Explicit memory as a framework for the neural correlate of consciousness

Joël Kuiper, s1609920, me@joelkuiper.eu,
Dr. S. M. van Netten*, Dr. F. A. Keijzer†
28th July 2011

Abstract

Recently neurobiological understanding of the formation of (explicit) memory has increased dramatically. We are well underway in defining a neural correlate of memory. But the vessel for experiencing explicit memories when, for example, reliving a past moment in great detail is less understood. Consciousness remains elusive within contemporary science. Here the possibility of explaining consciousness in terms of memory is given. On the background of dynamic systems theory and causality it will be proposed that there is an inherent conjunction between memory and consciousness. This conjunction is supported by findings in neuronal synchronization and computational modeling of recurrent networks.

Keywords: consciousness; explicit memory; plasticity; synchronicity; dynamic systems; cognitive neuroscience

Introduction

Memories are complex. The term memory encapsulates everything from a vivid or even traumatic flashbulb memory to the subtle motor adjustments needed to successfully run a marathon. Memories allow us to keep track of the world around us and, similarly important, of who we are and who we want to be. Indeed, it can be said that the ability to remember is at the core of our identity, and diseases that act on the declarative memories systems have been adequatly described as identity destroying.

The past decades substantial progress has been made in understanding the mechanisms that underly the ability to form memories. Modern electrophysiological, molecular and genetical and imaging techniques have allowed the pinpointing of both neural and synaptic plasticity that is fundamental to information storage, learning and adaptive behavior (Feldman, 2009; Caporale & Dan, 2008). On an anatomical level functional magnetic resonance imaging (fMRI) has allowed to accurately describe the relation between behavior (psychology) and neuroanatomy (physiology) in both healthy and pathological cases (Simons & Spiers, 2003; Frankland & Bontempi, 2005). It seems that, while important open questions remain, understanding memory in terms of its neural substrate (a neural correlate of memory) is well within the realm of possibilities.

For example, the withdrawal reflex of *Aplysia californica* (a type of sea slug) has been extensively studied as a type of rudimentary memory. The slug shows habituation and sensitization by modulating the amount of neurotransmitters released, making the reflex weaker or stronger, respectively. The molecular mechanisms of these effects in terms of protein and gene activity are now properly understood.

And, classical examples such as the bilateral removal of the temporal lobes of patient H.M., resulting in almost pure anterograde amnesia, have shown that brain areas like the hippocampus play an important role in the formation of memories in humans.

Furthermore, it is now known that large amounts of memories are implicit, like the withdrawal reflex or our motor programs, they happen without being consciously aware of them. These implicit memories evade awareness and can only be experienced indirectly through interactions with the environment. In the case of motor skills, for example, we are well aware of the progress we can make in learning such a skill and become aware that we have reached a certain aptitude, but what exactly the motor programs that control the muscles learn is unknown to ourselves.

Implicit memories are in sharp contrast with the explicit (declarative) memories that can be vividly experienced. However, if these explicit memories are to be understood completely in terms of a neural correlate, then the way they are experienced must also be understood.

Unfortunately, the apparatus for experiencing memories, consciousness, remains elusive within contemporary science leaving the neural correlate of explicit memory incomplete. It remains hard to imagine that our consciousness really finds its seat in the neuronal connections and other physical processes inside our body. Thus, while a naturalistic worldview is likely to be universally accepted, the problem still tempts us to more metaphysical explanations.

Consciousness its elusive nature can be explained in part because, in the spirit of logical positivism (see for a comprehensive review Creath, 2011), the term tends to point to a metaphysical concept and is therefore impossible to grasp empirically, even making it meaningless in itself. But, this problem can be circumvented by introducing a well defined, strictly empirical, notion of consciousness. Defining such a term universally is problematic, but an important goal of science is precisely that: to establish a universal language of thought.

Several attempts at a proper definition have been given from various perspectives, but it seems that approaching consciousness head on from the perspective of cognitive neuroscience has several problems, which caused reluctance to make attempts.

