracellular concentrations of amine/acetylcholine and the dynamics of neuronal activities will be investigated. For this purpose, we have developed the special technique combining a microdialysis and a unit recording (Nakamura, 1996).

Rapid eye movement sleep, where the single neuronal activity shows the 1/*f* fluctuation, is well known as a dream sleep. Crick and Mitchison (1983) postulated that dream sleep functions acts to remove certain undesirable modes of interaction in neural networks of the cerebral cortex by nonspecific random excitation (PGO waves), which are known to be generated in the brainstem and delivered to the cerebral cortex during dream sleep. Their idea was concurrent with the ''unlearning'' algorithm (Hopfield et al., 1983). However, its physiological reality has not yet been known. Concerning the relationship between the 1/*f* fluctuation of a single neuronal activity and PGO waves, we have the preliminary physiological evidence suggesting that there is no correlation between them (Yamamoto et al., 1993). Nevertheless, the metastable behavior of the network activity appears to be suitable for unlearning, because the depth of a potential well where a metastable state is located is reflected on the corresponding staying time. Appropriate unlearning could be performed by reducing synaptic weights every time the network remains in the vicinity of a metastable state. Based on a physiology-based concept similar to ours, Sutton et al. (1992) studied the state-dependent behavior of the associative sequencing asymmetry network which memorized limit cycles (Peretto, 1992). Although the implementation of the aminergic and cholinergic effects was different from ours, the network was assumed to be less excitable in the non-REM state (SWS in our case), and more excitable and randomly perturbed in REM. Therefore, from our previous and current results (Nakao et al., 1993; Yamamoto et al., 1995), the metastable property is expected to be pronounced in the simulated REM, which could underlie the diverse dynamics which they observed. These assumptions and the observed phenomena qualitatively coincide with ours. Although their network attractor is considered to have a more complex structure, the mechanism underlying the state-dependent behavior of their network could be understood within the same framework as ours.

Dreaming might be regarded as a random recalling process of memorized patterns without logical context. In this respect, the metastable behavior of the artificial neural network could be analogous to dreaming. Therefore, our model-based approach could provide novel information for investigating the functions of dreaming and REM through the 1/*f* fluctuations of neuronal activities. From the cognitive point of view, computational network models on dreaming have been proposed (Hinton et al., 1995; Antrobus, 1997). In addition, there have been many physiological and psychological studies which suggest an important role of a dream sleep in memory and learning processes such as Wilson and McNaughton (1994), Karni et al. (1994) and Smith (1995). The relationship between these ideas and the 1/*f* fluctuations merits further study.

Some researchers have reported dreamings during non-REM (e.g. Vogel et al., 1972). Currently, we do not have any appropriate explanations for this phenomena. According to our simulation results, metastable states with shallow potential wells and peculiar periodic events during non-REM such as spindling and slow wave activity may be a possibility. This will be the subject of a future paper.

REFERENCES

Amit, D. J. (1989). *Modeling Brain Function*. Cambridge: Cambridge University Press.

Antrobus, J. (1997). Dream theory 1997: toward a computational neurocognitive model. *Sleep Research Society Bulletin, 3,* 5–10.

Bulsara, A. R. & Ross, F. E. (1991). Cooperative stochastic processes in reduced neuron models. *Proceedings of the International Conference on Noise in Physical Systems and 1/f Fluctuations*, pp. 621–627.

Crick, F., & Mitchison, G. (1983). The function of dream sleep. *Nature, 304,* 111–114.

Gray, E. G. (1987). Synapse, morphology. In G. Adelman (ed.) *Encyclopedia of neuroscience* (pp. 1158–1162). Boston, MA: Birkhäuser.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The ''wake-sleep'' algorithm for unsupervised neural networks. *Sciences, 268,* 1158–1161.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA, 79,* 2254–2258.

Hopfield, J. J., Feinstein, D. I., & Palmer, R. G. (1983). ''Unlearning'' has a stabilizing effect in collective memories. *Nature, 304,* 158–159.

Jacobs, B. L. (1985). Overview of the activity of brain monoaminergic neurons across the sleep-wake cycle. In A. Wauquier et al. (Eds.) *Sleep: neurotransmitters and neuromodulators* (pp. 1–14). New York: Raven.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M., & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science, 265,* 679–682.

Kobuchi, Y. (1991). State evolution functions and Lyapunov functions for neural networks. *Neural Network, 4,* 505–510.

Kodama, T., Mushiake, H., Shima, K., Nakahama, H., & Yamamoto, M. (1989). Slow fluctuations of single unit activities of hippocampal and thalamic neurons in cats. I. Relation to natural sleep and alert states. *Brain Research, 487,* 26–34.

Kodama, T., Takahashi, Y., & Honda, Y. (1990). Enhancement of acetylcholine release during paradoxical sleep in the dorsal tegmental field of the cat brain stem. *Neuroscience Letters, 114,* 277–282.

Koyama, Y., & Kayama, Y. (1995). Properties of neurons participating in regulation of sleep and wakefulness. *Advances in Neurological Sciences, 39*(1), 29–40.

Maeda, T. (1995). Neural mechanisms of sleep. *Advances in Neurological Science, 39*(1), 11–19.

McCarley, R. W., Greene, R. W., Rainnie, D., & Portas, C. M. (1995). Brainstem neuromodulation and REM sleep. *Neurosciences, 7,* 341–354.

McCormick, D. A., & Price, D. A. (1987). Actions of acetylcholine in the guinea-pig and cat medial and lateral geniculate nuclei, in vitro. *Journal of Physiology, 392,* 147–165.

McKenna, T. M., Ashe, J.H., Hui, G. K., & Weinberger, N.W. (1988). Muscarinic agonists modulate spontaneous and evoked unit discharge in auditory cortex. *Synapse, 2*(1), 54–68.

Moss, F. & McClintock, P. V. E. (Eds.) (1989). *Noise in nonlinear dynamical systems,* Vols 1–3. Cambridge: Cambridge University Press.

Mushiake, H., Kodama, T., Shima, K., Yamamoto, M., & Nakahama, H. (1988). Fluctuations in spontaneous discharge of hippocampal theta cells during sleep-waking states and PCPA-induced insomnia. *Journal of Neurophysiology, 60,* 925–939.

Nakamura, K. (1996). *Development of neuropharmacological technique and its application to neuronal unit recording*. Master Thesis, Graduate School of Information Sciences, Tohoku University.

Nakao, M., Takahashi, T., Mizutani, Y., & Yamamoto, M. (1990). Simulation study on dynamics transition in neuronal activity during sleep cycle by using asynchronous and symmetry neural network model. *Biological Cybernetics, 63,* 243–250.

Nakao, M., Watanabe, K., Takahashi, T., Mizutani, Y. & Yamamoto, M. (1992). Structural properties of network attractor associated with neuronal dynamics transition. *Proceedings of the IJCNN*, Baltimore, MD, Vol. 3, pp. 529–534.

Nakao, M., Watanabe, K., Mizutani, Y., & Yamamoto, M. (1993). Metastability of network attractor and dream sleep. *Proceedings of the ICANN*, Amsterdam, 27–30.

Nakao, M., Honda, I., Musila, M., & Yamamoto, M. (1994). Metastable behavior of neural network under correlated random perturbations. *Proceedings of the ICONIP*, Seoul, pp. 1692–1697.

Peretto, P. (1992). *An introduction to the modeling of neural networks*. Cambridge: Cambridge University Press.

Sato, H., Hata, Y., Masui, H., & Tsumoto, T. (1987). A functional role of cholinergic innervation to neurons in the cat visual cortex. *Journal of Neurophysiology, 58*(4), 765–780.

Sei, H., Sakai, K., Yamamoto, M., & Jouvet, M. (1993). Spectral analyses of PGO-on neurons during paradoxical sleep in freely moving cats. *Brain Research, 612,* 351–353.

Shinomoto, S. (1987). A cognitive and associative memory. *Biological Cybernetics, 56,* 1–10.

Smith, C. (1995). Sleep states and memory processes. *Behavioral Brain Research, 69,* 137–145.

Stewart, M., & Fox, S. E. (1990). Do septal neurons pace the hippocampal theta rhythm?. *Trends in Neuroscience, 13,* 163–168.

Sutton, J. P., Mamelak, A. N., Hobson, J. A. (1992). Network model of state-dependent sequencing. In R. P. Lippmann, J. E. Moody, and S. J. Hanson (Eds.), *Advances in neural information processing systems* 4, *Proceedings of NIPS*'91 (pp. 283–290). San Meteo, CA: Morgan Kauffmann.

Vogel, G. W., Barrowclough, B., & Giesler, D. D. (1972). Limited discriminability of REM and sleep onset reports and its psychiatric implications. *Archives of General Psychiatry, 26,* 449–455.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265,* 676–679.

Yamamoto, M., Nakahama, H., Shima, K., Kodama, T., & Mushiake, H. (1986). Markov-dependency and spectral analyses on spike counts in mesencephalic reticular neurons during sleep and attentive states. *Brain Research, 366,* 279–289.

Yamamoto, M., Arai, H., Takahashi, T., Sasaki, N., Nakao, M., Mizutani, Y., & Kodama, T. M. (1993). Pharmacological basis of 1/*f* fluctuations of neuronal activities during REM sleep. *Sleep Research, 22,* 458.

Yamamoto, M., Nakao, M., Mizutani, Y., & Kodama, T. (1995). Dynamic properties in time series of single neuronal activity during sleep. *Advances in Neurological Science, 39,* 29–40

NOMENCLATURE

α	parameter of an autoregressive process
$\varepsilon_i(t)$	random noise mutually independent, zero-mean white Gaussian noise with an identical variance $\sigma^2(1-\alpha^2)$
σ^2	variance of the noise $\{\xi_i(t)\}$
$\xi_i(t)$	first-order autoregressive process
$u_i(t)$	internal state of the *i*th neuron at time *t*
w_{ij}	synaptic weight from the *i*th to *j*th neurons
$x_j(t)$	output of $j-$ the neuron at time *t*
h	global inhibitory input
N	the number of neurons contained in the network
M	the number of the memorized patterns
$r_\xi(k)$	autocorrelation function of the autoregressive process
$x_i^{(m)}$	the *i*th neuron's state of the *m*th memorized pattern
$\mathbf{x}(t)$	current state vector
$\mathbf{x}^*$	reference equilibrium state vector
$\mathbf{x}_0$	the ''0'' state vector
PSD	power spectral density
DC	direction cosine
MCS	Monte Carlo step

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1303–1316, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00063-4

1997 SPECIAL ISSUE

Consciousness and Neural Cognizers: A Review of Some Recent Approaches

CHRIS BROWNE, RICHARD EVANS, NICK SALES AND IGOR ALEKSANDER

Imperial College of Science, Technology and Medicine

(*Received* 26 *July* 1996; *accepted* 28 *June* 1997)

Abstract—*This paper synthesises three diverse approaches to the study of consciousness in a description of an existing program of work in Artificial Neuroconsciousness. The three approaches are drawn from automata theory (Aleksander, 1995, 1996), psychology (Karmiloff-Smith, 1992; Clark & Karmiloff-Smith, 1993) and philosophy (Searle, 1992). Previous work on bottom-level sensory-motor tasks from the program is described as a background to the current work on generating higher-level, abstract concepts which are an essential part of mental life. The entire program of work postulates automata theory as an appropriate framework for the study of cognition. It is demonstrated how both the bottom-level sensory-motor tasks and abstract knowledge representations can be tackled by a single neural state machine architecture. The resulting state space representations are then reconciled with both the psychological and philosophical theories, suggesting the appropriateness of taking an automata theory approach to consciousness.* © *1997 Elsevier Science Ltd.*

1. INTRODUCTION

This paper brings together three recent approaches in the study of consciousness. A neurally based theory of artificial consciousness (Aleksander, 1995, 1996), a philosophy of mind based on cognitive acts (Searle, 1992) and a theory of human development based on a redescription of learned competences (Clark & Karmiloff-Smith, 1993). The purpose of this paper is to present a neural automata theory approach to consciousness, rather than describe particular machine architectures, which are liable to undergo refinement as the work progresses. Particular details of the design of the experimental work described can be found in the appropriate references.

For continuity and completeness, the paper will commence with a brief recapitulation of the salient propositions of the Artificial Consciousness Theory which are later compared with Searle's philosophy of mind (Searle, 1992). Searle insists that the study of cognition is the study of consciousness, just as the study of biology is the study of life.

> [The brain's] special feature, as far as the mind is concerned, the feature in which it differs remarkably from other biological organs, is its capacity to produce and sustain all of the enormous variety of our conscious life. By consciousness I do not mean the passive subjectivity of the Cartesian tradition, but all of the forms of our conscious life-from the famous ''four f's'' of fighting, fleeing, feeding and fornicating, to driving cars, writing books and scratching our itches. All of the processes that we think of as especially mental—whether perception, learning, inference, decision making, problem solving, the emotions, etc.— are in one way or another crucially related to consciousness. Furthermore, all of those great features that philosophers have thought of as special to the mind are similarly dependent on consciousness: subjectivity, intentionality, rationality, free will (if there is such a thing), and mental causation. More than anything else, it is the neglect of consciousness that accounts for so much barrenness and sterility in psychology, the philosophy of mind, and cognitive science.
>
> The study of the mind is the study of consciousness, in much the same sense that biology is the study of life. Of course, biologists do not need to be constantly thinking about life, and indeed, most writings on biology need not even make use of the concept of life (Searle, 1992).

So, taking Searle's argument that the study of mind is the study of consciousness, this paper sets out to examine neural network models of mind. Having defined the position on consciousness adhered to and claimed that the study of mind is the study of consciousness, all that remains is to define the stance taken on mind in this paper. The mind is assumed to be directly responsible for cognition. Clark & Karmiloff-Smith (1993) make a

powerful distinction between complex information processors, which includes computers, and genuine cognizers.

> The sea slug and the VAX mainframe are both effective processors of information. Yet it is only human beings, and perhaps some higher animals, who are credited with genuine thoughts. Is this mere prejudice, or have we somehow latched on to a genuine joint in the natural order? If it is a genuine joint in nature, what feature of set of features mark it?
>
> The hypothesis to be considered is that there is indeed a joint in the natural order such that humans fall on one side and many other systems (including some quite sophisticated information processors) fall on the other. The joint, we argue, marks a pivotal difference in internal organisation. The representational redescription model embodies specific hypothesis about the nature of this joint (Karmiloff-Smith, 1979a, 1979b, 1986, 1990, 1992). For genuine thinkers, we submit, are endowed with an internal organisation which is geared to the repeated redescription of its own stored knowledge. This organisation is one in which information already stored in an organism's special-purpose responses to the environment is subsequently made available, by the RR process, to serve a much wider variety of ends. Thus knowledge that is initially embedded in special purpose effective procedures subsequently becomes a data structure available to other parts of the system (Clark & Karmiloff-Smith, 1993).

There has been some discussion about the boundaries between congizers and non-cognizers (Aleksander, 1996) which suggests that animals and even machines could redescribe their knowledge. Apart from noting this slight objection, the general principle that redescription is an essential feature of cognition is not at stake, there is only a slight difference in opinion over the position and abruptness of the divide.

In the search for artificial consciousness, the challenge, therefore, is to bestow a neural network with the ability to re-represent its own internal states. Previous work (Browne & Parfitt, 1997) suggests that such a system might well add further weight to the numerous refutations (Smolensky, 1987; Pollack, 1990; Van Gelder, 1990; Aleksander & Morton, 1993) of Fodor and Pylyshyn's attack on connectionist models of cognition (Fodor & Pylyshyn, 1988). The previous work further suggests that recursive redescription of system representations might provide a possible mechanism by which a pseudo-symbolic system, which many agree is the basis of cognitive function, could arise in a connectionist network. The general model of emergent hierarchical data structures presented here also has much in common with Harnad's proposals of the recursive grounding of language (Harnad, 1992).

Before the redescriptive process can be discussed in detail, the type of neural architecture and representational form being proposed must be described. Karmiloff-Smith's representational redescription model (Karmiloff-Smith, 1992) is then presented in detail, moving to a suggested neural architecture with the capacity for spontaneous re-representation of its own internal states. The paper concludes with a discussion of the relationship between Searle's philosophy, the Fundamental Postulate (Section 2.2) of the Artificial Consciousness Theory (Aleksander, 1996) and the proposed cognitive neural architecture which emphasises the importance of examining the mental processes occurring in the system as conscious in some form—the authors would stress an artificial, rather than biological form.

2. NEURAL STATE MACHINE MODELS OF COGNITION

For over 20 years, one of the authors (I.A.) has been suggesting that the capacity for thought can be encapsulated in a machine with an adaptable state structure (Aleksander & Hanna, 1975; Aleksander, 1996). An adaptable state machine can be implemented in a machine with some form of learning capacity. The contemporary versions of such machines have been dubbed as ''Neural State Machine Models (NSMMs)'' (Aleksander & Morton, 1993); the Multi-Automata General Neural Unit System (MAGNUS) (Aleksander et al., 1993) being the most widely used. Since the NSMM is about as far as one can get from much of the philosophical discussion concerning consciousness, it is worth pausing a moment, before pressing on with a description of the Artificial Consciousness Theory to place this proposition in relation to the philosophical discourse. A more detailed account can be found in Aleksander (1996).

2.1. Relationship to a Genealogy

The position taken in the program of work relating to the Artificial Consciousness Theory is that the term consciousness was first used by John Locke when in 1690 he wrote:

> First, I shall enquire into the origin of those ideas, notions, or whatever else you please to call them, which a man observes and is conscious to himself he has in his mind; and the way in which the understanding comes to be furnished with them (Locke, 1690, Ch. i. 4. Reproduced in Locke, 1975).

This can be distinguished from three millennia of concerns with the physical makeup of mind and its relationship to body culminating with Cartesian dualism (DesCartes, 1637) about which Locke and his successors for the next 300 expressed considerable skepticism. Locke's motivation for his study of knowledge is reputed to have come from his view that metaphysics was ultimately futile. He is reported to have turned to friends after an inconclusive debate on metaphysics and asked; ''*Shouldn't we first determine whether we are capable of answering such questions*?'' (Soloman, 1993).

Philosophy from Locke in the 17th century to Wittgenstein at the turn of the 20th attempted to refine the taxonomy of different ''ideas, notions, ...the understanding''. This in itself is not a matter for mathematical analysis but one of philosophical discourse. However, when suggesting that consciousness could be analysed in an artificial neural model it is precisely this taxonomy which has to be borne in mind, as the model must palpably be able to represent ''notions and their understanding'' which are as different as, say, feelings, objective, knowledge, actions, beliefs, etc. It is for this reason that a list of corollaries has been created in the work. Some of the corollaries are described in later sections of this paper; a complete list can be found in the Appendix.

The curious nature of contemporary analyses of consciousness is that they spread over a very wide methodological front. At one extreme is the natural philosophical descent of the lineage mentioned above, where it is the nature of consciousness and not its physical basis which is at stake. Many such discussions (Nagel, 1974; Lucas, 1994, for example) take the position that only the inexplicable subjective nature of consciousness is worth discussing. Giving it a neural basis is therefore at best uninteresting and at worst irrelevant to philosophical concerns. Others (Penrose, 1994, for example) believe that consciousness is beyond that which can be achieved through current mathematical analyses, particularly neuronal computations. He advocates a solution based on the function of microtubles in the brain cells causing effects which could only be described by a non-computable extension to quantum theory which includes the gravitational force. Most analyses, however, which actually attempt to find a neural basis of consciousness operate in a domain where effects observed in the human brain are correlated to conscious experience reported by or inferred in the owner of that brain. Greenfield (1995) and Crick (1994) are major contributors to this methodology. Even further away from classical philosophy is the work of computationalists such as Dennett (1991) and Sloman (1994) who argue that consciousness is a product of the architectural character of the brain which, for Dennett, can be expressed in computational terms such as virtual machines and interacting agents.

The approach taken in this paper, described in detail in (Aleksander, 1996), is founded on iconic representation and its consequences, which suggests that consciousness could be found in neuronal models which are neither structured like the brain, nor are they computationalist in the sense that they represent a programmer's idea of what the elements of consciousness are. The approach is constructivist only in the sense that it looks for emergent properties of a very large structure, but relates such properties to the global parameters of that structure. The only construction is the fundamental postulate, the rest is consequential or a matter of refinement. This is as far removed from the philosophical lineage as one can get, as a philosopher might immediately argue that such structures are disqualified for not being the flesh and blood for which consciousness makes sense. But the aim of the rest of this paper is to illustrate that this constructivist approach *does* provide a general perspective on the synthesis of consciousness, while respecting the genealogy of the subject. A synthesised consciousness might in itself, if Searle is correct, be at the root of building truly cognitive machines.

2.2. The Fundamental Postulate

The Fundamental Postulate (Aleksander, 1996) states the following:

> The personal sensations which lead to the consciousness of an organism are due to the firing patterns of some neurons, such neurons being part of a larger number which form the state variable of a neural state machine, the firing patterns having been learned through a transfer of activity between sensory input neurons and the state neurons.

The Fundamental Postulate posits, as an appropriate framework in which to examine cognitive processes, an automata theory in which the state variables have neural learning properties, and also that neural states which contribute to consciousness must be created through a process which has been described as *iconic learning*.

2.2.1. *Iconic Learning.* Iconic learning is a representational scheme proposed in previous work by one of the authors (Aleksander & Morton, 1993) and others (Harnad, 1987b). The representational scheme has two salient properties which are; (1) preservation of the structural form of the environment through the direct transfer of activity from sensory neurons to the state neurons responsible for consciousness and (2) the representations are concatinatively combinatorial.

Harnad proposes iconic representations for discrimination, which he defines as, ''*[T]he ability to judge whether two inputs are the same or different, and, if different, how different they are*'' (Harnad, 1992). Identification, on the other hand, is concerned with categorisation and requires a different representational form. As opposed to discrimination, which is a relative judgement between objects, identification is an absolute judgement as to which category a particular perception belongs to. Contrary to some theorists (Paivio, 1986, for example), Harnad maintains that iconic representations are insufficiently distinguishable for reliable identification (Harnad, 1982, 1987b). He proposes a system of ''categorical perception'' for reliable identification (Harnad, 1987a, b) which utilise categorical representations. Categorical representations are the product of category specific feature detectors which filter out information[1] in the iconic representations which are not

directly involved in the differentiation between members of a category and nonmenmbers (Harnad, 1990).

Harnad (1987b) proposes that the categorical representations only lose information which is irrelevant to the disambiguation between category members and nonmenbers within the existing context of alternatives. The issue here is that the features which differentiate a cat from say a dog are different from those which distinguish between a cat and a tiger, which are again different from the set of features which between a cat and an elephant. In other words, the category features vary depending on the particular features of the class of nonmembers with which the object for categorisation is compared. In order to accommodate a large range of confusable alternatives the category representations must maintain a great deal of the structural information about the environment, but in a compressed form. Since the structural form of the environment is to largely be maintained, then the resulting representational form is better dubbed as *reduced iconic representations* within the current discourse to highlight their compatibility with the Fundamental Postulate (Section 2.2). Harnad (1987b) describes the proposition as follows:

> Note that [the conversion from iconic to categorical representation] is still analogue, for it continues to preserve some of the physical ''shape'' of the input, albeit in a highly reduced and abstracted form. Iconic/Categorical conversion is really a transformation from macro-icons to micro-icons, the former preserving the context independent, holistic, configural properties of the proximal stimulus, the latter selectively reduced to only those properties that are invariant in a specific context of confusable alternatives (Harnad, 1987b).

2.3. State Space Representations of Knowledge

State spaces are the currency of automata theory. Two of the authors (Sales, 1996; Evans, 1996) have examined the properties of learned state spaces in neural state machines as models of cognition. The work was undertaken using a simulated, virtual, environment called *kitchenworld*.

2.3.1. *The Kitchenworld Environment*. The *kitchenworld* arrangement (Figure 1) consists of MAGNUS (Multi-Automata General Neural Unit System) exploring and naming objects in a virtual environment called *kitchenworld*. The *kitchenworld* consists of objects such as glasses, plates, cutlery, cups, bottles and the odd apple. The digitised image, available to the MAGNUS simulation, was created from a real image taken from a video camera. Since both MAGNUS and the *kitchenworld* are

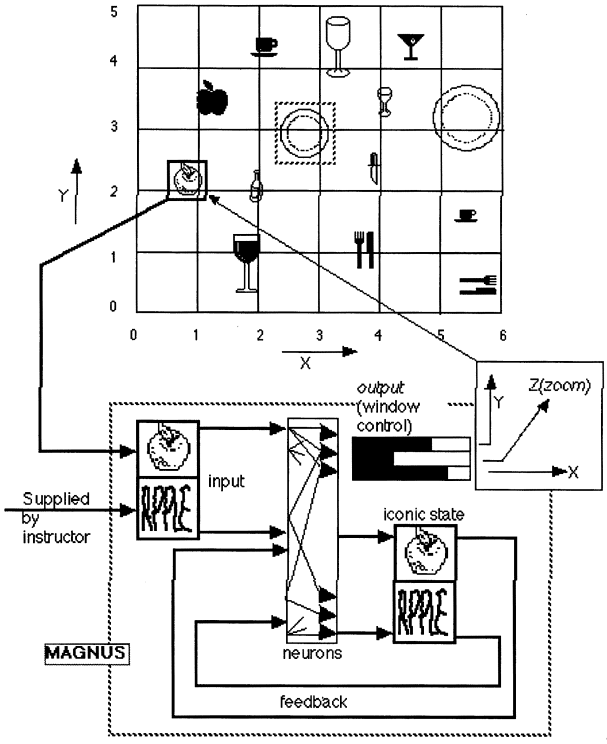


FIGURE 1. MAGNUS in a kitchen world.

simulated, it is important to distinguish between them. Everything within the large dotted frame (Figure 1) is the MAGNUS program. Everything else, including the picture, is another program—the *kitchenworld* program.

The two interact as follows. At its output, MAGNUS issues signals to *kitchenworld* which controls the position and size of a window placed over the picture. In the example (Figure 1) the window is shown to be over an apple. Another dotted position for the window is shown over a plate. So MAGNUS is to derive how much to move in the two directions X and Y, and how big to make the size (Z) of its window on the *kitchenworld*. The output of the MAGNUS (Figure 1) consists of three bars, one each for X, Y and Z. These can be thought of as "strengths of firing of three separate regions of the neural state machine which determine the position of the window.

Given the position and size of a window, the *kitchenworld* delivers to MAGNUS the content of the window as a thresholded binary pattern. This pattern forms one part of the neural state machine input. It can be seen as a visual input—MAGNUS current view of the *kitchenworld*. The second part of the system input can be seen as linguistic. It consists (Figure 1) of a binary pattern in the form of a word. This pattern is selected by an instructor and is considered to be a name for the object currently in view. The system is, therefore, exposed to two sensory events in different modalities, visual and linguistic, which occur in close time proximity; the word input is deliberately chosen and presented to coincide with the visual stimulus.

[1] Harnad (1990) actually uses the word features, but following correspondence between one of the authors (C.B.) and Harnad, the author believed that information is a more apt description.

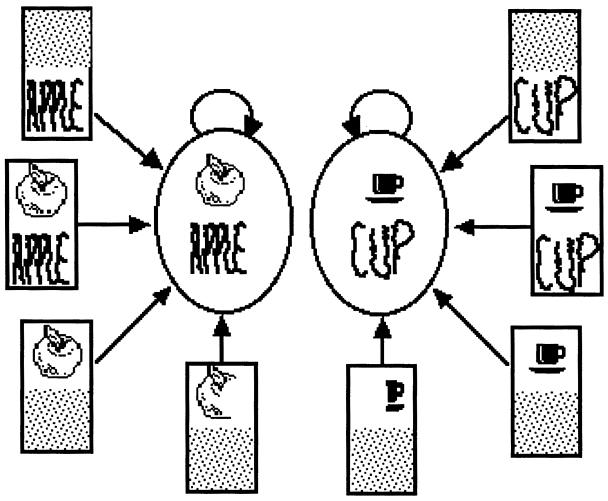


FIGURE 2. State characteristics for two named objects.

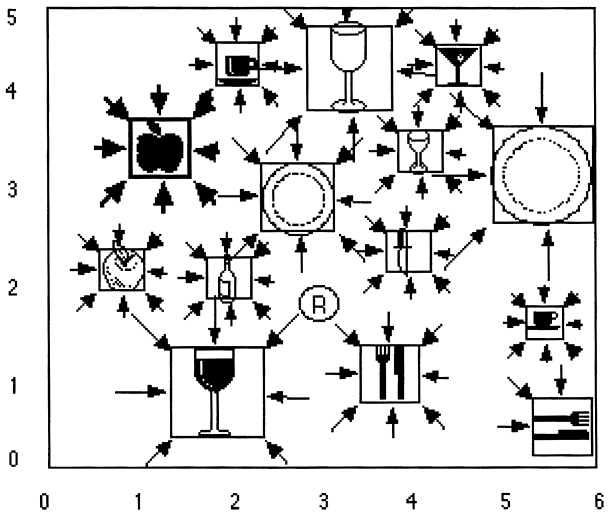


FIGURE 3. Attractor objects and a repeller background in the kitchen world.

2.3.2. *Simple Object Naming.* In previous work (Sales, 1996), a model of simple object naming was constructed, which verified Corollary 8 (Aleksander, 1996).

2.3.2.1. Corollary 8: Representation of meaning.

> When sensory events occur simultaneously or in close time proximity in different sensory modalities, iconic learning and generalisation of the neural state machine ensures that one can be recalled from the other (Aleksander, 1996).

The MAGNUS visual input was centred on an object in the *kitchenworld*, the framed apple (Figure 1) for example, and simultaneously presented patterns corresponding to the written name for the object to the linguistic input. MAGNUS, a generalising neural state machine, was then iconically trained on the pair of inputs. The neural state machine learned attractor states which consisted of both the visual and linguistic inputs (Figure 1). It is known that in common with any neural model, local generalisation in the logical neurons used in MAGNUS leads to the property of state reconstruction (Aleksander & Morton, 1991). This means that either the name input or the object input will lead to the reconstruction of the state representing both (Figure 2). In other words, an input from a single modality, with the other set to random noise, will lead to the recall of a state consisting of the iconic representations of the sensory experience of both modalities. The experience of each modality can be reconstructed from the other, as proposed in Corollary 8.

The learning of simple object names discussed here is open to the criticisms raised by philosopher's such as Wittgenstein (1953) and Quine (1960). The problem, in short, concerns the underdeterminance of language. The issue at stake is how a child learning a language is to determine which part of the visual scene relates to the object being named. The refutations of these criticisms lie in the formation of group nouns for objects, which is a more complex extension of the work described in this section and is discussed in the section relating to that (Section 3.1.1).

2.3.3. *Locating Objects.* Previous work (Evans, 1996) used the MAGNUS system in the *kitchenworld* environment to examine the capacity of an iconically trained neural state machine to locate objects. Object location is crucial if it is to be understood how a conscious organism might partition its sensory world into objects that need naming. It has been demonstrated (Evans, 1996) that it is possible to train a window to locate on a particular object such as an apple (Figure 1). Location on an object is achieved by displacing a window, which is perfectly positioned over an object, in both location and size, and training the network to output a move towards the correct position. The interesting part of this work is that having done this for the apple, as shown, not only the apple becomes an attractor in the image space, but so does any other object. A close look shows that the system is interpreting *any disturbance from the background* in the visual input as a pattern which causes an output that moves the movable input window towards the object. So when the input window covers a part of an object, it will tend to be attracted towards that object. Curiously, this creates not only attractors where the objects are, but also ''repellers'', such as R (Figure 3) in areas where there are no objects.

The experiments demonstrate that it comes naturally for a neural state machine with a visual input and a motor output to develop a general notion of finding an object by treating it as a disturbance in a background after minimal instruction.

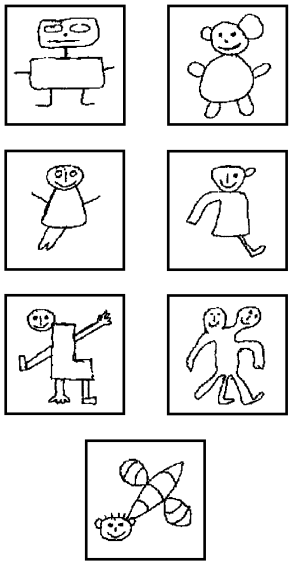


FIGURE 4. Example drawings inspired by children's drawings of a man who does not exist. After (Clark & Karmiloff-Smith, 1993). Taking the drawings in rows from the top left and denoting the ages of the artists as years; months, then the ages are as follows (a) 4;8 (b) 5;9 (c) 5;3 (d) 9;0 (e) 9;8 (f) 9;6 (g) 9;8.

3. REPRESENTATIONAL REDESCRIPTION

The tasks presented so far have all concerned bottom-level, sensory-motor processing. Earlier, we suggested that cognition only arose through a process of redescription of lower-level representations into higher-level ones. That process is now described in detail.

The representational redescription (RR) model has been developed by Karmiloff-Smith over a period of time. A comprehensive description is given in Karmiloff-Smith (1992). At the heart of the RR model is an internal process of representational change by which the mind enriches itself.

The RR model posits a number of levels. It should be stressed that RR is not a Piagetian stage model. RR is presented as a domain general process, but importantly it is posited to act on domain specific representations. Three levels of the RR model are of interest to the connectionist modeller.

The first level of implicit, or level-I, representations is the level achieved by connectionist models of behaviour. At this level stored representations are activated by external stimuli, but are not available to the system in an explicit re-usable format (Karmiloff-Smith, 1992; Browne & Parfitt, 1997). Beyond level-I is the first level of explicit representation of knowledge, level-E1. At level E1 representations are not available to consciousness or verbal report, but they are available without the need for external stimuli and to other domains. At higher levels of redescription, levels E2/E3 [also denoted by E + (Clark & Karmiloff-Smith, 1993)] the knowledge is available to conscious access and verbal report.

Karmiloff-Smith has undertaken a number of developmental studies in order to investigate changes in children's representations (Karmiloff-Smith, 1992). These studies bear out the predictions of the RR model. A particularly salient study is one concerning drawings (Karmiloff-Smith, 1990).

Six drawings were produced by four- to 11-year-olds. The children were told the subjects of the drawings. First a house was requested, then a house which does not exist. Similar pairs of drawings for a man and an animal were asked for. Only the drawings of men (Figure 4) are considered here.

The motivation behind asking the children to draw a basic object, succeeded by a version of that object that did not exist was as follows. The standard object could be produced by the simple execution of an action schema. To produce the fictitious drawing, however, the children would be forced to break down their action schema for drawing the regular objects in order to resize, reshape, omit or swap components. The hypothesis, which was borne out, was that young children with only level-I representations would be unable to draw the fictional manifestations of the objects. Their implicit only representations constrained these children to merely reproducing the entire schema, despite verbal report that they were about to draw ''a man who does not exist''. The components of the schema were not available as explicit information available to the children for manipulation.

Of greater interest was the difference in drawings produced by four- to six-year-olds and eight- to 10-year-olds. The imaginary drawings from the former were created by altering the size and shape of components and some deletions of components. The deleted components were often the last one to be drawn in the undistorted pictures. The later group's fictitious drawings showed a much greater variety of manipulations from the original. The components of the original images were manipulated as independent entities, leading to insertion of additional elements from the same conceptual category, change of position and/or orientation of whole elements, or even the insertion of components from a different conceptual category.

The children's drawing study demonstrates the underlying thesis of RR, which is that knowledge which is implicit in the original action schema, in this case the action schemas for drawing body components, gradually becomes more explicit through an internally generated process of representational redescription.

Another important feature of the RR process is that it is recursive. The redescribed representations must themselves be capable of undergoing further redescription. As more levels of redescription occur, the knowledge becomes more abstract and more accessible. As stated above, previous work suggests that it is the recursive nature of RR which may lead to an insight as to how

a pseudo-symbolic system of cognitive representation is constructed from low-level processing (Browne & Parfitt, 1997).

3.1. Redescription in State Spaces: Complex States

Previous work (Aleksander, 1996) suggests that the bottom-level sensory-motor, simple states learned in tasks such as those studied above (Section 2.3) [Evans, 1996; Sales, 1996] are equivalent to Karmiloff-Smith's bottom-level, implicit (level-I) representations (Karmiloff-Smith, 1992). He further suggests that higher-level, complex states can be formed which are equivalent to the first level of explicit, abstract knowledge; level-E1.

3.1.1. *First Level of Redescription.* To illustrate the notion of complex state representations of abstract concepts, a particular example which makes reference to the *kitchenworld* environment will be used. Within the *kitchenworld* there are a number of repeated instances of various objects (Figure 1). There are two apples, four glasses, two cups and two plates. When the objects are named like objects will receive identical names; both cups will be labelled as ''cup'', each glass as ''glass'', etc. Since the object naming function of the system allows the visual input to be recalled from the linguistic label, the multiple visual input for each linguistic label might now appear to cause a conundrum. Which visual input is to be recalled in association with the linguistic label? The dilemma can be resolved by adding a probabilistic element to the system (Browne & Aleksander, 1996). This probabilistic element allows the system to explore regions of state space where the input does not explicitly define a particular state. In the case where the visual input is undefined and the linguistic input is a recognised label, the probabilistic state space exploration allows all the states which match the linguistic input to be explored (Figure 5). During the exploration of state space the states corresponding to several, or even all, the visual inputs are visited. The linguistic labels have now become class nouns, representing the commonality between a cluster of bottom-level, learned states. These clusters of state can now be represented as single, complex states [Figure 6(a)].

There is some controversy as to whether the complex states described are adequate representations of class nouns. The deficiencies of this approach are discussed in a later section. For the current argument, however, which is based on the linguistic construction of class nouns the described complex states are arguably sufficient, if recourse is made to Wittgenstein's proposition of the non-essentialism of language (Wittgenstein, 1953). Wittgenstein argued that there are no hidden, or ultimate, universals in language. He suggests that a word such as ''dog'' derives its meaning from a mental image, the

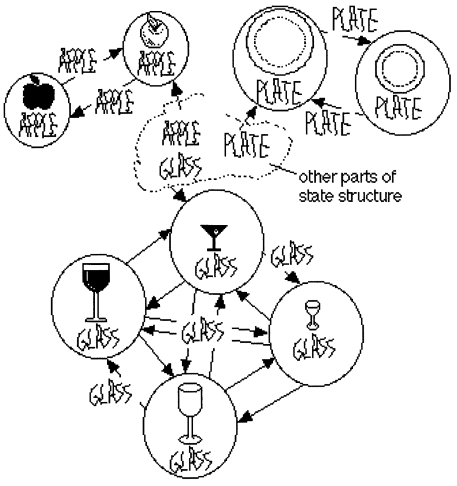


FIGURE 5. Concepts in state space. The first level of redescription.

visual modality region of the described complex states, and that the image must be derived from only the set of dogs which the person in question has experienced. The simple states which together form the complex state are all derived directly sensory experience. This is ensured by the insistence on iconic learning. It is in this retreat from the realism of word meaning, which is arguably what prompted Wittgenstein to write ''*Philosophical Investigations*'' (Wittgenstein, 1953), that the concerns of Quine (1960) and particularly Wittgenstein (1953), noted above (Section 2.3.2) are refuted. The argument stated here is expanded in detail by both Harnad (1995) and Searle (1992). Searle argues in particular that Quine's theory (Quine, 1960) is only appropriate for a zombie robot, which conscious human beings are clearly not.

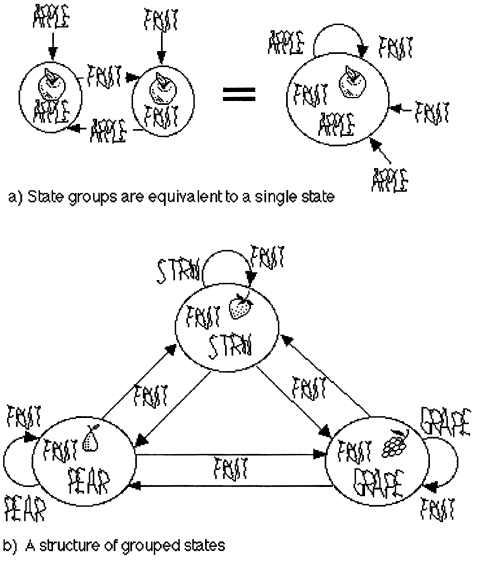


FIGURE 6. State grouping to form concepts (a) and higher-level concepts (b).

3.1.2. *Second Level of Redescription.* Karmiloff-Smith (1992) demands the redescribed states must themselves be available for redescription. Continuing with the previous example where *kitchenworld* states are redescribed through the addition of linguistic tags, the second level of redescription can be driven by providing new, equally valid but broader domain, linguistic tags for known objects. Moving slightly away from the actual *kitchenworld*, assume that a system is exposed to an environment consisting of a number of fruits; strawberries, pears and grapes for instance. Assume also that the neural state machine has learned the names of the individual fruits, so it has complex states representing the class nouns strawberry, pear and grapes. Now say the system's visual input is centred on one of the strawberries and the word fruit is input to the linguistic modality. The visual input of the strawberry will cause a region of the complex state representing the noun strawberry to be reconstructed and learning will combine the new linguistic tag, fruit into that state. A new complex state with two linguistic tags, strawberry and fruit, for the associated visual stimuli will be created. If the word ''fruit'' were then associated, in an identical manner, with the complex states representing pear and grapes, then a set of states which are unified by a common linguistic tag [Figure 6(b)] is created, which is identical in form to the set of simple states which combined to create the initial complex states [Figure 6(a)] except that the lower-level states are themselves complex states, rather than the bottom-level, sensory-motor states of the previous example. In other words, the more general group noun ''fruit'' has allowed the first level complex states representing the group nouns for individual fruits to be recursively redescribed at a higher level. This recursive redescription can be seen to be constructing a general, hierarchical information structure.

3.2. Inadequacy of Language Based Redescription

The proposed method of representational redescription gives an excellent demonstration of the applicability of complex states to the task. The redescription process, as currently described, however, requires external linguistic input. Karmiloff-Smith is quite prepared to accept that explicit knowledge can arise through verbal instruction, but she insists that this is not true representational redescription. That, she insists is driven by an internal, spontaneous process.[2]

The possibility of deriving internally driven group representations, their form and the implications on the system are the topics of the next section.

[2] Interestingly, Harnad has recently shown the use of language in conveying explicit knowledge to a group which creates an evolutionary advantage over groups with no language where each member must derive all their explicit knowledge for themselves (Harnad, 1996).

4. INTERNALLY DRIVEN REDESCRIPTION IN A SINGLE NEURAL ARCHITECTURE

The task of building a single neural architecture with an internally driven redescriptive mechanism can now be split into two questions. (1) Given the explanatory power of the suggested complex state approach to redescription, what form of internal representation might replace the externally provided class nouns? (2) How could a single neural architecture derive the appropriate representations?

4.1. Representational Replacement for Class Nouns

Since representational redescription is posited as a recursive process, the appropriate representational form for the abstracted concepts must be highly related to that proposed for the bottom-level; i.e. iconic learning. Fodor and Pylyshyn argue, correctly, that the token[3] representing an higher level concept should consist of the systematic combination of the items from which it is comprised (Fodor & Pylyshyn, 1988).What is at stake is whether systematic combination of tokens can only be achieved in computational data structures, or whether a neural network is capable of such functionality. Pollack has clearly demonstrated that a connectionist system is capable of systematic combination (Pollack, 1990). The suggestion presented here is an extension and refinement of Pollack's recursive auto-associative memory (RAAM).

The concept of glass has already been established to be grounded in the sensory experience of all the glasses a person has seen. It would seem appropriate to construct the representation of the concept from this sensory experience. Doing so would satisfy both the criteria for systematic combination of the tokens from which the concept is constructed and the demand for iconic representation.

The most obvious technique of combining the sensory iconic representations of the four glasses (Figure 5) which have been experienced would be to superimpose them. This technique has two serious inadequacies; (1) the resulting representation is liable to be amorphous and (2) the amorphous nature of the representation makes it even more open to Harnad's criticism that it is insufficient for identification (Harnad, 1990).

4.2. Internally Generating the Appropriate Representations

Since class nouns represent categories, the most obvious candidate for the appropriate representational form is

[3] Token, rather than icon, is used in this context as reference is being made to Fodor and Pylyshyn's work which is symbol based. The author would strenuously argue that arbitrary symbols are insufficient for cognition and that iconic representations are appropriate (Aleksander & Morton, 1993; Aleksander, 1996). Throughout this section, if the word token is used in relation to a connectionist system it should be read as icon. Token is used for consistency and chosen for its generality. Icons are a subset of tokens, thus token is universally applicable.

Harnad's categorical representations (Harnad, 1987b) described above (Section 2.2.1) and dubbed as reduced iconic representations for clarity. Such reduced descriptions are entirely commensurate with Hinton's proposals (Hinton, 1988).

4.2.1. *Learning Rules.* Two learning rules, Bienenstock, Cooper and Munro (BCM) [Bienenstock et al., 1982], and Local Error-driven and Associative Biologically Realistic Algorithm (LEABRA) [O'Reilly, 1996] overtly display representational reduction whilst maintaining structural integrity. Previous work (Browne & Parfitt, 1997) describes the features of both these learning rules and posits them as ideal candidates for the learning rules underlying a neural system capable of representational redescription. O'Reilly gives an excellent account of the inadequacy of backpropagation in maintaining the structural form of the environment (O'Reilly, 1996) and so it is not considered further as a learning rule.

Intrator (Intrator, 1992) has applied exploratory projection pursuit methods in order to derive a modification to the original BCM learning rule (Bienenstock et al., 1982). The technique results in an unsupervised learning rule for input to output dimensionality reduction which seeks projections that emphasise distinguishing features. Such a learning rule is ideally suited to creating reduced iconic representations.

4.2.2. *Approximationism.* It is important to note that the features extracted by Intrator's (Intrator, 1992) learning rule emphasise distinguishing features in only the data presented. It is therefore possible that a feature which distinguishes two categories for a portion of the data is later proven to be inadequate for the task and therefore requires refinement. Harnad (1987b) frames the problem as follows:

> Consider a simple problem in machine vision: Suppose all that a visual scene categoriser had to do was to tell apart trees from animals, that is, to categorise all instances of trees as trees and all instances of animals as animals. Suppose, by way of further simplification, that trees and animals were the only patterns the analyser ever encountered, and suppose (to simplify still further) that its input patters were already suitably smoothed and parsed so that they only appeared in standard positions, with figure and ground, parts and whole, already appropriately sorted.
>
> Now it is evident that if there were nothing to worry about but sorting trees and animals under these canonical conditions, a very simple rule would work quite well, for example, counting the number of legs, L, and calling an instance a tree if L was less than or equal to 1 and an animal otherwise. Obviously such a rule could only work with well smoothed and parsed inputs, free of anomalous instances (such as storks standing on one leg, trees with split trunks, or tables). As an approximation, however, the rule would sort the standard cases described. As anomalies were introduced, the rule could be revised and elaborated so as to tighten the approximation in accordance with the new contingencies (Harnad, 1987b).

Harnad (1987b) refers to this effect as ''Approximationism''. The suggestion being that a categorical representation is not a fixed absolute, but an ever improving approximation driven by the nature of the context; being the noncategory members the category members have to be differentiated from.

4.2.3. *Construction of Category Representations Occurs Over Time.* The implication of approximationism for the internally generated category representation is that this must occur over time. The feature extraction system will develop features which distinguish category members from previously encountered nonmembers, but as anomalies or new items which from other nonmember categories are experienced the set of reduced features which defines the category representation will require refinement or addition.

4.2.4. *Neural Architecture Capable of Building Appropriate Category Representations.* Previous work (Browne & Parfitt, 1997) has suggested a two unit neural system (Figure 7) for producing internally driven category representations. The first network extracts the category features from sensory stimuli using Intrator's modification (Intrator, 1992) to the BCM rule (Bienenstock et al., 1982). The second network then recombines these extracted features as they are extracted over time by means of its feedback loop in much the same manner as Elman's Simple Recurrent Network (Elman, 1990) constructs temporal relations in the context units. If, in contrast to Elman (1990), who uses error backpropagation (Rummelhart et al., 1986), a structure preserving algorithm, such as the two suggested above (Section 4.2.1) is used in the second network, then the evolved category representation maintains its iconic form as desired.

5. COMPARISON WITH OTHER APPROACHES

It is certainly not intended to claim in this paper that automata theory is the only approach to understanding consciousness. The aim of this discussion is to argue that the described automata based theory does provide a useful theoretical framework for exploring consciousness, or at least the particular aspects focussed on here; redescription and abstraction. It is, therefore, important to note the essential features of iconically trained neural state machines that differentiates them from other approaches. The authors are perfectly willing to accept that other approaches may satisfy the fundamental postulate and corollaries. The theory is, however, couched in terms of state spaces which fits naturally with neural state machines. The appropriateness of

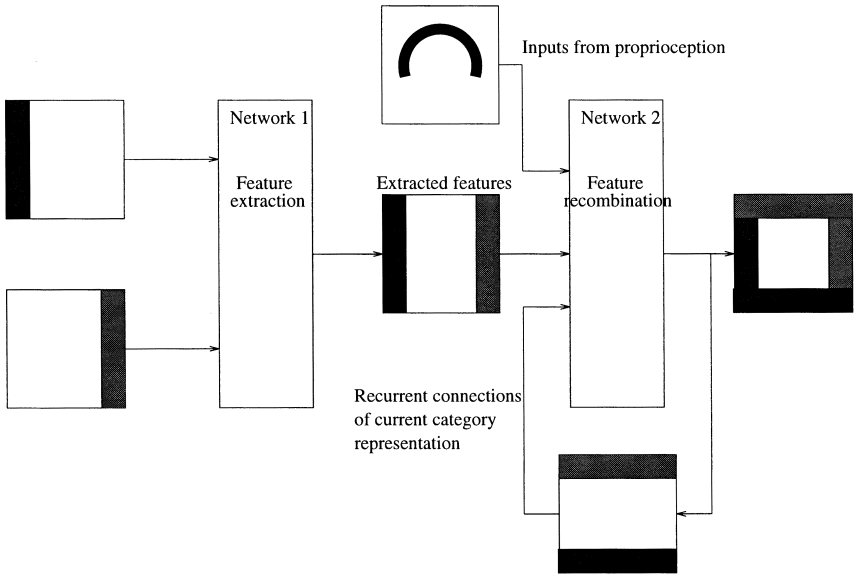


FIGURE 7. A schematic of a suggested system to learn the concept of rotation through the redescription of iconically learned states. The first network extracts features from the currently presented views (in this case a bar down the left in the upper image and a bar down the right in the lower image). An intermediate state is formed from the combination of the features extracted from the current training example. These sets of features extracted from a number of individual pairs of training examples are combined, over time, in the second (right hand) network, to form a representation of the overall concept. This representation consists of the combination of all the features extracted from individual instances of the concept.

automata theory to the study of consciousness is of little surprise considering the fact that it was developed against a background of the Macy Conferences on Feedback Mechanisms and Circular Causal Systems in Biology and the Social Sciences, at which the question of mechanistic consciousness was raised at an early stage in the development of computing (Asprey, 1990).

It is first important to point out that the iconic training is an essential feature. Iconic training ensures that the system is not subject to the symbol grounding problem. The symbol grounding problem relates to the absence of intrinsic meaning in a system of arbitrary symbols, the currency of classical Artificial Intelligence. Searle (1980) provided one of the original demonstrations of the issue in what he called ''*The Chinese Room*'' thought experiment. Harnad has reformulated the idea in a clearer form by describing the never ending search for meaning by a non speaker attempting to translate a Chinese text from a Chinese-to-Chinese dictionary (Harnad, 1990). Both these arguments are considered to be powerful criticisms of any claim computationalism has to modelling cognition or consciousness. Harnad takes the ideas a stage further using them to refute the claim that cognition is computation (Harnad, 1994). The authors, therefore, suggest that symbolic Artificial Intelligence, as it currently stands, is not an appropriate model of consciousness.

Neural networks, on the other hand, derive their representations directly from the environment and are therefore less susceptible to the symbol grounding problem. It is not true to say that neural networks are immune from the problem. If the neural network uses a localist representation at any stage of processing, then the environmental structure present at the input is lost and the symbol grounding problem re-emerges. However, there are a large number of neural network models which employ distributed representations throughout. Further discussion is limited to these systems.

Above (Section 1), it was stated that Fodor and Pylyshyn's (Fodor & Pylyshyn, 1988) objections to connectionism had been refuted. Although this is true, it has not as yet been fully demonstrated how a connectionist network might generate flexible representations which are available throughout a system (Clark & Karmiloff-Smith, 1993). Current neural networks are capable of generating concept trees (Pollack, 1990), but this still does not describe the full capacity for internally driven, recursive concept abstraction demanded by Clark & Karmiloff-Smith (1993).

The failure of both traditional Artificial Intelligence and connectionist approaches in cognitive modelling has lead, in an attempt to find a successful compromise, to the development of a number of hybrid systems. Implicit in hybrid systems is the hope that the juxtaposition of the two approaches will alleviate the inadequacies of both. Unfortunately, what is equally as likely is that the resultant system will suffer from the deficiencies of both connectionism and computationalism. Symbol grounding demands that there is a clear, consistant process by which higher-level, more abstract symbols are generated from lower-level ones. The break in the representational scheme between the symbolic and

connectionist regions of a hybrid system means that they are still susceptible to the symbol grounding problem, if at a higher-level. Also, the tokens utilised by symbol processing region of the system are constrained by the connectionist system, so may well be far from ideal for the required computation.

What is required is for either a symbol system to uneqivocably overcome the symbol grounding problem, or the development of a connectionist system with the capacity to flexibly manipulate its representations. It is suggested that the automata theory approach described here does provide a clear view as to how the latter type of system might be constructed. That is not to say that other approaches are incapable of some, or even all, of the functions outlined by the corollaries. The automata theory framework taken as an overall approach, however, does stand out, as it does provide useful insight into cognitive functioning which other approaches have so far failed to clearly reveal.

What the automata theory approach described in this paper offers can be summarised as a clear view as to how a neural system might build up a set of internally generated, hierarchical, flexible representations in a number of domains. Such representations have been claimed to be at the heart of cognition (Clark & Karmiloff-Smith, 1993) and are later argued to be commensurate with a contemporary philosophy of mind (Section 6).

6. CONCLUSIONS: WHERE DOES CONSCIOUSNESS COME IN?

A fair critique of the above would be that all that has been suggested is method for learning to represent worlds containing nested concepts, so why have a fanciful reference to the difficult and woolly concept of consciousness? In fact, why refer to a program of work which goes under the heading of Artificial Consciousness at all? The answer lies in the fundamental postulate and therefore its associated corollaries. A neural state machine has been shown to satisfy the requirements of corollaries 4, iconic learning and 8, the representation of meaning. Corollary 4 relates to consciousness by suggesting that iconic transfer creates ''sentience'' by making neural firing patterns meaningful in terms of the organism's sensory world. It is this corollary which gives the proposals immunity from the symbol grounding problem, discussed above.

Corollary 8 is discussed as being the key to the representation of abstractions such as place and category within a neural state machine.

Of course, any attempt at a neural construction of the properties of consciousness can be dismissed purely on the grounds that some definition of consciousness can be found which excludes the activity altogether either for not being biologically relevant or for being a philosophical error of category as is the belief of some of the authors discussed above (Section 2.1). It is arguable, however, that the model developed in this paper is true to the style of enquiry advocated by Locke [as quoted above (Section 2.1)]. The principle of an iconic transfer comes from the ''consciousness to one's self'' of ''ideas, notions or whatever else you please to call them...''

Of all the contemporary philosopher's, Searle is possibly the closest to a continuation of the line of enquiry commenced by Locke. Searle's position is well defended in Searle (1992). It can be summarised as follows. Consciousness is an emergent property of the brain. We do not understand how this occurs as yet, but advances in science will unravel the mystery in time. For the present we should just accept that one of the functions of the brain is to cause consciousness. Consciousness is inextricably bound to mental life. The contents of consciousness are mental phenomena. Unconscious mental phenomena are only describable in terms of their similarity to conscious ones. Although unconscious mental phenomena may never become the content of consciousness they must have an identical form. Attributes of the brain which can never enter consciousness are nonconscious. Searle differentiates unconscious from nonconscious mental phenomena as follows:

> Think of the difference, for example, between my belief (when I am not thinking about it) that the Eiffel Tower is in Paris, and the myelination of the axons in my central nervous system. There is a sense in which both are unconscious. But there is a big difference between them in that the structural states of my axons couldn't themselves be conscious states, because there isn't anything mental about them. I assume for the sake of this argument that myelination functions essentially in the production of my mental states, but even if myelinated axons were themselves objects of experiences, even if I could feel inwardly the state of the myelin sheaths, still the actual structures are not themselves mental states. Not every unconscious feature of my brain that (like myelination) functions essentially in my mental life is itself a mental feature. But the belief that the Eiffel tower is in Paris is a genuine mental state, even though it happens to be a mental state that is most of the time not present to consciousness. So here are two states in me, my belief and my axon myelination: both have something to do with my brain and neither is conscious. But only one is mental, and we need to get clear about what makes it mental and the connection between that feature—whatever it is—and consciousness. Just to keep this distinction clear, I propose in this chapter to call phenomena like myelination, which are not in the mental line of business at all, ''nonconscious'' and phenomena like mental states that I am not thinking about or have repressed ''unconscious'' (Searle, 1992).

In the neural state machine approach to consciousness, the fundamental postulate defines the current system state as the existing conscious state. Other iconic states in the learned state structure are possible candidates for becoming the system state, so in Searle's terminology are

unconscious mental phenomena. Since the neural state machine approach to consciousness is defined in terms of state spaces, little reference need be made to the actual structure of the system, although it is worth mentioning that the MAGNUS simulation (Section 2.3) described above is extremely adept for the task, which comprise Searle's nonconscious mental phenomena need not be studied in detail.

The insistence on iconic learning provides the system with both symbol grounding (Searle, 1980; Harnad, 1990) at the bottom-level and a maintenance of the aspectural shape (Searle, 1992) of the higher-level mental states.

In this light, the items discussed above are well within the realms of the study of consciousness and might prove to be a fruitful approach to a complex problem.

REFERENCES

Aleksander, I. (1995). Artificial neuroconsciousness: An update. In J. Mira & F. Sandoval (Eds.), *From natural to artificial neural computation*. (pp. 566–583). Heidelberg: Springer.

Aleksander, I. (1996). *Impossible minds: my neurons, my consciousness*. IC Press.

Aleksander, I., Evans, R. & Penny, W. (1993). MAGNUS: An iconic neural state machine. In *Proceedings of the Weightless Neural Networks Workshop*.

Aleksander, I. & Hanna, F. K. (1975). *Automata theory: An engineering approach*. New York: Crane Russak.

Aleksander, I., & Morton, H. (1991). General neural unit: retrieval performance. *Electronic Letters*, *27*, 1776–1777.

Aleksander, I. & Morton, H. (1993). *Neurons and symbols*. Chapman and Hall.

Asprey, W. (1990). *John von Neumann and the origins of modern computing*. MIT Press.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specifity and binocular interaction in the visual cortex. *Journal of Neuroscience*, *2*, 32–48.

Browne, C., & Aleksander, I. (1996). Digital general neural units with controlled transition probabilities. *Electronics Letters*, *32*(9), 824–825.

Browne, C. J. & Parfitt, S. (1997). Iconic learning and epistemology. In *Does representation need reality? Proceedings of the International Conference New Trends in Cognitive Science*. Vienna: Austrian Society of Cognitive Science Report 97-01.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards. *Mind and Language*, *8*(4), 487–519.

Crick, F. (1994). *The astonishing hypothesis*. New York: Scribner's.

Dennett, D. C. (1991). *Consciousness explained*. Penguin Books.

DesCartes, R. (1637). *Discourse on method*.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Evans, R. G. (1996). *A neural architecture for a visual exploratory system*. PhD thesis, Imperial College of Science, Technology and Medicine.

Fodor, Z. W., & Pylyshyn, J. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Greenfield, S. (1995). *Journeys to the centres of the mind*. New York: Freeman.

Harnad, S. (1982). *Metaphor and mental duality*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Harnad, S. (1987a). *Categorial perception: a critical overview*. New York: Cambridge University Press.

Harnad, S. (1987b). *Category induction and representation*. New York: Cambridge University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Harnad, S. (1992). *Connecting object to symbol in modelling cognition*. Springer.

Harnad, S. (1994). Computation is just interpretable symbol manipulation: Cognition isn't. *Minds and Machines*, *4*(Special issue on ''What is Computation?''), 379–390.

Harnad, S. (1995). *The origins of words: A psychological hypothesis*. Muenster: Nodus Publishers.

Harnad, S. (1996). On the virtues of theft over honest toil: Grounding language and thought in sensorimotor categories. In *Proceedings of the Hang Seng Centre Conference on Language and Thought*.

Hinton, G. E. (1988). Representing part-whole hierarchies in connectionist networks. In *Proceedings Tenth Annual Conference of the Cognitive Science Society* (pp. 48–54). Montreal, Quebec.

Intrator, N. (1992). Feature extraction using an unsupervised neural network. *Neural Computation*, *4*, 98–107.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*, 57–83.

Karmiloff-Smith, A. (1992). *Beyond modularity*. MIT Press.

Locke, J. (1975). *Essay concerning human understanding*. Oxford: Clarendon Press.

Lucas, J. R. (1994). A view of one's own. *Philosophical Transactions of the Royal Society, London. A*, *349*, 147–152.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–450.

O'Reilly, R. C. (1996). *The LEABRA model of neural interactions and learning in the neocortex*. PhD thesis, Carnegie Mellon University, Pittsburgh.

Paivio, A. (1986). *Mental representation: A dual coding approach*. Oxford, New York.

Penrose, R. (1994). *Shadows of the mind*. Oxford University Press.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1-2), 77–105.

Quine, W.V.O. (1960). *Word and object*. MIT Press.

Rummelhart, D.E., Hinton, G.E., & Williams, R. (1986). *Learning internal representations by error propagation* (pp. 45–76). Cambridge, MA: MIT Press.

Sales, N. J. (1996). *Aspects of lexical development in artificial neural networks*. PhD thesis, Imperial College of Science, Technology and Medicine.

Searle, J. R. (1980). Minds, brains and programs. *Behavioural and Brain Sciences*, *3*, 417–424.

Searle, J. R. (1992). *The rediscovery of the mind*. The MIT Press.

Sloman, A. (1994). Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society, London A*, *349*, 43–58.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *The Southern Jnl of Philosophy*, *XXVI*(Supplement), 137–161.

Soloman, R. (1993). *Introducing philosophy*. Harcourt Brace College Publishers.

Van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.

APPENDIX A: A SUMMARY OF THE ARTIFICIAL CONSCIOUSNESS FRAMEWORK

A.1. The Fundamental Postulate

The personal sensations which lead to the consciousness of an organism are due to the firing patterns of some neurons, such neurons being part of a larger number which form the

state variables of a neural state machine, the firing patterns having been learned through a transfer of activity between sensory input neurons and the state neurons.

A.1.1. Corollary 1: The Brain is a State Machine

The brain of a conscious organism is a state machine whose state variables are the outputs of neurons. This implies that a definition of consciousness be developed in terms of the elements of automata theory.

A.1.2. Corollary 2: Inner Neuron Partitioning

The inner neurons of a conscious organism are partitioned into at least three sets:

Perceptual Inner Neurons: responsible for perception and perceptual memory;
Auxiliary Inner Neurons: responsible for inner ''labelling'' of perceptual events;
Autonomous Inner Neurons: responsible for ''life-support'' functions—not involved in consciousness.

A.1.3. Corollary 3: Conscious and Unconscious States

The contribution to consciousness of the inner neurons and the sensory neurons has three major modes:

Perceptual: which is active during perception—when sensory neurons are active;
Mental, conscious: which is an act of thinking in the same neurons even when sensory neurons are inactive or otherwise engaged;
Mental, unconscious: which is activity generated by the neurons involved in conscious activity, but which does not cause sensations of consciousness.

A.1.4. Corollary 4: Iconic Learning

To qualify for a contribution to consciousness, firing patterns in the inner perceptual/conscious neurons need to be created through dominant neural inputs which sample the activity of outer sensory neurons and influence the function of the inner neurons. This has been dubbed ''iconic learning''.

A.1.5. Corollary 5: Prediction

Relationships between world states are mirrored in the state structure of the conscious organism enabling the organism to predict events.

A.1.6. Corollary 6: The Awareness of Self

As a result of iconic learning and feedback between physical output and the senses, the internal state structure of a conscious organism carries a representation of its own output and the effect that such an output can have on world states.

A.1.7. Corollary 7: Will

The organism, in its mental mode, can enter state trajectories according to need, desire or in an arbitrary manner not related to need. This gives it its powers of acting in both a seemingly free and purposeful manner.

A.1.8. Corollary 8: Representation of Meaning

When sensory events occur simultaneously or in close time proximity in different sensory modalities, iconic learning and generalisation of the neural state machine ensures that one can be recalled from the other.

A.1.9. Corollary 9: Learning Utterances

The feedback loop responsible for the creation of ''self'' representations is also responsible for the creation of state representations of the basic utterances of the organism which are retrieved in response to the utterances of other, ''adult'', organisms and may be used by the adult to teach the organism more complete utterances such as the words of a language.

A.1.10. Corollary 10: Learning Language

Language is a result of the growth process of a social repository from which it can be learned by a conscious organism, given the availability of knowledgeable ''instructors''. The structure of such language is a process of the social evolution of a best-match to the development of state structure in the organism.

A.1.11. Corollary 11: Iconic Transfer of Qualitative Sensory Properties (Qualia?)

Iconic transfer operates on all aspects of sensory perception discriminated by receptors. Therefore qualitative properties of objects such as colour become candidates for iconic transfer and representation in recallable state structure.

A.1.12. Corollary 12: Instinct

To enhance survival, an organism needs a substrate of output actions that are linked to inputs or inner autonomic neurons and which are or become independent of the state of inner neurons. These may be inborn or become independent of state variables as a result of development.

A.1.13. Corollary 13: Emotions

Emotion in a neural state machine is an iconic encoding of instinctive sensations. It leads to behaviour which enhances the avoidance of danger and the protection of survival.

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1317–1331, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00050-6

1997 SPECIAL ISSUE

Learning, Action and Consciousness: A Hybrid Approach Toward Modelling Consciousness

RON SUN

Department of Computer Science, The University of Alabama

(*Received* 2 *July* 1996; *accepted* 4 *December* 1996)

Abstract—*This paper is an attempt at understanding the issue of consciousness through investigating its functional role, especially in learning, and through devising hybrid neural network models that (in a qualitative manner) approximate characteristics of human consciousness. In doing so, the paper examines explicit and implicit learning in a variety of psychological experiments and delineates the conscious/unconscious distinction in terms of the two types of learning and their respective products. The distinctions are captured in a two-level action-based model* CLARION. *Some fundamental theoretical issues are also clarified with the help of the model. Comparisons with existing models of consciousness are made to accentuate the present approach. © 1997 Elsevier Science Ltd.*

Keywords—Neural networks, Hybrid systems, Consciousness, Implicit learning, Reinforcement learning, Procedural knowledge, Rule extraction, Dual representation.

1. INTRODUCTION

Amidst the widespread enthusiasm of recent years concerning the scientific study of consciousness, there are a large number of models being proposed (including computational models, which in turn include neural network models), and various claims have been made about them. These models capture to various extents experimental findings and pretheoretical intuitions about consciousness (see, e.g. Taylor, 1994; Schacter, 1990; Jackendoff, 1987; Shallice, 1988; Baars, 1988; Dennett & Kinsbourne, 1992; Penrose, 1994). Unfortunately, however, some of these models (such as Dennett & Kinsbourne, 1992; Shallice, 1988; Jackendoff, 1987) are aimed at a very high and gross level of explanation (e.g. overall architectures) and thus unable to provide more detailed predictions and explanations. On the other hand, existing computational, especially neural network, models tend to rush directly into complex neural physiological thickets (Taylor, 1994; Edelman, 1989) and thus may lose sight of forests. In addition, most existing models do not deal adequately with one crucial aspect of human consciousness: learning. In contrast to these approaches, we intend to stay at an intermediate and functional level; investigating the detailed functional roles of consciousness and determining how various aspects of the conscious and the unconscious should figure into the architecture of the mind (in terms of learning as well as performance). In other words, we posit a middle level between phenomenology and physiology/neurobiology, which might be more apt at capturing fundamental characteristics of consciousness. We will also link computational models of consciousness to parts of (phenomenological) philosophy that are concerned with consciousness.

As we will focus mainly on the learning aspect in consciousness, let us briefly describe the learning settings that we examined (from Sun et al., 1995). These settings are more complex than simple categorisation/classification and though action-based, involve more than just simple stimulus–response pairing. Psychological experiments involving dynamic decision making or artificial grammars will be discussed. In dynamic decision making (Berry & Broadbent, 1988), subjects were required to control the levels of an output variable by manipulating levels of an input variable. In one instance, subjects were to manage a simulated sugar production factory and the goal was to reach and maintain a particular level of productivity by manipulating the size of the workforce. In another instance, subjects were to interact with a computer simulated ‘‘person’’ and to maintain the behaviour of the person at ‘‘very friendly’’ by manipulating his/her own behaviour. In artificial grammar

Acknowledgements: This work was supported in part by Office of Naval Research grant N00014-95-1-0440. Thanks to Ed Merril and Diana Gordon for their comments.

Requests for reprints should be sent to Ron Sun, Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487, USA; Tel.: (205) 348-1667; e-mail: rsun@cs.ua.edu.

learning (Reber, 1989), subjects were presented with a string of letters that were generated in accordance with a simple grammar. Although subjects were unaware of the underlying grammars, they were asked to judge the grammaticality of novel strings. In addition, two navigation tasks were used for both psychological experiments and computational modelling in our lab (see Sun et al., 1995, 1996a, b). One is maze running, in which from a starting location, a subject has to find a goal location, using only local sensory information. The other is navigation through minefields, in which a subject is required to go from a starting location to a goal location within a given short period of time by navigating through a densely packed field of mines (which will explode if the subject gets too close). I will later show the relevance of these learning tasks to the study of consciousness.

In the rest of this paper, I will first present a cognitive architecture CLARION[1] for accounting for the distinction of the conscious and the unconscious (in Section 2). I will then show how it accounts for a number of phenomena related to the conscious/unconscious distinction in learning (which may also be referred to as the explicit and implicit distinction) (in Section 3). A discussion of some fundamental theoretical issues will take place after that (Section 4). A comparison to existing models of consciousness such as Baars (1988), Schacter (1990), and Damasio (1994) will follow, which will show the commonalities shared by some of these models and CLARION and the unique features of the present model (in Section 5). Some concluding remarks (Section 6) will complete the paper.

2. A HYBRID NEURAL NETWORK MODEL

A computational model that can tackle the learning tasks mentioned above is needed. It needs to satisfy some basic requirements as follows. It must be able to learn from scratch on its own (as human subjects often do in the learning tasks outlined earlier; Berry & Broadbent, 1988; Reber, 1989; and also Sun et al., 1996a). The model also has to perform concurrent, on-line learning. That is, it has to learn continuously from on-going experience in the world; for, as indicated by Medin et al. (1987), Nosofsky et al. (1994) and others, human learning is often gradual, on-going and concurrent, which is true of all the aforementioned tasks. As suggested by Anderson (1983) and many others, there are clearly two types of knowledge involved in human learning — procedural and declarative: while one is generic and easily accessible, the other is embodied and specific. Moreover, different types of learning processes are involved in acquiring different types of knowledge (Anderson, 1983; Keil, 1989; Smolensky, 1988; Stanley et al., 1989). Humans are able to learn procedural knowledge through trial and error (without a priori knowledge) in the aforementioned tasks. On top of low-level procedural skills, declarative knowledge can be acquired also through on-going experience in the world (see Stanley et al., 1989). Furthermore, it is important for declarative knowledge to be learned through the meditation of low level skills (i.e. bottom-up learning; see Sun et al., 1996a).

Procedural knowledge (skills) can be captured by subsymbolic distributed representation, such as that provided by a backpropagation network. Because of the implicit nature of procedural skills, details of such skills are in general inaccessible to consciousness (Anderson, 1983; Reber, 1989). A distributed representation naturally captures this property with representational units that are capable of accomplishing tasks but are in general uninterpretable and subsymbolic (Sun, 1994, 1995). (A symbolic representation may be used, but then this would require an artificial assumption that these representations are not accessible, while other similar representations are accessible — such a distinction is arbitrary.)

Procedural knowledge can be learned in a couple of different ways. In the case where correct input/output mappings are provided, straight backpropagation can be used on a neural network. Otherwise, reinforcement learning can be used (Sutton, 1990; Watkins, 1989). This is preferred because there is often no uniquely correct action in the aforementioned tasks, although feedback is usually available. Using reinforcement learning in neural networks, we can measure the goodness of an action through a payoff-reinforcement signal. An adjustment can be made to weights to increase the chance of selecting the actions that receive positive reinforcement and to reduce the chance of selecting the actions that receive negative reinforcement.

This level can be modular; that is, a number of small networks can co-exist each of which is adapted to specific modalities, tasks, or groups of input stimuli. This coincides with the well known modularity claim (Fodor, 1983; Karmiloff-Smith, 1986; Cosmides & Tooby, 1994), in that much processing in the human mind is done by limited, encapsulated (to some extent), specialized processors that are highly efficient. It is also similar to the idea of Shallice (1988) that a multitude of ''action systems'' compete with each other. There also has been some work in neural network and machine learning communities in developing modular systems, which are equally relevant.

On the other hand, declarative knowledge can be captured by a symbolic or a ''localist'' representation (Clark & Karmiloff-Smith, 1993), in which each unit has a clear conceptual meaning or interpretation. This allows declarative knowledge to be highly accessible and inferences to be performed explicitly (Smolensky, 1988; Sun, 1994, 1995).

[1] It was originally developed for modelling human skill learning; see Sun et al. (1995).

Declarative knowledge can be learned in a variety of ways. In this work, because of the dynamic on-going nature of the learning tasks, we need to be able to dynamically acquire a representation and to modify the representation subsequently if necessary, in an efficient or even one-shot fashion.

The difference in representing procedural and declarative knowledge revealed by the above discussion leads naturally to a two-level architecture, in which one level is procedural and the other declarative. This structuring can be argued on both psychological and philosophical grounds. Anderson (1983) put forward the dichotomy of separate yet interconnected declarative and procedural knowledge bases to account for a variety of learning data. Smolensky (1988) suggested that the separation of conceptual-level and subconceptual-level processing. The conceptual level possesses three characteristics: (1) public access; (2) reliability; and (3) formality. It can thus be modelled by symbolic processing. In contrast, skills, intuition, and the like are not expressible in linguistic forms and do not conform to the three criteria prescribed. Hence, skills and intuition constitute a different type of capacity, reflecting the ''subsymbolic'' processing at the subconceptual level (see also Shiffrin & Schneider, 1977). In a similar vein, Dreyfus and Dreyfus (1987) contrasted analytical and intuitive thinking, from a phenomenological analysis of human cognitive skill learning in which the fluent, holistic and situation sensitive way of solving problems (intuition) as observed in master level performers is in sharp contrast with the slow, deliberate thinking that often occurs in the novices (analytical thinking). Models have been proposed to account for such two-tiered structures, which often posit the existence of at least two separate components, each of which responds to one side of a dichotomy (e.g. Posner & Snyder, 1975; Schacter, 1990; Murphy & Medin, 1985; Keil, 1989; Sun, 1992a, 1994; Sun & Bookman, 1994). The dual representation hypothesis put forth in Sun (1994) stated that:

> It is assumed in this work that cognitive processes are carried out in two distinct levels with qualitatively different processing mechanisms. Each level encodes a fairly complete set of knowledge for its processing, and the coverage of the two sets of knowledge encoded by the two levels overlaps substantially.

Based on the above considerations, we developed CLARION: *Connectionist Learning with Adaptive Rule Induction ON-line*. It consists of two main components: the top level encodes explicit declarative knowledge, and the bottom level encodes implicit procedural knowledge. In addition, there is an episodic memory, which stores recent experiences in the form of ''input, output, result'' (i.e. stimulus, response, and consequence) that are recently-filtered (episodic memory will not be used in this paper and therefore will not be further discussed here) (see Figure 1).

An overall pseudo-code algorithm that describes the operation of CLARION is as follows:

1. Observe the current state x (in a proper representation).
2. Compute in the bottom level the Q-values of x associated with each of the possible actions a_i's: $Q(x, a_2), ..., Q(x, a_n)$. Select one action or a few based on Q-values.
3. Find out all the possible actions $(b_1, b_2, ..., b_m)$ at the top level, based on the input x (sent up from the bottom level) and the rules in place.
4. Compare the values of the selected a_i's with those of the b_j's (sent down from the top level), and choose an appropriate action b.

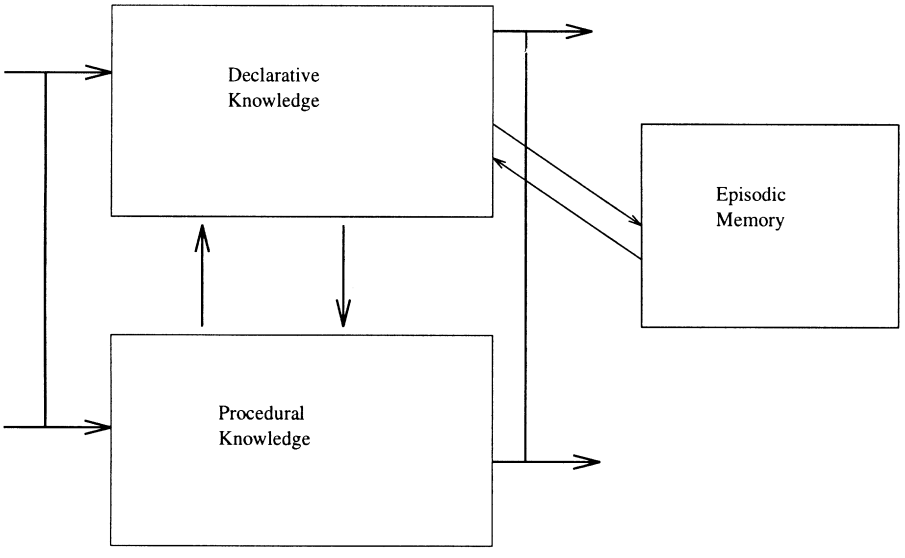


FIGURE 1. The CLARION architecture.

5. Perform the action b, and observe the next state y and (possibly) the reinforcement r.
6. Update Q-values at the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm.
7. Update the rule network at the top level using the *Rule-Extraction-Refinement* algorithm.
8. Go back to step 1.

In the bottom level, a Q-value is an evaluation of the ''quality'' of an action in a given state: $Q(x, a)$ indicates how desirable action a is in state x (which consists of some sensory input). We can choose an action based on Q-values, e.g. by choosing the one that has the maximum Q-value in the current state or by choosing an action probabilistically based on Q-values. To acquire the Q-values, one option is to use the *Q-learning* algorithm (Watkins, 1989), a reinforcement learning algorithm.[2] In the algorithm, $Q(x, a)$ estimates the maximum discounted cumulative reinforcement that the agent will receive from the current state x on:

$$\max\left(\sum_{i=0}^{\infty}\gamma^i r_i\right) \tag{1}$$

where γ is a discount factor that favours reinforcement received sooner relative to that received later, and r_i is the reinforcement received at step i (which may be 0). The updating of $Q(x, a)$ is based on minimising

$$r + \gamma e(y) - Q(x, a) \tag{2}$$

where γ is a discount factor and $e(y) = \max_a Q(y, a)$. Thus, the updating is based on the *temporal difference* in evaluating the current state and the action chosen. In the above formula, $Q(x, a)$ estimates, before action a is performed, the (discounted) cumulative reinforcement to be received if action a is performed, and $r + \gamma e(y)$ estimates, after action a is performed, the (discounted) cumulative reinforcement that the agent will receive; so their difference (the temporal difference in evaluating an action) enables the learning of Q-values that approximate the (discounted) cumulative reinforcement. Using Q-learning allows sequential behaviour to emerge. Through successive updates of the Q-function, the agent can learn to take into account future steps in longer and longer sequences.[3]

We chose to use a four-layered network for implementation (see Figure 2), in which the first three layers form a (either recurrent or feedforward) backpropagation network for computing Q-values and the fourth layer (with only one node) performs stochastic decision making. The network is internally subsymbolic and implicit in representation (in accordance with our previous considerations). The output of the third layer (i.e. the output layer of the backpropagation network) indicates the Q-value of each action (represented by an individual node), and the node in the fourth layer determined probabilistically the action to be performed based on the Boltzmann Distribution (i.e. Luce's choice axiom; Watkins, 1989):

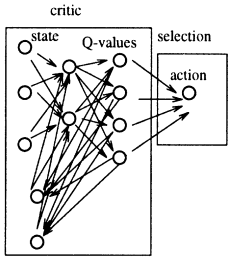


FIGURE 2. The *Q*-Learning method.

$$p(a|x) = \frac{e^{1/\alpha Q(x, a)}}{\sum_i e^{1/\alpha Q(x, a_i)}} \tag{3}$$

Here, α controls the degree of randomness (temperature) of the decision making process.[4] The combination of Q-learning and backpropagation facilitates the development of procedural skills in the bottom level, which can potentially be done solely on the basis of acting and exploring in the real world. This learning process performs both structural credit assignment and temporal credit assignment.

In the top level, declarative knowledge is captured in a simple propositional rule form. To facilitate correspondence with the bottom level and to encourage uniformity and integration (Clark & Karmiloff-Smith, 1993), we chose to use a localist network model for representing these rules. Basically, we connect the nodes representing conditions of a rule to the node representing the conclusion. However, we need to determine how we wire up a rule involving conjunctive conditions. There are a number of previous attempts (e.g. Sun, 1992b; Towel & Shavlik, 1993) that we can draw upon. For each rule, a set of links can be established, each of which connects to a concept in the condition of a rule to the conclusion of the rule. So the number of incoming links to the conclusion of a rule is equal to the number of conditions of the rule. If the concept in the condition is

[2] Supervised learning methods can also be applied, when correct mappings of an input and output are available.

[3] In terms of both simplicity and performance, Q-learning is best among similar reinforcement learning methods (Lin, 1992; Sun et al., 1995).

[4] The training of the backpropagation network is based on minimising the following:

$$err_i = \begin{cases} r + \gamma e(y) - Q(x, a) & \text{if } a_i = a \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where i is the index for an output node representing the action a_i. Backpropagation is then applied as usual to adjust the weights. Or, when correct mappings are available for each step, backpropagation can be directly applied.

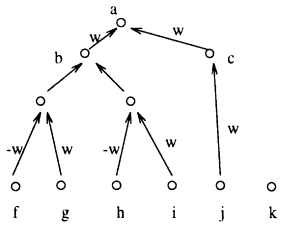


FIGURE 3. A network for representing rules: (1) $b\ c \rightarrow a$; (2) $\neg f\ g \rightarrow b$; (3) $\neg h\ i \rightarrow b$; (4) $j \rightarrow c$.

in a positive form, the link carries a positive weight w; otherwise, it carries a negative weight — w. Sigmoidal functions are used for node activation (as an obvious choice; other functions are also possible):

$$\frac{1}{1+\mathrm{e}^{\sum_i i_i w_i - \tau}} \tag{5}$$

The threshold τ of a node is set to be n times $w - \theta$, where n is the number of incoming links (the number of conditions leading to the conclusion represented by this node), and θ is a parameter, selected along with w to make sure that the node has activation above 0.9 when all of its conditions are satisfied, and has activation below 0.1 when some of its conditions are not met. (Activations above 0.9 are considered 1, and activations below 0.1 are considered 0; so rules are crisp/binary.) In addition, if there is more than one rule that leads to the same conclusion, an intermediate node is created for each such rule: all of the conditions of a rule are linked to the same intermediate node, and then all the intermediate nodes are linked to the node representing the conclusion (see Figure 3). (For more complex rule forms including predicate rules and variable binding, see Sun, 1992b).

To fully capture bottom-up learning processes, we devised an algorithm for learning declarative knowledge (rules) using information in the bottom level. The basic idea is as follows: if an action decided by the bottom level is successful (here, being successful could mean a number of different things, including the difference between the Q-value of the state before an action is performed and that after the action is performed, which comes from the bottom level; the details are specified in Sun et al., 1995), then the agent extracts a rule that corresponds to the action selected by the bottom level and adds the rule to the network. Then, in subsequent interactions with the world, the agent verifies the extracted rule by considering the outcome of applying the rule: if the outcome is not successful, then the rule should be made more specific and exclusive of the current case; if the outcome is successful, the agent may try to generalise the rule to make it more universal (Mitchell, 1982). (The detail of the algorithm can be found in Sun et al., 1995.)

At the top level, after rules have been learned, backward and forward chaining reasoning, means-ends analysis, counterfactual reasoning and explicit hypothesis testing can be used. These rules, expressed in the ''state → action result'' form (which constitutes a *schema*; cf. Waltz, 1991). allows powerful operations to be performed. Backward chaining means-ends analysis is accomplished at the top level through backward tracing of rule links from the ''result'', which is the new state entered after the ''action'' is performed, to the state, which is the state before the action is performed. This process is successively applied, with the derived ''state'' as the new state (the result) to be back traced, until reaching an initial state. All of the actions involved in the derivation are collected, which form a plan for accomplishing the desired final result. Counterfactual reasoning can also be applied (because we have information concerning conditions, actions, and results readily available): one can thus hypothetically alter either the conditions or the action of a rule, and see the change in (immediate or final) results. Such counterfactual reasoning can be used to justify (or explain) a chosen action, or a sequence of such actions (a plan). Explicit hypothesis testing, similar to techniques discussed by e.g. Bruner et al. (1956), Nosofsky et al. (1994), and Michalski (1983), can also be applied at this level of CLARION.

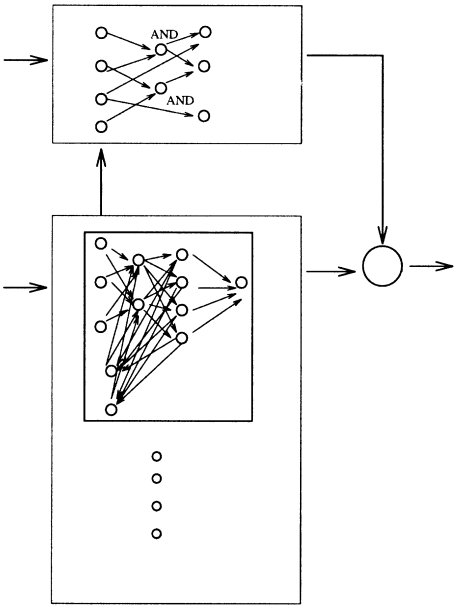


FIGURE 4. The implementation of CLARION.

The algorithm (at step 4) makes the final decision on which action to take by incorporating influences from both levels (as has been shown by Willingham et al., 1989, in humans, declarative knowledge can influence procedural performance). It allows different operational modes: (1) relying only on the top level; (2) relying only on the bottom level; or (3) combining the outcomes from both levels weighing them differently. The weights can change over time and in different situations. (These different operational modes roughly correspond to the

folk psychological notions of the intuitive mode, the deliberative mode, and the various mixtures of the two with different percentages of each; more later; see Sun et al., 1995.) Figure 4 shows the details of the two levels of the model.

The necessity of a two level architecture that incorporates two types of processes can be summed up as follows:

- In terms of representation, without the bottom level, the model will not be able to represent procedural skills properly. Such skills may involve graded, uncertain and inconsistent knowledge and autonomous stochastic exploration (with numeric calculation and probabilistic firing).
- In terms of learning, without learning in the bottom level, the model will not be able to learn from experience dynamically to acquire procedural skill. The bottom level captures *gradual* learning of skills, which is different from one-shot rule learning at the top level.
- Without the top level, the model will not be able to (1) represent generic, easily accessible, and crisp knowledge and (2) explicitly access and communicate that knowledge. When precision, consistency, and certainty are needed, declarative knowledge is preferred.
- Without rule learning, the model will not be able to acquire quickly and dynamically explicit knowledge for the top level from experience, and therefore have to resort to externally given declarative knowledge or to procedural knowledge exclusively.

There is ample biological evidence that indicates the existence of multiple pathways (in visual, linguistic, and other processing modes) some of which lead to conscious awareness, while others do not (e.g. one type is cortical while the other is subcortical), as described in Damasio (1994) and LeDoux (1992). For example, LeDoux (1992) described a cortical pathway from stimulus to thalamus to cortex, which produces conscious thoughts, and a subcortical pathway from stimulus to thalamus then to amygdala, which can lead directly to brain stem and effect actions without any explicit process. A two-level model such as CLARION approximates the separation of the two kinds of pathways to a certain degree, and suggests, in a concrete and tangible way, how subcortical processes can play a fundamental role in supporting and initiating consciousness in a bottom-up direction (more later).

3. PSYCHOLOGICAL DATA AND THE MODEL.

The crucial link between this model of procedural/declarative knowledge and the conscious/unconscious distinction in humans is in the psychological work on implicit learning (by e.g. Reber, 1989; Lewicki et al., 1992; Berry & Broadbent, 1988; Stanley et al., 1989; Willingham et al., 1989). Such work shows the dissociation between conscious and unconscious learning. Human knowledge, and its acquisition process, could be partially or completely unconscious. The connection from such illustrative data to our model lies in the ability of the model to account for some of the most important characteristics of human implicit/explicit learning, as will be sketched below. (More detailed comparisons and data can be found in Sun et al., 1996a, b)

3.1. The Difference of Conscious/Unconscious Learning

In the psychological data, there is a clear demonstration of the difference between conscious and unconscious (or, explicit and implicit) learning. Berry & Broadbent (1988) demonstrated this through an experiment using two similar dynamic decision tasks differing in the degree of saliency of the input/output relation. Human subjects were required to maintain the behaviour of a computer person at a ''very friendly'' level through their inputs. In the salient version, the computer responded in accordance with the subjects immediately preceding input. In the non-salient version, the computer responded in accordance with the input prior to that. Results suggested that subjects in the two conditions learned the tasks in very different ways: subjects in the non-salient condition learned the task implicitly while subjects in the salient condition learned the task explicitly, as demonstrated by tests of their explicit knowledge. Reber (1989) described a similar situation in artificial grammar learning. When complex hierarchical relations were needed to judge grammaticality, subjects tended to use implicit, unconscious learning; for example, when a sequence consisted of pairings of adjacent symbols that were ambiguous pair-wise but unambiguous when the entire sequence was viewed through hierarchical relations, such as in the case of 101110, implicit learning was preferred by the subjects. When only pair-wise relations were needed, such as in the case of 101010, subjects were more likely to use explicit, conscious learning by inducing an explicit rule. In other tasks, Cohen et al. (1990) also expressly demonstrated a dissociation between learning simple (pairwise) relations and learning complex hierarchical relations. A pattern emerging from the human data is that, if the to-be-learned relationships are simple, usually explicit/conscious learning prevails, while, when more complex relationships are involved, implicit/unconscious learning becomes more prominent. The implicit learning mechanism appears to be more structurally sophisticated and able to handle more difficult situations (Lewicki et al., 1992). It is important to note the inability of human subjects to articulate their implicitly learned knowledge, no matter how hard they tried (this is especially true in Lewicki et al., 1992). The subjects were often not even aware that they were learning. Nevertheless their performance improved over time, which demonstrated that their knowledge was unconscious.

This accords well with the CLARION model. In the model, one can freely move from one type of process to another, by engaging or disengaging the top level and its associated learning mechanisms (explicit processes that are consciously accessible, as discussed before), or the bottom level and its associated mechanisms (implicit processes that are not consciously accessible). Furthermore, in general, the bottom level is used to tackle more complex relations while the top level takes on simpler and crisper relations (cf. Reber, 1989; Seger, 1994). This is because the top level does not lend itself easily to the learning of complex structures due to its crisp, individuated, and symbolic representation and rigorous learning process. The bottom level, with its distributed network representation that incorporates gradedness and temporal information, handles complex relations better.

A specific example of this complexity difference is as follows. Implicit learning of sequences (e.g. artificial grammar sequences) is biased towards sequences with a high level of statistical structure with much correlation (Stadler, 1992). As has been demonstrated by Elman (1990) and by Cleeremans and McClelland (1991), recurrent backpropagation networks, as used in the bottom level of CLARION (in conjunction with *Q*-learning), can handle sequences with complex statistical structures, given proper training procedures. Dienes (1992) reported similar results, in which a simple network model outperformed other models in capturing sequence learning data. The rule learning mechanism, as used in the top level of CLARION, clearly has trouble handling such sequences. Therefore, in the circumstances in which a high level of statistical structure is involved in sequences, the bottom level prevails.

Note that there has been other work that demonstrated the distinction and dissociation of the two types of knowledge and proposed models based on that (e.g. Schacter, 1990; Shallice, 1988). However, some of the empirical work on which these models are based is concerned with abnormal subjects, most typically patients with brain damages. For example, Schacter (1990) discussed the following types of patients: amnesia (a selective inability to remember recent experience and to learn new information, typically due to lesions in the medial temporal lobe), blindsight (the inability to make certain responses in the absence of conscious perceptual awareness due to damages in the visual cortex), aphasia (impairment in processing syntactic or semantic information due to damages to a particular brain region), hemineglect (an impaired ability to attend to the side contralateral to the damaged hemisphere), and so on, all of which were characterised by dissociation of different types of information/knowledge somewhat similar to situations discussed above. Schacter (1990) proposed a model for accounting for the dissociation (see Section 5 for details). Although the model was, I believe, on the right track, the support for it was not as strong as it could have been, because in brain damaged patients, it was possible that certain reconfiguration and reallocation might have taken place (Shallice, 1988) and thus rendered the findings less applicable to normal human subjects. In this work, I only examine experimental findings from normal human subjects and thus results obtained may be generalised to a wider range of settings.

3.2. Delayed Explication of Unconscious Processes

In the implicit learning literature, implicit performance typically improves earlier than explicit knowledge that can be verbalised by the subject (Stanley et al., 1989). For example, in dynamic decision tasks, although performance quickly rises to a high level, subjects' verbal knowledge improves far slower; the subjects cannot provide usable verbal knowledge until near the end of their training (Stanley et al., 1989). Bowers et al. (1990) also showed delayed explication of implicit processes. When subjects were given patterns to complete, they showed implicit recognition of what a proper completion might be even though they did not have explicit recognition of a correct completion. The implicit recognition improved over time and eventually, an explicit recognition was achieved. In all of these cases, as suggested by Stanley et al. (1989) and Seger (1994), we may hypothesise that, due to the fact that explicit knowledge lags behind but improves along with implicit knowledge, explicit knowledge is in a way extracted from implicit knowledge. Cleeremans and McClelland (1991) also pointed out this possibility in discussing their data and models.

Several developmental theorists have considered a similar process in child development. Karmiloff-Smith (1986) suggested that developmental changes involve representational redescription. In young children, first low level implicit representations of stimuli were formed and used, then, when more knowledge was accumulated and stable behaviours developed, through a redescription process, more abstract representations were formed that transformed low-level representations and made them more explicit and usable. Based on data on perceptual analysis and categorization in infancy, Mandler (1992) proposed that relatively abstract ''image-schemas'' were extracted from perceptual stimuli, which coded several basic types of movements. On top of such image schemas, concepts were formed using information therein. She suggested that it was likely that an infant gradually formed ''theories'' of how his/her sensorimotor procedures work and thereby gradually made such processes explicit and accessible. Finally, Keil (1989) suggested that conceptual representations were composed of an associative component (with frequency and correlational information; Hasher & Zacks, 1979) and a ''theory'' component (with explicit knowledge; Murphy & Medin, 1985). Developmentally, there was a clear shift from associative to theory based representations in children. In data concerning learning

concepts of both natural and nominal kinds, simple similarity-based or prototype representations dominated at first, but gradually more explicit and focused theories developed and became more prominent. Keil (1989) pointed out that it was unlikely that theories developed independently, but rather they developed somehow from associative information that was already available. These findings further testify to the ubiquity of an implicit-to-explicit transition (Clark & Karmiloff-Smith, 1993).

CLARION readily captures this kind of bottom-up process. The bottom level develops implicit, embodied skills on its own (Section 2, eqn (2)), while the top level extracts explicit rules using algorithm *Rule-Extraction-Refinement* (Section 2). Thus, the delayed bottom-up learning naturally falls out of the model.[5]

3.3. Differences in Conscious/Unconscious Processes: Flexibility, Generalizability, and Robustness.

It has been shown that implicit learning produces less flexible knowledge than explicit knowledge (Seger, 1994; Berry & Broadbent, 1988; Stanley et al., 1989; Karmiloff-Smith, 1986). Seger (1994) argued that implicit learning results in knowledge that was more tied to the specific stimulus modality of the learning environment and less manipulable. Based on psycholinguistic data, Karmiloff-Smith (1986) observed that with the growth of explicit representations, more and more flexibility was shown by subject children. CLARION can account for the higher degree of flexibility of explicit, conscious knowledge relative to implicit, unconscious knowledge. Due to the explicit (i.e. localist) representation used at the top level of CLARION (which stores explicit knowledge), a variety of explicit manipulations can be performed that are not available to the bottom level. For example, backward and forward chaining reasoning, counterfactual reasoning, explicit hypothesis testing learning, and so on can be used individually or in combination. These capacities lead to heightened flexibility in the top level. The bottom level employs only backpropagation networks and thus cannot have the same flexibility.

As observed in many experiments, following explicit learning, subjects are able to handle novel stimuli in a similar way (or in other words, to generalise). In artificial grammar learning, Reber (1967, 1976) found good transfer to strings using different letters but based on the same grammar. Berry and Broadbent (1988) showed that subjects trained on a dynamic decision task could transfer to another task with a similar cover story and identical underlying relations. Generalisation has been demonstrated in neural network models by e.g. Elman (1990) and many others. Elman (1990) reported good generalisation of sequences by recurrent backpropagation networks in grammar learning. Pollack (1991) found generalisation of such networks to arbitrarily long sequences. As in human learning, generalization in neural networks is based in part on similarity of old and new sequences but also in part on certain structures exhibited by the sequences. Thus, the bottom level of CLARION, which incorporates a backpropagation network, has the capability to capture the generalization exhibited in human implicit learning. (Explicit processes, as in the top level of CLARION, can also generalise, albeit in a different way as discussed in Sun et al., 1995.)

It has also been observed that implicit processes are more robust than explicit processes (Reber, 1989) in the face of internal disorder and malfunctioning. For example, Hasher and Zacks (1979) found that encoding of frequency information (an implicit process) was correctly performed by clinically depressed patients, even though they could not perform explicit tasks consciously. Warrington and Weiskrantz (1982) found that amnesics were more successful in performing implicit rather than explicit memory tasks. This effect is consistent with the dual representation framework of CLARION: while the top level employs localist representation and is thus more vulnerable to malfunctioning, the bottom level utilises a distributed representation that is more resistant to damages and faults, as demonstrated amply in neural network models.

3.4. Unconscious Initiation of Action

Existing evidence indicates that unconscious processes often (if not always) initiate actions in skilled performance in advance of conscious awareness. Libet (1985) reported that electrophysiological ''readiness potentials'' (RPs) always precede conscious initiation of an act that is fully endogenous and voluntary. After a conscious intention to act appears, whether the action actually takes place or not can still be decided consciously by a subject within a time period of somewhere between 100 and 200 ms. As suggested by Libet (1985), the role of the conscious mind is not to initiate a specific course of action, but to control and influence (implicitly selected and initiated) actions.

This view is consistent with that of Willingham et al. (1989) that the role of explicit processes is to influence the implicit process but not to directly take control of skill learning or performance. Willingham et al. (1989) posited that the effects from the two processes are ''superimposed'' on each other, so that each type complement each other. Kelley and Jacoby (1993) also insisted that an important function of the explicit mind is to oppose, or counterbalance, the influence of the implicit mind.

The aforementioned view is also consistent with

[5] There is also evidence that explicit knowledge may develop independently. Willingham et al. (1989) reported such data. These data rule out the possibility that one type of knowledge is *always* preceded by the other type, at least under their experimental conditions. To account for this phenomenon, in CLARION, explicit hypothesis testing can be employed in the top level for learning rules, independently of the bottom level, as mentioned before.

voluminous data on the ever-present role of unconscious processes in all kinds of tasks: lexical priming, semantic processing, visual attention, unconscious perception and so on (as discussed in Velmans, 1991; Marcel, 1983). Velmans (1991) summarized evidence for the existence of implicit (preconscious) analysis of input stimuli, implicit processing of semantic content of word pairs in ''shadowing'' experiments, and implicit processing of bilingual messages in similar experimental settings. Most of these findings support the possibility that unconscious processes start before conscious processes take hold.

CLARION can readily accommodate this phenomenon, in that the bottom level, which captures unconscious processes, can work independently and initiate processing without the involvement of the top level. However, after the initiation of action and, consequently, the activation of the relevant nodes, the corresponding representations at the top level can then be activated by the bottom-up information flow (see Section 2). The activated explicit representations and their associated processes at the top level will in turn influence the implicit processing at the bottom level, in way of modifying and rectifying its outcomes and decisions (through the combination mechanism; see Section 2). Thus the implicit processes, which directly control actions in skilled performance, incorporate the results of explicit processes from the top level.

3.5. Synergy Between the Conscious/Unconscious Processes

Why are there two separate (although interacting) systems, one conscious and the other unconscious? Based on earlier discussions, we may hypothesize that each system serves a unique function and the two are complementary to each other; that is, there may be a synergy between the conscious and the unconscious. Such a synergy may show up by speeding up learning, improving learned performance, and facilitating transfer of learned skills.

In terms of speeding up learning, Stanley et al. (1989) reported that in a dynamic decision task (the sugar factory task), subjects' learning improved if they were asked to generate verbal instructions for other subjects along the way during learning. Willingham et al. (1989) found that those subjects who acquired full explicit knowledge appeared to learn faster.

In terms of learned performance, Willingham et al. (1989) found that subjects who verbalized while performing were able to attain a higher level of performance, because the requirement that they verbalised their knowledge prompted the formation and utilization of explicit knowledge. In high-level skill acquisition, Gick and Holyoak (1980) found that good problem solvers could better state rules that described their actions in problem solving. This phenomenon may be related to the self-explanation effect (Chi et al., 1989):

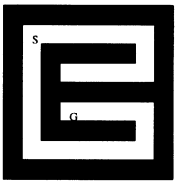


FIGURE 5. The Maze. The starting position is marked by ''S'' in which the agent faces upward to the upper wall. The goal is marked by ''G''.

subjects who explained the examples in textbooks more completely did better in solving new problems. In all these cases, it may well be the explication and the use of explicit knowledge that helped the performance.

In terms of facilitating transfer of skills, Willingham et al. (1989) obtained some suggestive evidence that explicit declarative knowledge facilitated transfer of skilled performance. It was reported that: (1) subjects who acquired explicit knowledge in a training tasks tended to have faster response times in a transfer task; (2) these subjects were also more likely to acquire explicit knowledge in the transfer tasks. In high-level domains, Ahlum-Heath and DiVesta (1986) also found that the subjects who were required to verbalize while solving the Tower of Hanoi problem performed better on a transfer task after training.

Sun et al. (1995) reported some simulation experiments that demonstrated CLARION was able to exhibit analogous synergy effects in learning, performance, and transfer through the interaction of the two levels. The simulation experiments were conducted in two domains: maze running and navigation through minefields. The details of the experiments and complete data can be found in Sun et al. (1995). Briefly, in the maze task, a subject/agent was to find a unknown target in the maze and had only rudimentary sensory inputs regarding its immediate left, front and right side, indicating whether there was a wall, an opening, or the goal; the agent could move forward, turn to the left, or turn to the right, until it found the target (see Figure 5). In terms of speeding up learning, the differences in learning speeds between *Q*-learning (which in CLARION captures unconscious learning at the bottom level) and CLARION (which includes both unconscious and conscious learning) were very significant. In terms of trained performance (measured by the average number of steps needed to reach the target in one episode), CLARION outperformed pure *Q*-learning by large margins again. We also compared the trained performance of the bottom level of CLARION alone (after the training of the entire system together, including *Q*-learning) with the performance of pure *Q*-learning, and discovered that the explication of skills not only improved the performance of the whole system, but it also improved the *Q*-learning part when included as part of CLARION. We also assessed the performance of trained models in a new and larger maze. CLARION transferred much better than *Q*-learning alone

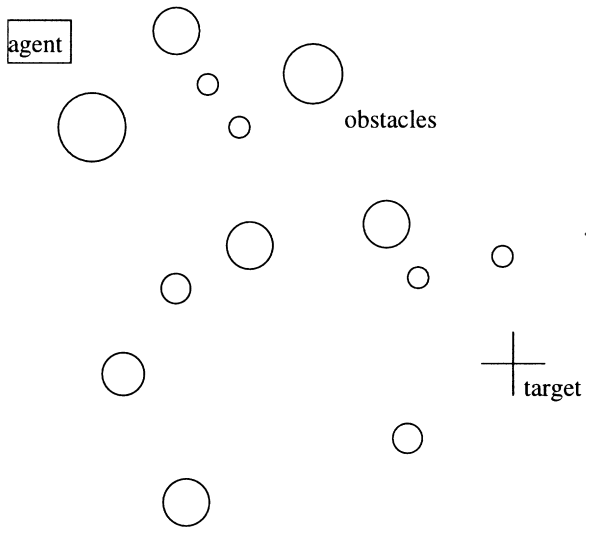


FIGURE 6. Navigating through mines.

(in terms of number of steps to reach the goal in one episode). Furthermore, by comparing the corresponding performance of the top level, the bottom level and the whole CLARION model, it was clear that often learned rules alone (the top level) performed better in transfer than the bottom level, as well as than the whole CLARION model, which showed that explicit knowledge facilitated transfer.

In the simulated navigation task shown in Figure 6, the subject/agent had to navigate an underwater vessel through a minefield to reach a target location. The agent received only local information from a number of instruments, as shown in Figure 7. Using only this information, the agent decided (1) how to turn and (2) how fast to move, and within an allotted time period, could either (a) reach a target (which is a success), (b) hit a mine (a failure), or (c) run out of fuel (a failure). In terms of learning speeds, the superiority of CLARION over *Q*-learning was statistically significant. To access transfer, after training models on 10-mine minefields, we assessed performance of these models in new minefields that contained 30 mines. CLARION outperformed *Q*-learning. The difference between the best transfer of Q-learning and the best transfer of CLARION was statistically significant. In sum, CLARION is able to replicate similar findings in human conscious/unconscious learning.

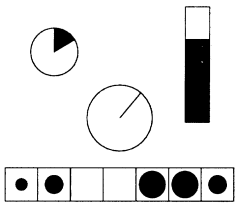


FIGURE 7. The navigation input. The display at the upper left corner is the fuel gauge; the vertical one it the upper right corner is the range gauge; the round one in the middle is the bearing gauge; the 7 sonar gauges are at the bottom.

4. THEORETICAL ISSUES

Some theoretical issues concerning consciousness will be discussed below in relation to the CLARION model.

4.1. Casual Efficacy of Consciousness

Is consciousness epiphenomenal as some have claimed? To see the casual role of consciousness, it is useful to examine the available cognitive. The following has been observed from patients who suffered the loss of some of the capacities of their consciousness due to brain damages (Marcel, 1988):

- They lost the ability to act on parts of the environment that were not accessible to their explicit/conscious mind (as in the case of blindsight patients who could not grasp objects in the blind field).
- They lost the ability to form an integrated self-concept (as in the case of amnesiac patients).
- They lost the ability to learn new complex tasks that required explicit, verbal instructions.
- They lost the ability to form explicit plans of actions before acting on them.

These effects indicate certain casual efficacy of consciousness, and are consistent with CLARION. In the model, these effects follow from the loss of the mechanisms in the top level of CLARION, such as backward chaining reasoning (planning), verbal instruction taking, and the use of explicit knowledge (regarding self and objects in the world).

Through contrasting ''aware'' vs ''unaware'' conditions on their experiments with human subjects, Kelley & Jacoby (1993) showed that conscious awareness per se had a distinct effect in subsequent behaviour. The two different conditions produced two different causal attributions in the subjects: one to true causes (in the ''aware'' condition) and the other to spurious causes (in the ''unaware'' condition; see Nisbett & Wilson, 1977 for causal attribution as error-prone, post hoc interpretation); consequently, different causal attributions led to different actions on the part of the subjects making the attributions. This kind of causal role in consciousness is consistent with that of the top level of CLARION, which shares the responsibility of controlling actions using explicit knowledge. A second mechanism on top of unconscious processes can offer counterbalance, and thus can have clear survival values to the agent possessing it.

4.2. Human Intuition

While the top level of CLARION captures conscious processes, the bottom level may capture *intuition* to some extent (as a form of skill): This level has the characteristics of being implicit, inaccessible, and holistic, which are also characteristics of human intuition (James, 1890; Dreyfus & Dreyfus, 1987). According to Smolensky

(1988), as mentioned before, intuition and skill are not expressible in linguistic forms and constitute a different kind of capacity, reflecting ''subsymbolic'' processing. Dreyfus & Dreyfus (1987) suggested that intuition is manifested in the fluent, holistic and situation sensitive way of dealing with the world, unique to humans and not captured by conventional symbolic computation. These identified characteristics can be found in the bottom level of CLARION to some extent.

It was hypothesised by Reber (1989) that human intuition may be the direct result of implicit, unconscious learning: Through the gradual process of implicit learning, ''tacit'' (implicit) representations emerge that capture environmental regularities and are used in direct coping with the world (without the involvement of any introspective process). Intuition is the end product of this process of unconscious and bottom-up learning (Reber, 1989). Bowers et al. (1990) also suggested that intuition is the outcome of an unconscious, implicit process (which later becomes explicit due to the emergence of a coherent pattern of activation) in the context of discovery. CLARION indeed uses implicit learning to develop tacit (implicit) representations in the bottom level and thus acquires intuition in the sense identified above.

4.3. Two Types of Consciousness

We can also examine the two levels of CLARION using the perspective of phenomenological philosophy. Heidegger (1927) emphasised a basic mode of existence, that is, the immediate comportment with the world. Normally, when going about its daily business, an agent is not *thematically* conscious of routine activities. Everyday routine activities are mostly made up of non-deliberate ''primordial'' coping. For example, in normal perception, we are usually not having thematic experience of the world (Dreyfus, 1992). An agent's ''openness onto the world'' is fundamental and makes possible the secondary experience of deliberate looking or trying to see (Dreyfus, 1992). Comportment is prior to any (explicit) belief, (explicit) knowledge, or (explicit) representation; it is a direct connection between an agent and its existential context. It is comportment with the world that is in fact a more fundamental kind of consciousness, according to Heidegger, and in this view, consciousness is non-representational (i.e. without explicit representation).

This is in contrast with the notion of ''thematic'' consciousness, which involves a focused, meditated awareness of the object of consciousness (akin to the common notion of consciousness). Thematic consciousness is representational because it treats awareness itself as an object (see also Clark & Karmiloff-Smith, 1993). As has been argued extensively by Heidegger (1927), thematic consciousness can indeed arise, but it presupposes a nonthematic, nondeliberate, direct, and on-going way of dealing with the world (i.e. comportment); for direct comportment with the world is a *necessary* means for coping with a complex world that exhibits complex regularities. Explicit representations are derived from direct comportment with the world. Derived representation and thematic consciousness come into play, e.g. during breakdown in which established routines get disrupted and thus alternative ways are necessary (see also Sun, 1994, Chapter 8). In the light of the above, the distinction between the two levels in CLARION can be corresponded roughly to the distinction between comportment and thematic consciousness. Specifically, the bottom level captures the implicit and routine activities and embodies ''comportment''. This is because the bottom level embodies skills resulting from and used for directly coping with the world and involves distributed representation, which is hence unable to present explicit traces of its processes (Smolensky, 1988). However, some events (e.g. certain implicit processes or explicit verbal inputs; Velmans, 1991) may lead to activation of corresponding explicit representations and processes at the top level and therefore lead to (thematic) conscious awareness. This is because at the top level explicit/localist representation is used, which makes it possible to articulate the content that is present and trace the processes as they are occurring (Sun, 1994, Chapter 2). Furthermore, due to the fact that the top level of CLARION is derived from, mediated by, and grounded in the bottom level, which has direct interactions with the external world and develops bottom-up from such interactions, (thematic) consciousness is clearly grounded, in this model as in humans, in the interaction between the agent and the external world.

4.4. Qualia

But what about qualia? Qualia refer to the phenomenal quality of conscious experience. Block (1994) distinguishes access consciousness and phenomenal consciousness, whereby access consciousness refers to the utilization of the content of consciousness while phenomenal consciousness refers to the subjective feel of conscious experience. Although it has been a major difficulty to understand phenomenal consciousness/qualia (''the hard problem''; Chalmers, 1992), some speculations may be made here in relation to CLARION: qualia (phenomenal consciousness) may be accounted for by the totality of a multi-modal (see the next section regarding modalities), multi-level organization and its total collective states, which are of extremely high complexity involving external perception (of many modalities), internal perception, action decision making, explicit concepts, etc. The complexity of this organisation may explain the difficulty (or impossibility) of describing phenomenal qualities (qualia) of consciousness, which is the most striking feature of phenomenal consciousness (as argued by Nagel, 1974). In this approach, a particular kind of phenomenal quality may be accounted for by a

particular region of total-state space (involving the totality of all the aforementioned aspects) or the manifold as termed by Van Gulick (1993), which gives rise to the sense of what something is like (Nagel, 1974). Clearly, such regions depend on particular *functional* organizations of modules and levels (Chalmers, 1992) that support such a space of total-states. Qualia are thus (partially) the result of functional organisations (architectures) of cognitive apparatuses. In CLARION, qualia are (partially) the result of the two-level organization, on top of all the detailed, intricate structures involved in various fine-grained modalities (the detail of which is not covered here). Equally important, such regions arise from the interaction of the agent and the world, as the consciousness is grounded in learning, action, and the world in which the agent exists. Thus, phenomenal consciousness is in general derived from a complex integrative organisation of cognitive apparatuses that develop as a result of the interaction between the agent and the world. This is true both in an ontological and in an ontogenetic sense. Note that, though access consciousness has a clear causal role in behaviour (as discussed earlier), phenomenal consciousness is less clear in this regard (Block, 1994).

4.5. Functionalism

An important question regarding consciousness is whether there is a physical basis for consciousness and what it is. It is pretty much agreed upon among contemporary philosophers that there is indeed a physical basis but they disagree on what constitutes that physical basis. In the CLARION model, we basically stipulate that the physical basis of consciousness is made up of the detailed architecture, the fine-grained functional organisation (e.g. the two-level modular framework with detailed sensory modality structures at a very fine level), of one's cognitive apparatus, in interaction with one's world. The distance between the physical and the phenomenological/psychological is so great that intermediate levels (functional levels) are necessary for studying cognition, and this is especially true in the case of studying consciousness. The present approach is a form of fine grained functionalism (Chalmers, 1992), which states that consciousness is invariant across systems with the same functional organization at a sufficiently fine-grained level (i.e. the principle of organizational invariance), as argued for elegantly by Chalmers (1992). It is a weak form of functionalism, in that it is not just causal connections between functional states that are important, but also the level at which we identify functional states. This approach is also interactionalism in the sense that the interaction of internal and external systems (the agent and its world), on the basis of internal (developing) fine-grained functional organizations, is crucial in giving rise to conscious experience (Heidegger, 1927).

5. COMPARISONS

The following comparisons will further explicate and accentuate the CLARION model in accounting for issues of consciousness. CLARION captures important features of existing models. The comparisons reveal that CLARION has the potential to account for the integration of sensory modalities, global consistency, and the unity of consciousness.

First we can compare CLARION with the model of Baars (1988) (see Figure 8), in which a large number of specialist processors perform unconscious processing and a global workspace coordinates their activities through global broadcasting to achieve consistency and thus conscious experience. The model bears some resemblance to CLARION, in that unconscious specialist processors in that model can be roughly equated to modules in the bottom level of CLARION, and the global workplace may be roughly captured by the top level, which ''synthesizes'' the bottom level modules and is essential in conscious processing. One difference is that CLARION does not emphasise as much internal consistency (Marcel, 1983): it is believed to be limited as a phenomenon in consciousness and may have only limited roles in the emergence of consciousness. Global broadcasting in Baars' model (Baars, 1988; Revonsuo, 1993) can be viewed as the integration of the two levels of representations (with the bottom-level representations dispersed within multiple modules) in CLARION, which does produce somewhat consistent outcomes (which lead to the unity of consciousness; Baars, 1988; Marcel, 1983; more on this later).

We can also compare CLARION with the model of Schacter (1990) (see Figure 9), which is based on neurophysiological findings of the dissociation of different types of knowledge in brain damaged patients as mentioned earlier. It is similar to CLARION, in that it includes a number of ''knowledge modules'' that perform specialized and unconscious processing (analogous to bottom-level modules in CLARION) and send their

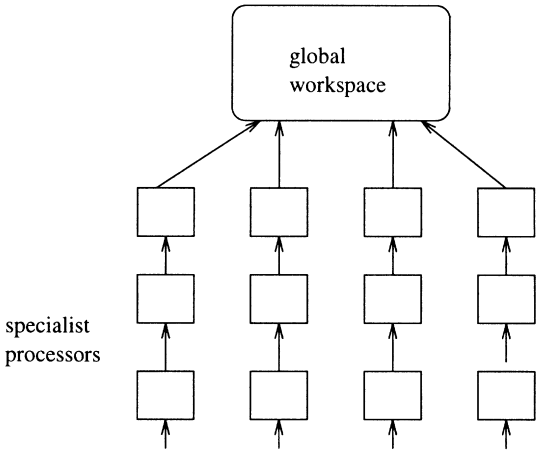


FIGURE 8. Baars' model of consciousness.

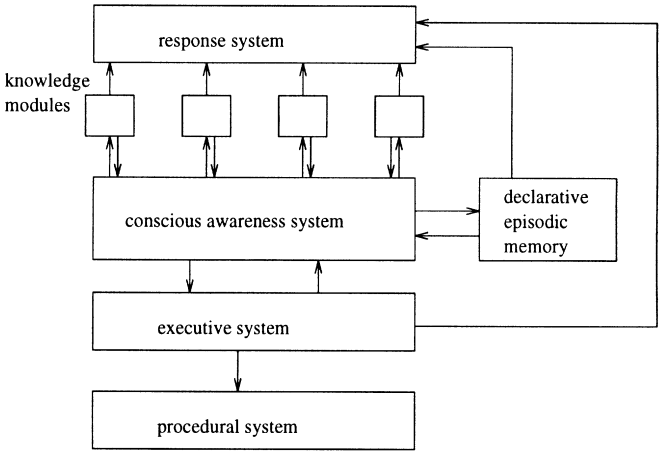


FIGURE 9. Schacter's model of consciousness.

outcomes to a ''conscious awareness system'' (analogous to the top level in CLARION), which gives rise to conscious awareness. Schacter's explanation of some disorders (e.g. the loss of short-term memory or explicit learning abilities, as mentioned earlier) is that certain brain damages result in the disconnection of some of the modules from the conscious awareness system, which leads to their inaccessibility to consciousness. An alternative explanation offered by CLARION is that disorders may not be due to disconnected modules, but the loss of some explicit learning and performance mechanisms at the top level (resulting from brain damages etc.).

Finally, we can examine Damasio's neuroanatomically motivated model (Damasio, 1994; Revonsuo, 1993). The model (see Figure 10) hypothesised the existence of many ''sensory convergence zones'' that integrated information from individual sensory modalities through forward and backward synaptic connections and the resulting reverberations of activations, without the need for a central location for information storage and comparisons; it also hypothesised the global ''multi-modal convergence zone'', which integrated information across modalities also through reverberation via recurrent connections. In CLARION, different sensory convergence zones may be roughly captured by bottom-level modules, each of which takes care of sensory inputs of one modality (at a properly fine level), and the role of the global multi-modal convergence zone (similar to the ''global workspace'' in a way) may be played by the top level of CLARION, which has the ultimate responsibility for integrating information (and serves as ''conscious awareness system''). The widely recognised role of reverberation (Damasio, 1994; Taylor, 1994) may be captured in CLARION through using recurrent connections within modules at the bottom level and through multiple top-down and bottom-up information flows across the two levels, which lead to the unity of consciousness that is the synthesis and integration of all the information present (Marcel, 1983; Baars, 1988).

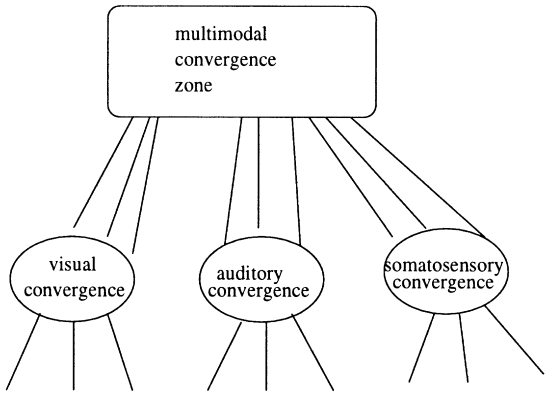


FIGURE 10. Damasio's model of consciousness.

6. CONCLUDING REMARKS

This paper presented a hybrid neural network model for learning that incorporated the distinction of declarative and procedural knowledge, and succeeded to some extent in accounting for the distinction of the conscious and the unconscious (or the explicit and the implicit). More specifically, the CLARION model applied neural network and machine learning techniques to explain complex human learning and consciousness in normal human subjects. It accounted for phenomena in psychological literature on learning and development in terms of the two levels in the model and their associated mechanisms. The model readily accommodated important features of existing models of consciousness. We also had something to say about theoretical issues such as qualia and intuition on the basis of the model, which helped clarify complex issues in a tangible way. The key issue for future research is scaling up the model by incorporating a variety of sensory information and dealing with complex environments, in order to allow rudimentary forms of consciousness to emerge, because, I believe, complexity is a necessary condition for consciousness.

REFERENCES

Ahlum-Heath, M., & DiVesta, F. (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition*, *14*, 281–285.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Berry, D., & Broadbent, D. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, *36A*, 209–231.

Block, N. (1994). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*.

Bowers, K., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, *22*, 72–110.

Bruner, J., Goodnow, J., & Austin, J. (1956). *A study of thinking*. New York: Wiley.

Chalmers, D. (1993). *Towards a theory of consciousness*. Ph.D Thesis, Indiana University.

Chi, M., Bassok, M., Lewis, M., Reimann, P., & Glaser, P. (1989). Self-explanation: how students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*, *8*(4), 487–519.

Cleeremans, A., & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cohen, A., Ivry, R., & Keele, S. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 17–30.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. *Cognition*, *50*, 41–77.

Damasio, A. (1994). *Decartes' error*. New York: Grosset/Putnam.

Dennett, D., & Kinsbourne, M. (1992). Time and the observer. *Behavioral and Brain Science*, *15*, 183–200.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, *16*, 41–79.

Dreyfus, H., & Dreyfus, S. (1987). *Mind over machine: the power of human in tuition*. New York, NY: The Free Press.

Dreyfus, H. (1992). *Being-in-the-world*. Cambridge, MA: MIT Press.

Edelman, G. (1989). *The remembered present: a biological theory of consciousness*. New York: Basic Book.

Elman, J. (1990). Finding structures in time. *Cognitive Science*, *14*, 179–211.

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

Jackendoff, R. (1987). *Consciousness and the computational mind*. MIT Press.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Hasher, J., & Zacks, J. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356–358.

Heidegger, M. (1927). *Being and time* (English translation, 1962). New York: Harper and Row.

James, W. (1890). *The principles of psychology*. New York: Dover.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition*, *23*, 95–147.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kelley, C., & Jacoby, L. (1993). The construction of subjective experience: memory attribution. In M. Davies & G. Humphries (Eds.), *Consciousness*. Oxford, UK: Blackwell.

LeDoux, J. (1992). Brain mechanisms of emotion and emotional learning. In *Current opinion in neurobiology*, **2**(2) (pp. 191–197).

Lewicki, P., Hill, T., & Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist*, *47*, 796–801.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*, 529–566.

Lin, L. (1992). Self-improving reactive agents based on reinforcement learning, planning, and teaching. *Machine Learning*, *8*, 293–321.

Mandler, J. (1992). How to build a baby. *Psychology Review*, *99*(4), 587–604.

Marcel, A. (1983). Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, *15*, 238–300.

Marcel, A. (1988). Phenomenal experience and functionalism. In A. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science*. Oxford, UK: Oxford University Press.

Medin, D., Wattenmaker, W., & Michalski, R. (1987). Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

Michalski, R. (1983). A theory and methodology of inductive learning. *Artificial Intelligence*, *20*, 111–161.

Mitchell, T.(1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Nisbett, R., & Wilson, T. (1977). Lelling more than we can know: verbal reports on mental processes. *Psychological Review*, *84*(3).

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *4*, 435–450.

Nosofsky, R., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Penrose, R. (1994). *Shadows of the mind*. Oxford, UK: Oxford University Press.

Posner, M., & Snyder, C. (1975). Facilitation and inhibition. In P. Rabbit & S. Dornick (Eds.), *Attention and performance*. Academic Press.

Pollack, J. (1991). The induction of dynamic recognizers. *Machine Learning*, *7*(2/3), 227–252.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, A. (1976). Implicit learning of synthetic languages: the role of instructional set. *Journal of Experimental Psychology. Human Learning and Memory*, *2*, 88–94.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental psychology: General*, *118*(3), 219–235.

Revonsuo, A. (1993). Cognitive models of consciousness. In M. Kamppinen (Ed.), *Consciousness, cognitive schemata and relativism* (pp. 27–130). Dordrecht: Kluwer.

Schacter, D. (1990). Toward a cognitive neuropsychology of awareness: implicit knowledge and anosagnosia. *Journal of Clinical and Experimental Neuropsychology*, *12*(1), 155–178.

Seger, C. (1994). Implicit learning, Psychological Bulletin, *155*(2), 163–196.

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

Shiffrin, R., & Schneider, W. (1977). Controlled and Automatic human information processing II. *Psychological Review*, *84*, 127–190.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–74.

Stadler, M. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 318–327.

Stanley, W., Mathews, R., Buss, R., & Kotler-Cope, S. (1989). Insight without awareness: on the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41A*(3), 553–577.

Sun, R. (1992a). A connectionist model for commonsense reasoning incorporating rules and similarities. *Knowledge Acquisition*, *4*, 293–321.

Sun, R. (1992b). On variable binding in connectionist networks. *Connection Science*, *4*(2), 93–124.

Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York, NY: John Wiley and Sons.

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*(2), 241–296.

Sun, R., & Bookman, L. (Eds.). (1994). *Computational architectures integrating neural and symbolic processes*. Norwell, MA: Kluwer Academic Publishers.

Sun, R., Peterson, T., & Merrill, E. (1995). *Hybrid architecture and situated learning* (Technical Report TR-CS-96-0019). University of Alabama. submitted to a journal.

Sun, R., Peterson, T., & Merrill, E. (1996a). *Bottom-up skill learning* (Technical Report TR-CS-96-0021). University of Alabama. submitted to a journal.

Sun, R., Peterson, T., & Merrill, E. (1996b). Bottom-up skill learning in reactive sequential decision tasks. *Proc. of 18th Cognitive Science Society Conference*. Hillsdale, NJ: Lauwrence Erlbaum Associates.

Sutton, R. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proc. of 7th International Conference on Machine learning*. San Mateo, CA: Morgan Kaufmann.

Taylor, J. (1994). Goal, drives and consciousness. *Neural Networks*, *7*(6/7), 1181–1190.

Towel, G., & Shavlik, J. (1993). Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning. 7*.

Van Gulick, R. (1993). Understanding the phenominal mind. In M. Davies & G. Humphries (Eds.), *Consciousness*. Oxford, UK: Blackwell.

Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences*, *14*, 651–726.

Waltz, D. (1991). How to build a robot. In S. Wilson (Ed.), *Proc. on Conf. on Simulation of Adaptive Behaviors*. Cambridge, MA: MIT Press.

Warrington, E., & Weiskrantz, L. (1982). Amnesia: a disconection syndrome?. *Neuropsychologica*, *20*, 233–248.

Watkins, C. (1989). *Learning with delayed rewards.* Ph.D Thesis. Cambridge, UK: Cambridge University.

Willingham, D., Nissen, M., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060

Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1333–1343, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00057-9

1997 SPECIAL ISSUE

Consciousness in a Self-Learning, Memory-Controlled, Compound Machine

ROBERT ALAN BROWN

8 Foster Street, Mattapoisett, Massachusetts

(*Received* 6 *July* 1996; *accepted* 15 *April* 1997)

Abstract—*A memory-controlled, sensor/actuator machine senses conditions in its environment at given moments, and attempts to produce an action based upon its memory. However, a sensor/actuator machine will stop producing new behavior if its environment is removed. A sensor/sensor unit can be added to the sensor/actuator machine, forming a compound machine. The sensor/sensor unit produces a stream of internally created sensed conditions, which can replace the sensed conditions from the environment. This illusion of an environment is similar to consciousness. In addition, actuator/sensor and actuator/actuator units can be added to this compound machine to further enhance its ability to function without an environment. Predetermined and empirical memory cells can be distributed throughout the control units of this compound machine to provide instinctive and learned behavior. The internal and exterior behavior of this compound machine can be modified greatly by changing the cycle start and ramp signals that activate these different kinds of memory cells. These signals are similar in form to brain waves. © 1997 Elsevier Science Ltd.*

Keywords—Memory-controlled machines, Conditional memory cells, Logarithmic changes in memory, Trinary encoders and decoders, Interrogation signals, Brain waves, Anesthesia, Coma.

1. MEMORY-CONTROLLED, SENSOR/ ACTUATOR MACHINE

The diagram in Figure 1 can represent a typical machine. A memory-controlled, sensor/actuator machine senses conditions in its environment at given moments, and converts these sensed conditions into memory addresses. It then attempts to produce an action based upon the content of these memory addresses, completing a transition cycle (Ashby, 1960).

The behavior of a memory-controlled, sensor/actuator machine can be predetermined (programmed or set beforehand), or it can be established over time by an empirical (self-learning) process. The output of the controller of a memory-controlled machine also can be absolute (determined by individual memory cells), or conditional (determined by the relative values of many memory cells).

1.1. Absolute-Predetermined, Memory-Controlled Machine

An absolute-predetermined, memory-controlled machine can be programmed to produce a specific action for each specific sensed condition. This is done by placing the toggle switch in each of its absolute-predetermined (AP) memory cells in a ‘‘1’’ (conducting) or ‘‘0’’ (non-conducting) position, as shown in the scalar matrix in Figure 2.

At the beginning of a transition cycle, the value of the sensor variable (A) is identified and held by a latched connect. This energizes the input disconnect, which prohibits any other value of the sensor variable (A) from being identified in that transition cycle. The voltage from the output enable appears at the output terminal of the AP cell with a closed toggle switch and the energized latched connect. This energizes an output connect belonging to the actuator variable (Z). The actuator delay timer (t_d) in the cycle timing circuit delays the movement of the actuator to the position of the energized output connect, and the cycle end timer (t_e) terminates the transition cycle, releasing the latches and timers. A brake holds the actuator in a fixed position between transition cycles.

1.2. Conditional-Predetermined, Memory-Controlled Machine

Complex behavior may be difficult to program into an absolute memory-controlled machine. A cumulative programming process can be achieved by changing the stored memory values logarithmically between zero and

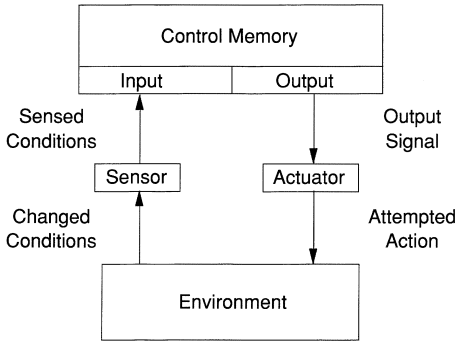


FIGURE 1. A memory-controlled, sensor/actuator machine produces an ongoing line of behavior when each sensed condition produces a new action, and each action creates a new sensed condition through the environment in each transition cycle.

1 in steps. This is done by means of the logarithmic subtraction mechanism (LSM) (Brown, 1994) in each conditional-predetermined (CP) memory cell in the controller of the conditional-predetermined, memory-controlled machine shown in Figure 3.

The CP memory cells are interrogated by the ramp signal (R). The ramp signal (R) causes a CP cell to become conductive when the ramp signal reaches the sensitivity value of that CP cell. The output from the matrix is terminated by the output disconnect when an output latch is energized. This prohibits any other less sensitive CP cells from selecting another value of the actuator variable (*Z*).

1.3. Logarithmic Subtraction Mechanism

The logarithmic response curves in Figure 4 show how the sensitivity of a single conditional-predetermined (CP) memory cell approaches its maximum or minimum value asymptotically.

An equal input of 1's and 0's leaves the sensitivity of a CP cell near the mid-position after a sufficient number of programming steps.

2. SELF-LEARNING, SENSOR/ACTUATOR MACHINE

A self-learning, memory-controlled machine measures the action that is actually carried out in each transition cycle. It then adjusts its memory to increase the likelihood of selecting the actual action when the same conditions reoccur (Brown, 1994), as shown in Figure 5.

The behavior of a self-learning machine will change until it is no longer interfered with by outside influences. Thus, the behavior of a self-learning, memory-controlled machine will come to reflect the intentions of its operators, and the constraints of its environment.

2.1. Absolute-Empirical, Memory-Controlled Machine

A circuit can be added to the absolute-predetermined controller shown in Figure 2 that measures the actual action that takes place in a given transition cycle. This actuator feedback circuit can be used to change the state of absolute-empirical (AE) memory cells, forming the absolute-empirical memory controlled machine shown in Figure 6.

The memory state of each absolute-empirical (AE) memory cell is set by changing its value to 0 when it selects an output, and changing its state to 1 if the output its selects actually takes place in a given transition cycle.

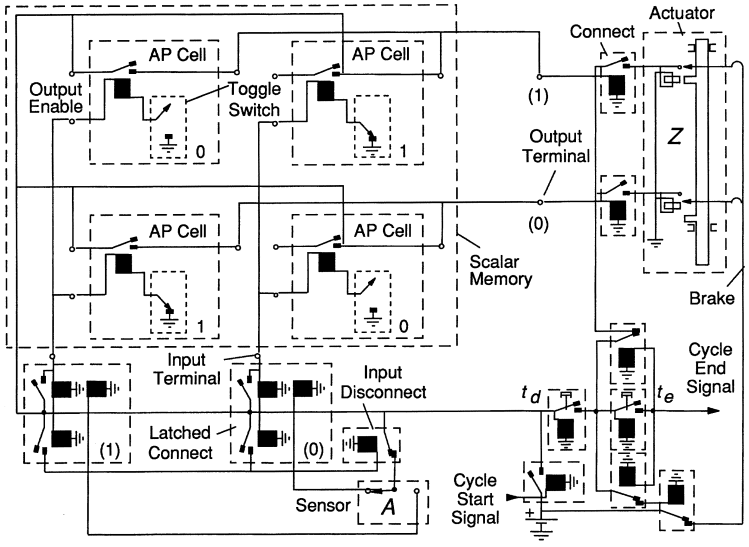


FIGURE 2. The binary memory matrix of an absolute-predetermined, memory-controlled machine is similar to a computer memory matrix.

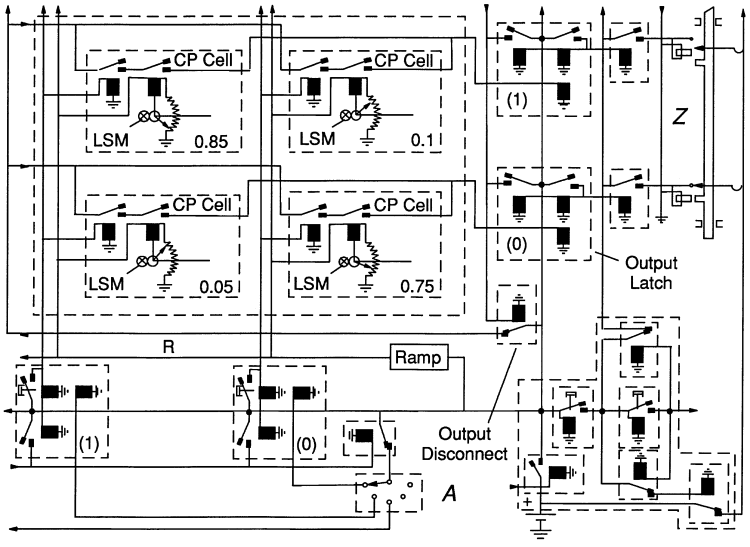


FIGURE 3. The analog sensitivity of each conditional-predetermined (CP) memory cell in the controller of a conditional-predetermined memory-controlled machine is determined by the position of the wiper arm on its variable resistor.

2.2. Absolute-Empirical Algorithm

The operation of the absolute-empirical, memory-controlled machine in Figure 6 is shown in the absolute-empirical algorithm in Figure 7.

An absolute-empirical, memory-controlled machine loses all traces of its previous experience when it encounters new learning.

2.3. Conditional-Empirical, Memory-Controlled Machine

The actuator feedback circuit shown in the absolute-empirical (AE) cell memory-controlled machine in Figure 6 can be added to the conditional-predetermined memory-controlled machine shown in Figure 3, forming the monolithic conditional-empirical memory-controlled machine shown in Figure 8.

The sensitivity of the selecting CE memory cell is fractionally decreased. Then, the actuator feedback circuit measures the actual action that takes place in a given transition cycle. It then fractionally increases the sensitivity of a feedback-selected CE cell, which is at the intersection of the actual value of the actuator co-variable (Z') and the value of the input variable (A) in that transition cycle. This makes the feedback-selected CE cell more likely to select the value of the actuator variable (Z) that takes place with the value of the sensor variable (A) in that transition cycle, as explained in the next section.

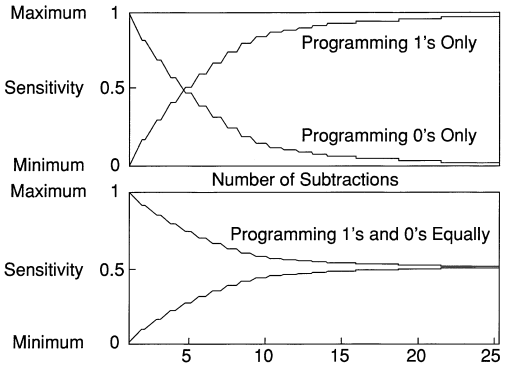


FIGURE 4. A falling logarithmic response curve is produced when the LSM subtracts a fixed percentage of the sensitivity each time the conditional-predetermined (CP) memory cell is programmed with a "0". A rising logarithmic response curve is produced when the LSM adds the same fixed percentage of the difference between the sensitivity and the maximum sensitivity each time the conditional-predetermined (CP) memory cell is programmed with a "1".

2.4. Logarithmic Response Curves of a CE Cell

The logarithmic subtraction mechanism (LSM) in each

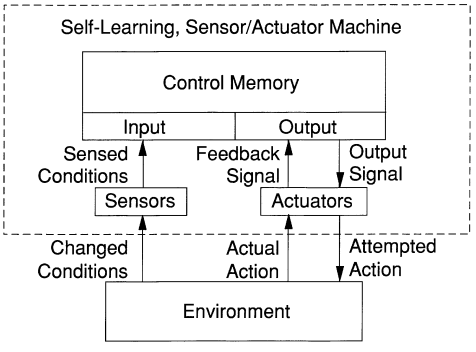


FIGURE 5. A self-learning, sensor/actuator machine learns to produce the action that is most likely to be carried out with each sensed condition in a given transition cycle.

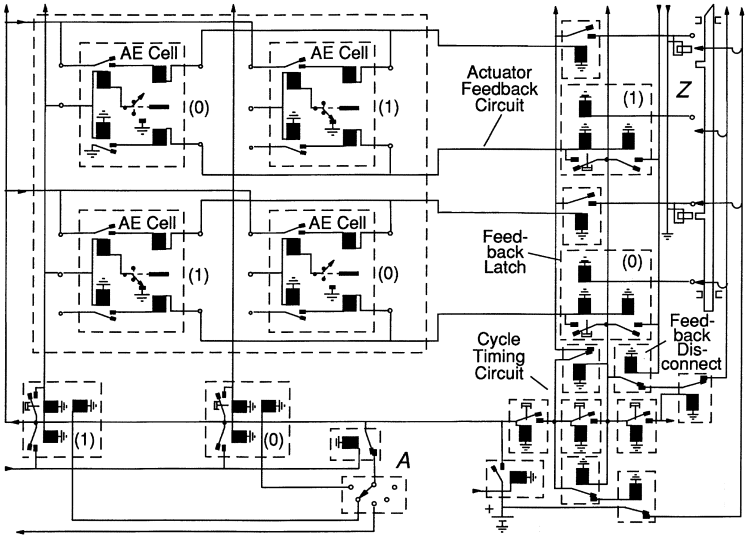


FIGURE 6. An absolute-empirical, memory-controlled machine learns the last action that is actually carried out with each sensed condition.

conditional-empirical (CE) memory cell decreases its sensitivity by a fixed percentage of its existing sensitivity value when it selects an action. It then increases its sensitivity by a fixed percentage of its difference from 1 if that action actually takes place in that transition cycle, creating the logarithmic response curves shown in Figure 9.

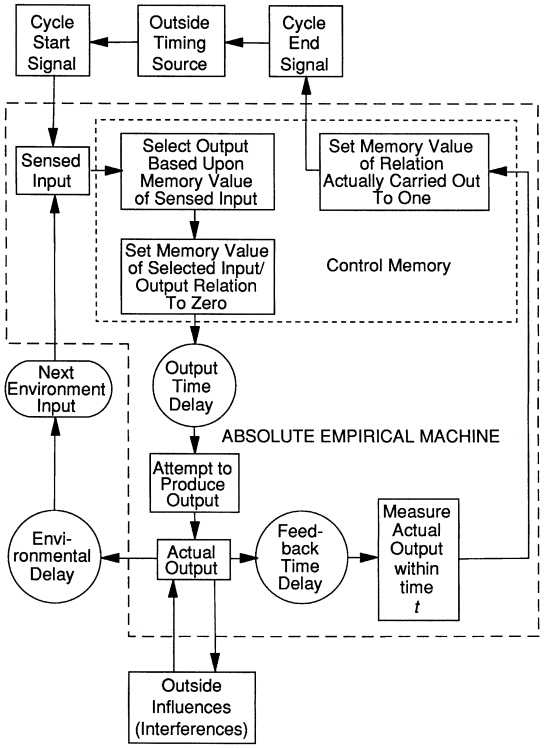


FIGURE 7. The absolute-empirical algorithm is typical of a binary computer memory, which leaves no trace of a recorded value once it is erased.

If more selection events occur than feedback events, the sensitivity of a CE cell will reach a value closer to the minimum sensitivity. If more feedback events occur than selection events, the sensitivity of the CE cell will reach a position closer to the maximum sensitivity. If an equal number of selection and feedback events occur for a sufficient number of times, indicating that the selected action always occurs, the sensitivity of the CE cell will approach the mid-position. Usually, there are more selection events than feedback events, because not every selected action is carried out successfully. Thus, a successful selection/feedback event increases the sensitivity of a CE cell toward the mid-position asymptotically, which is the maximum sensitivity a cell can have if every selected action is carried out.

2.5. The Conditional-Empirical Algorithm

The conditional-empirical algorithm shown in Figure 10 is similar to the accounting system used in financial markets.

The conditional-empirical algorithm causes the most sensitive memory CE cell connected to a sensed condition in a given transition cycle to select an action, but at the price of losing some of its sensitivity. If its action takes place, it regains more sensitivity than it lost. Thus, successful CE cells become more likely to select actions. However, the memory of an unsuccessful CE cell is reduced only slightly, leaving a permanent record of previous successes.

2.6. Duplex Control Memory

The control memory of a self-learning, memory-controlled machine works best if it is divided into two

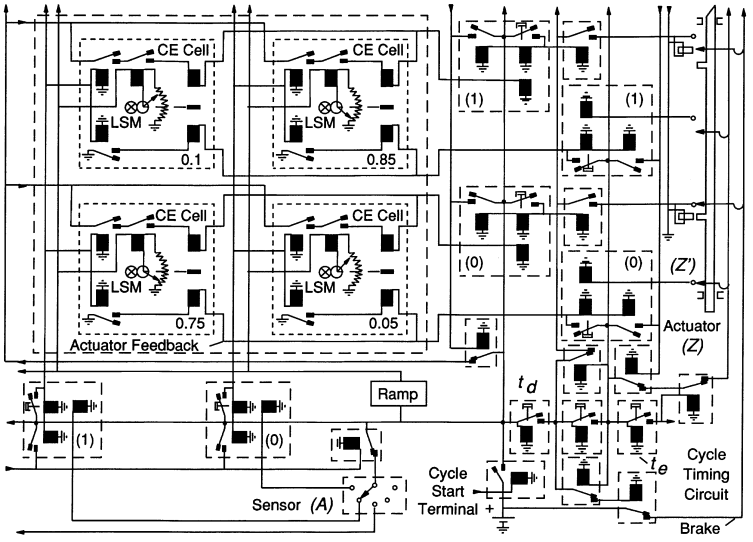


FIGURE 8. In a conditional-empirical, memory-controlled machine, a ramp signal causes the most sensitive CE cell connected to the existing value of the sensor variable (*A*) to select its value of the actuator variable (*Z*).

memories. An input decoder unit is connected to an output encoder/decoder unit by an intermediate variable (I) and intermediate co-variable (I'). The two units form the duplex control memory shown in Figure 11. This configuration allows a value of the intermediate variable (I) to represent a specific actuator state, and greatly reduces the number of required memory cells when a limited number of sensor/actuator relations are required.

The self-learning duplex sensor/actuator machine measures the actual values of its actuator variables, and uses this feedback signal to decode a back-selected value of its intermediate co-variable (I'). This back-selected value of the intermediate co-variable (I') is then used to increase the likelihood that the programmable input decoder will select the value of the intermediate variable (I) that is most likely to produce the actual values of the actuator variable that take place with the sensed conditions in that transition cycle.

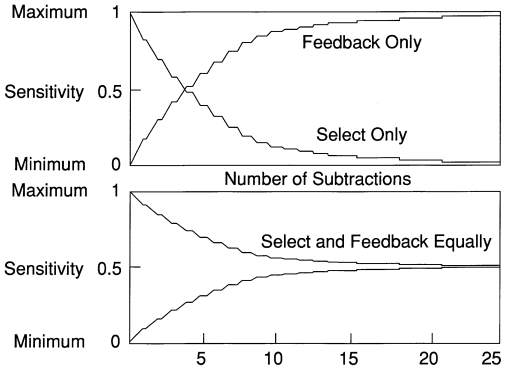


FIGURE 9. The logarithmic response curves show that the sensitivity stored in a conditional-empirical (CE) cell approaches the mid-position when there are an almost equal number of selection and feedback events.

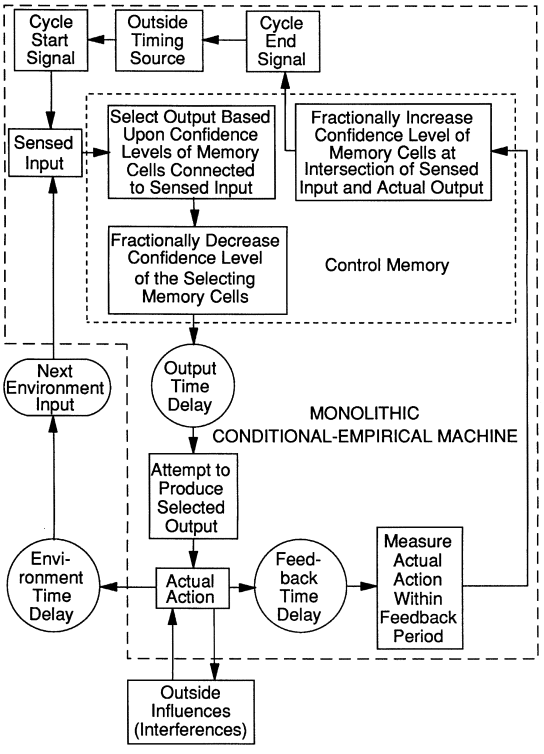


FIGURE 10. According to the conditional-empirical alogrithm, a buyer's liquid assets are reduced when a investment is made. If the investment is successful, the assets are increased more than they were decreased, when the investment is liquidated.

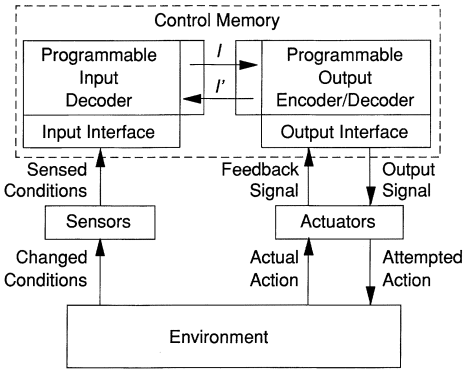


FIGURE 11. A duplex control memory must be programmed to decode the values of its sensor variables into a forward-selected value of its intermediate variable (*I*). Also, it must be programmed to encode this forward-selected value (*I*) into forward-selected values of its actuator variables. The duplex network is similar to a computer memory system. The sensor matrix acts as a programmable address decoder, and the actuator matrix acts as a set of data (memory) registers.

2.7. Conditional-Empirical, Bidirectional Actuator Unit

The CE cell matrix shown in Figure 8 can be used as a self-learning input decoder. However, it cannot be used as the output encoder/decoder in a self-learning duplex machine because it cannot feed back the information about the actual behavior of its actuators to its sensor unit. Thus, the conditional-empirical, bidirectional actuator unit shown in Figure 12 is required in a self-learning duplex network.

Each bidirectional (BD3) memory cell uses the same LSM and variable resistor (memory element) in the forward-selection and back-selection process. The forward-selecting ramp (F-R) is isolated from the back-selecting (B-R) ramp by the diodes shown. A bidirectional actuator matrix is not used as a stand-alone machine like the monolithic machine in Figure 8, rather as a unit in an empirical network.

2.8. Conditional-Empirical, Duplex Network

Two or more of the scalar conditional-empirical matrices in Figure 8 can be connected to form a multivariable sensor matrix. Two or more of the scalar bidirectional matrices in Figure 12 can be connected to form a multivariable actuator matrix. The multivariable sensor and actuator matrices can be connected to form the multivariable, conditional-empirical, duplex network shown in Figure 13.

The conditional input matrix forward-selects a value of the intermediate variable (I) where, in the set of CE cells connected to the latched terminal of each input variable (A and B) and the forward-selected value of the intermediate variable (I), the least sensitive CE cell in the set is more sensitive than the least sensitive CE

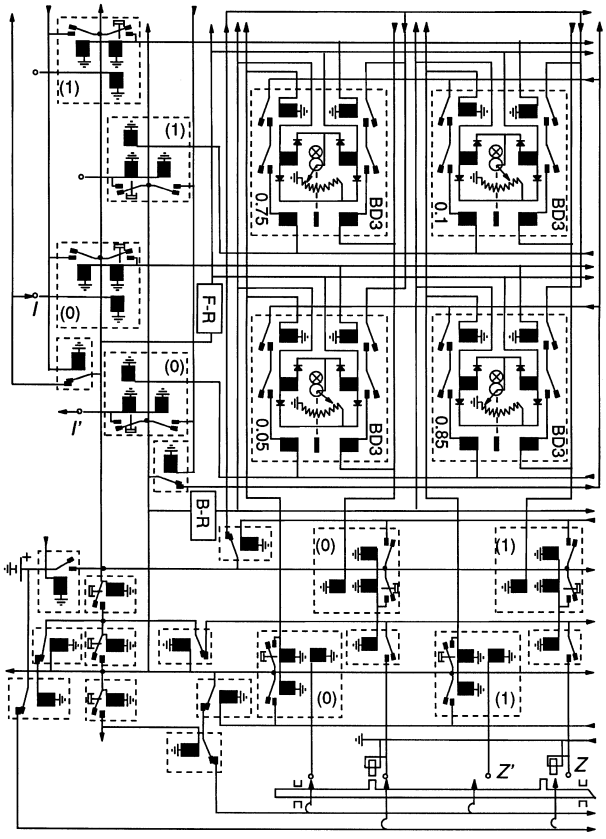


FIGURE 12. A conditional-empirical, bidirectional actuator unit forward-selects a value of its output variable (*Z*) based upon the value of its intermediate variable (*I*) and the sensitivity of its bidirectional (BD3) memory cells. It then back-selects a value of its intermediate co-variable (*I'*) based upon the value of its actuator co-variable (*Z'*) and the sensitivity of its bidirectional (BD3) memory cells.

cell in any other set connected to the latched input terminals and an unselected value of the intermediate variable (I). The bidirectional actuator matrix then forward-selects the value of the actuator variables (Y and Z) connected to the most sensitive bidirectional memory (BD3) cells connected to the forward-selected value of the intermediate variable (I). The sensitivities of the forward-selecting memory cells in both matrices are reduced according to the conditional-empirical algorithm (Figure 10).

The bidirectional actuator matrix then back-selects a value of the intermediate co-variable (I') where, in the set of BD3 cells connected to the actual latched terminal of each actuator co-variable (Y' and Z') and the back-selected value of the intermediate co-variable (I'), the least sensitive BD3 cell in the set is more sensitive than the least sensitive BD3 cell in any other set connected to the latched actuator co-terminals and an un-back-selected value of the intermediate co-variable (I'). The sensitivities of the back-selecting BD3 cells are increased logarithmically. Then the sensitivities of the conditional-empirical (CE) memory cells in the

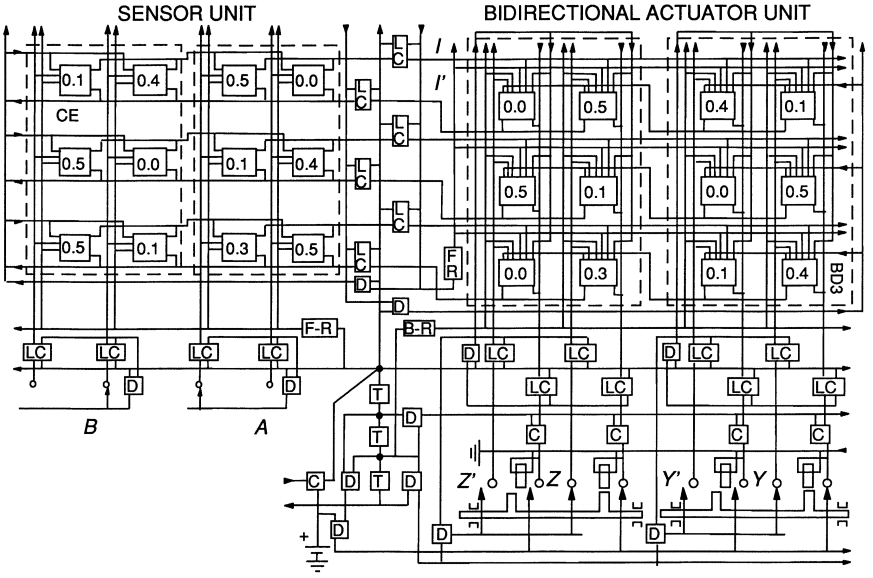


FIGURE 13. The conditional-empirical (CE) and bidirectional (BD3) cells in a multivariable, conditional-empirical, duplex network contribute to the forward-selection and back-selection of a value of the intermediate variable (*I*) that represents a sensor/actuator relation with the highest historical probability.

input matrix connected to the latched values of the sensor variables (A and B) and the back-selected value of the intermediate co-variable (I') are increased logarithmically according to the conditional-empirical algorithm (Figure 10).

2.9. Network Algorithm

The network algorithm shown in Figure 14 describes the operation of a duplex network of empirical memory matrices.

2.10. Digitized Self-Learning Machine

The empirical machine shown in Figure 13 has to learn behavior for each value of the sensed conditions encountered, even though the same output behavior may apply to many values of these sensed conditions. This problem can be overcome by connecting an encoder to the sensor variable and the sensor matrix through additional aggregate variables $A(a)$ and $A(b)$. A decoder and feedback encoder can be connected to additional aggregate actuator variables $Z(a)$ and $Z(b)$ and the bidirectional actuator matrix, as shown in Figure 15.

A generally high or a generally low value of the sensor variable (A) can select a unique value of the intermediate variable (I) through the most significant aggregate variable ($A(b)$) when high sensitivity values appear in both values of the least significant aggregate variable ($A(a)$). Alternatively, each specific value of the sensor variable (A) can each select a unique value of the intermediate variable (I) through each combination of values of the most significant and least significant aggregate variables ($A(b)$ and $A(a)$).

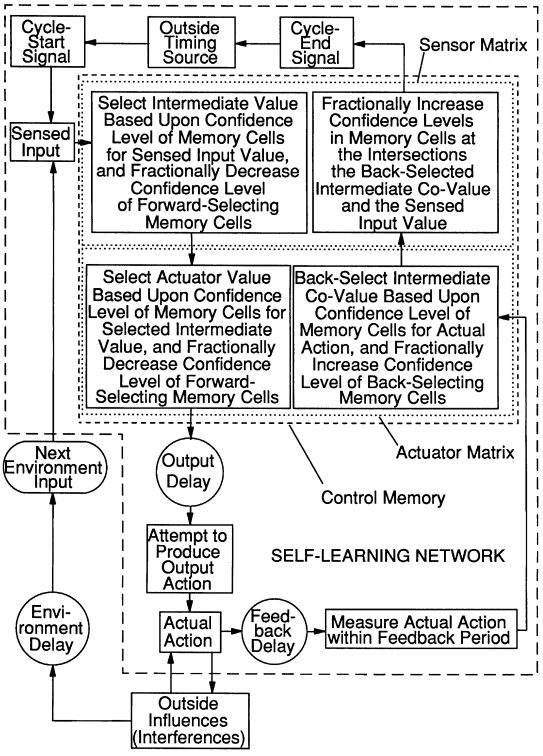


FIGURE 14. A duplex network uses the network algorithm to forward-select values of its actuators according to the forward-selected value of its intermediate variable. It then back-selects a value of the intermediate co-variable (*I'*) that represents the value of the intermediate variable (*I*) most likely to forward-select the actual values of the actuator variables that occurs in that transition cycle. The back-selected value of the intermediate co-variable (*I'*) is then used to increase the sensitivity of the conditional-empirical (CE) cells in the sensor matrix.

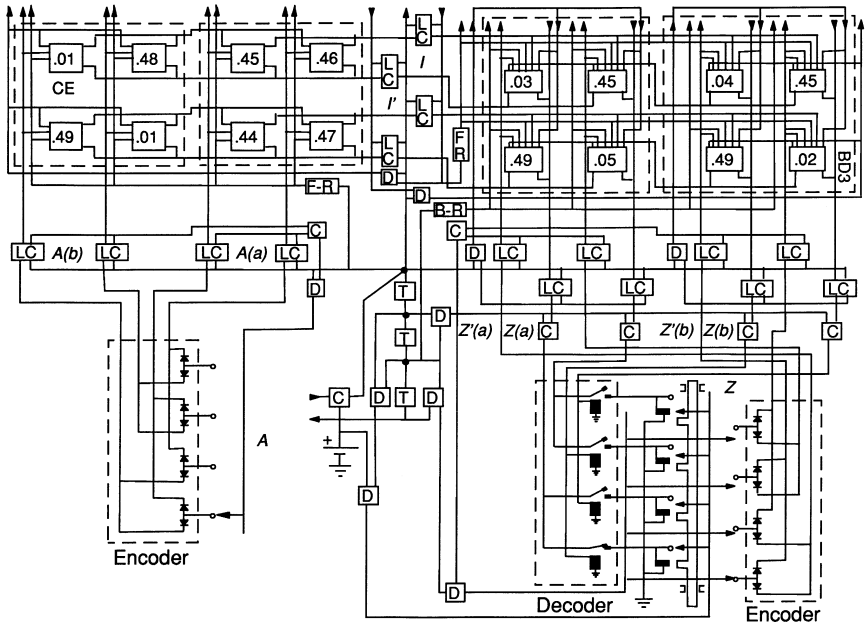


FIGURE 15. A digitized self-learning machine can learn to provide general or specific representations of its sensor and actuator variables (self-partitioning), and can do so with a minimum number of memory cells and a minimum learning time.

A binary sensor encoder, actuator decoder, and feedback encoder operating upon two aggregate variables is shown for the sake of simplicity. Additional aggregate variables must be added if the number of values of a sensor or actuator variable exceeds four. Trinary (three-value) encoders and decoders provide higher resolution with fewer memory cells than binary or quadrary encoders and decoders (Brown, 1994). The trinary digitized duplex network is the most effective configuration of the empirical machines presented.

3. COMPOUND MACHINE

If the environment of a sensor/actuator machine is removed, it will stop running. However, a sensor/sensor unit can supply new sensed conditions even if the real sensed conditions are suspended. The sensor/sensor unit can be connected to the sensor variables of a sensor/actuator machine, forming the compound machine (Brown, 1994) shown in Figure 16.

The four units of a compound machine make a unique contribution to its overall success. The sensor/sensor unit can provide the illusion of an environment by producing a stream of new sensations. The actuator/sensor unit may help the machine anticipate certain events by supplying new sensed conditions based upon the actual behavior of its actuators. The actuator/actuator unit can supply routine actuator behavior, which may contribute to the physical skills of the machine without the need to involve the sensor systems.

3.1. Cycle Control Signal

Since these memory-controlled machines operate in discrete steps at distinct moments, they may be called discontinuous machines, in contrast to continuous machines, such as analog computers. These discontinuous, memory-controlled machines require a cycle control

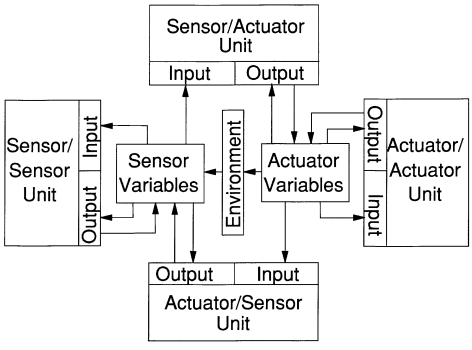


FIGURE 16. A compound machine can keep running even if its environment is removed.

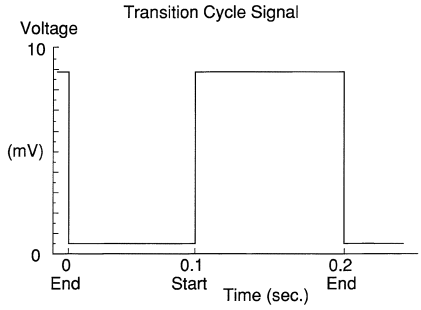


FIGURE 17. The cycle control signal determines the period between each input/output relation (transition).

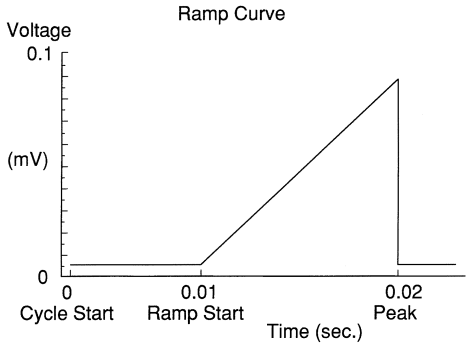


FIGURE 18. The voltage ramp signal is used to find the CP cell or CE cell with the highest sensitivity in the set of conditional cells connected to the sensed value at the start of a transition cycle.

signal to start and end each sense/action (transition) cycle, as shown in Figure 17.

Suspending the cycle control signal stops the behavior of a discontinuous machine.

3.2. Ramp Signal

Conditional-predetermined and conditional-empirical control memories require the additional ramp (search) signal shown in Figure 18 to find the action that occurs most likely with a sensed condition.

Loss of the ramp signal results in the suspension of all behavior produced by the conditional (CP and CE) memory cells, but does not influence the behavior of the absolute (AP and AE) memory cells shown in Figs 2, and 6.

3.3. Mixed Memory Matrix

Absolute-predetermined (AP) cells in Figure 2 and absolute-empirical (AE) cells in Figure 6 can be modified to operate with conditional-predetermined (CP) cells in Figure 3 and conditional-empirical (CE) cells in Figure 8, as shown in the mixed memory matrix in Figure 19.

The values of the mixed absolute-predetermined (MAP) cells and conditional-predetermined (CP) cells may be determined beforehand, and be programmed into mixed matrices to take care of self-regulation (homeostatic) functions of the machine. These predetermined memory cells may provide for innate behavioral mechanisms such as ''ducking a brick'' (Dennett, 1991). Behavior that must be learned can be established in conditional-empirical (CE) memory cells and in mixed absolute-empirical (MAE) memory cells in an input matrix, and bidirectional empirical (BD3) cells in the remaining matrices after the machine is put into use.

The behavior of the empirical cells can override and supersede the behavior of the predetermined cells if consistent behavior is found and learned. This is because the maximum sensitivity of the empirical cells can be designed to be greater than the sensitivity of the predetermined cells. The predetermined cells can also provide fallback behavior in cases where consistent behavior cannot be found and learned. All of these cells can be distributed throughout mixed memories of sensor/actuator, sensor/sensor, actuator/sensor, and actuator/actuator units of the compound machine shown in Figure 16.

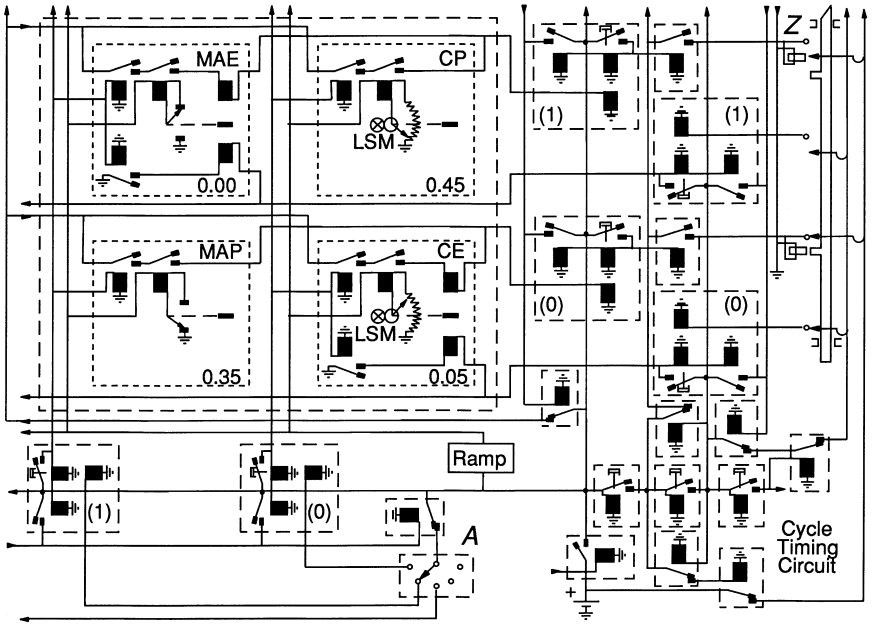


FIGURE 19. In a mixed memory matrix, the mixed absolute-predetermined (MAP) cells and conditional-predetermined (CP) cells can produce instinctive behavior, while the mixed absolute-empirical (MAE) cells and conditional-empirical (CE) cells can produce learned behavior.

4. CONSCIOUSNESS

The sensor/sensor unit of a compound machine may serve the function of consciousness. That is, it may provide an ongoing stream of internally-generated sensations that carries the machine through states that lack meaningful or useful external stimulation. This process may be extremely valuable in situations where interruptions occur in the normal stream of sensed conditions. For example, too much light may saturate the visual sensors of an autonomous guided vehicle (AGV) if it looks directly at the Sun. Alternatively, the visual sensors may be unable to see in an environment with too little light, say when it enters a dark room. In these cases, the sensor/sensor system of this AGV may be able to imagine the normal view that would be expected under the circumstances that preceded the loss of the image, and it could proceed as if it could see. Without the contribution of the sensor/sensor unit (consciousness unit) the AGV would have to be designed to stop when its image sensor fails. Otherwise, it may produce uncontrolled and/or inappropriate behavior.

Most of the memory cells in a sensor/sensor control memory of a compound machine may be conditional-empirical (CE and BD3) cells. This is because it is difficult or impossible to determine the sequence of sensor values that may occur in a given task beforehand. However, it is conceivable that predetermined memory cells may be used in a sensor/sensor controller to provide innate illusions (Dennett, 1991) that are useful to the operation of the machine.

4.1. Changing Levels of Consciousness

Changing the distribution of the ramp signals throughout a compound machine will result in major changes in its behavior. If the ramp signals are suspended to all the conditional memory cells in a compound machine with a mixed memory, the machine will go into what appears to be a coma. In this case, the memory cells involved in learning will not operate. Only the absolute-predetermined (AP) cells used for self-regulation will continue to operate, producing a vegetative state. If the ramp signal is discontinued to the conditional memory cells of its sensor/sensor controller only, the compound machine will appear to be in a hypnotic state. It will appear to react to all outside suggestions, having lost its internal sense of intentionality. Other physical and mental states are produced when the cycle start and ramp signals are discontinued to the other controllers in a compound machine. For example, a dreaming, sleep-like state can be produced if the ramp signal is discontinued to all of the controllers except the sensor/sensor controller. In this case, the machine can think, but cannot move. The trajectory of the sensor/sensor controller would correspond to a dream. If the ramp signals were suspended to all of the controllers except the actuator/actuator controller of a compound machine, the machine might go into sleep walking line-of-behavior involving routine movements, such as getting out of bed. If the cycle start signal were to be suspended to all of the controllers of a compound machine, a non-living machine would go into a state of suspended animation. Its memories would not change, and it would start right up where it left off when the cycle start signals are reintroduced. However, a living being would die if its cycle start signals were discontinued. This is because its self-regulation functions would cease.

4.2. Tests of Consciousness

Physicians test the degree of consciousness of patients by asking them if they know what day it is, or what city they are in. Since there are no obvious clues in the patients' room from which they can react responsively, they must dig into their conscious mind to find the answer to these types of questions. Thus, their answer provides a behavioral clue of consciousness (Herbert, 1993). Likewise, a compound machine can be placed in a task that involves sensor input. Then the sensors can be disconnected intermittently. If the compound machine is able to cope with the task successfully, a strong case is made for the machine having an internal representation (consciousness) of the external environment normally seen by its sensors. This internal representation is not a model, but a dynamic line-of-behavior that is similar to our intuitive understanding of our conscious thought process. We seem to live in (act through) our conscious thoughts.

4.3. A Cycle and Ramp Signal Mode Controller

Since there may be many memory units in a complex compound machine, there may be many overlapping cycle start signals and ramp signals. Each may operate in phase and out of phase with other signals. Some signals may be present at certain times, and be suspended at other times. The changing pattern of these signals greatly alters the type of behavior produced by a compound machine. These changes in behavior occur without changes in the basic memory state of the machine. This is similar to the changes in behavior that occur with changing patterns of brain waves in living beings. For example, patterns associated with sleep create different behaviors than patterns associated with wakefulness. Since these different types of behavior may be useful in different situations, a separate mode controller may be required. The mode controller can be used to control when and where the cycle start and ramp signals operate. Since each matrix is influenced by the

cycle timing circuit, the mode controller can control each cycle timing circuit, thus controlling the mode of behavior of a compound machine. For example, the mode controller can determine the level of physical action required for a given task, and set the sensor/actuator and actuator/actuator ramps to produce the appropriate level of arousal. Likewise, the mode controller can shut down the cycle start signals to an AGV used in a space voyage, thus immobilizing the AGV. It can then restart the cycle start signals when the AGV is ready to be deployed.

5. CONCLUSION

The local determinations of success rate by the logarithmic subtraction mechanism (LSM) in each conditional memory cell allow the digitized, self-learning, compound machine described in this paper to learn the behavior allowed by its operators and its environment. A useful enhancement of this compound machine is a sensor/sensor unit that produces internal signals similar to our conscious or internal sensing process. This Neural activity (Freemam, 1995) can occur without the machine sensing or acting in a real environment. Altering the cycle control and/or ramp signals of this compound machine can further enhance its behavior. This non-computational memory system may correspond to the living brain to a greater degree than conventional neural networks, which must sum currents, or fuzzy systems that depend upon global software calculations.

5.1. Questions Raised by the Design of Memory-Controlled, Compound Machines

- Does neural activity cause brain waves, as is commonly believed, or do brain waves cause the activity of neural memory cells in the same way that ramp signals cause conditional memory cells to operate, as described above? If neural activity causes brain waves, why are there no brain waves in the spinal cord, where we know there is much neural activity, but few if any memory cells?
- Do the cycle start signals and ramp signals, with their overlapping phase and different periods, correspond to brain waves? If so, the source of these signals in living beings may be found. If chemicals can be found that change the phase and/or relative strength of these signals, it may be possible to suspend and re-establish consciousness in a living being in a controlled manner. This may provide new ways of creating anesthesia and/or reviving individuals from coma.

REFERENCES

Ashby, W. Ross (1960). *Design for a brain*. New York: John Wiley and Sons.

Brown, Robert Alan (1994). *Machines that learn*. New York: Oxford University Press.

Dennett, Daniel C. (1991). *Consciousness explained*. New York: Little, Brown.

Freemam, Walter J. (1995). *Society of brains*. New Jersey: Lawrence Erlbaum Associates.

Herbert, Nick (1993). *Elemental mind: Human consciousness and the new physics*. New York: Dutton.