Crick and Koch (1990, 2003), proposed that there might be a legacy issue carried over from behaviorism, the idea that we

*University of Groningen, Dept. of Artificial Intelligence
†University of Groningen, Dept. of Theoretical Philosophy

should simply not be concerned with such black-box issues. Or, that contemporary neuroscience lacks a useful vector in approaching the problem in terms of neural processes.

Indeed, explaining consciousness in terms of interacting proteins or fMRI-BOLD measurements in particular brain areas on its own does seem to be a widely, or publicly, accepted route.

But what is not accepted for consciousness seems to be for explicit memory, namely finding a neural correlate. It is accepted that explicit memories reside in our brain, and that the experience of remembering is a conscious one. This offers an interesting line of reasoning, since the neural correlate of explicit memory is one that needs to include the conscious experience of remembering to be comprehensive.

So instead of finding a Neural Correlate of Consciousness (NCC) head on it might be possible to explain or even define *in terms of the explicit memory process*. This is a two way approach, both a systematic search for the neural correlate of the experience of explicit memory is undertaken, as well as proposing that this neural correlate might be closely related to that of consciousness.

It seems reasonable that integral in this systematic search would be the advances in the field of computational neuroscience. The possibility to simulate hypothetical processes and perform otherwise technically challenging experiments *in silico* has undoubtedly expanded our understanding of the relation between mind and matter.

The aim of this paper is therefore to create a unifying view of explicit memory and consciousness. While no novel ideas or results will be introduced per se, it is the authors opinion that the proposed conjunction between the philosophical, neuroscientific and computational approach will provide new insights.

It will be proposed that both memory and consciousness rely on the same mechanisms, and that the NCC is closely related to the neural correlate of memory. Furthermore, a short overview will be given of memory, consciousness and synchronous activity. These ideas are often correlated (e.g. Edelman et al., 2011; Fries, 2009; Klimesch et al., 2010; Herrmann et al., 2004), but there is no consensus yet on their precise relation.

Firstly a definition and origin of memory will be given from a dynamical systems perspective. Herein, coincidence detection and causality will play an important role. Secondly an overview of neural network topology and spike activity will provide the necessary neuroscientific background to this perspective. Next, an overview of recent discoveries in Spike Time Dependent Plasticity (STDP) will be given. And, finally an integrative perspective will be proposed in which it will be argued that consciousness relies on the same mechanisms as memory.

Memories as dynamic and future predicting

To use explicit memory as a way to explain consciousness, and to explain the conscious experience of remembering, a few ideas must first be elaborated. This is necessary to formulate an appropriate and useful definition of explicit memory.

Briefly, instead of looking at memory as a storage device it must be considered as a continuous process, which will be supported by the concept of *dynamic systems*. Furthermore for memories to be evolutionary advantageous there must a certain order in things, such that remembering a past moment will be beneficial in the future, which will be elaborated as the concept of *causality*. These two concepts will then be used to propose some criteria for which a neural correlate compatible with one for consciousness can be found.

While the concepts are unlikely to be refuted, it must be noted that if they are, the henceforth proposed conjunction fails to produce an empirically useful notion of consciousness and memory.

Dynamic Systems

The acceptance that we are subjected to same laws of physics as everything else clearly set course away from the Cartesian mindset of substance duality. With this acceptance not only did psychology find its way into neuroanatomy, the conceptual link between our surroundings and our inner mental world could also be made.

It was realized that our surroundings can be used to remember things and to complement intelligence. One example is that of making notes, or more recently, the internet.

Key is that our mental representations arise from being situated in the very dynamic world around us. This connection between inside and outside, has made the groundwork for the use of dynamic systems theory within neuroscience.

The dynamic systems approach applied this way has given rise to the idea that instead of complicated abstract symbol manipulations, human and animal behavior can be better explained with the differential equations that arise from the laws of physics[1]. And following from that notion, that much of our cognitive capabilities are in a sense simpler than they seem, because we can exploit exactly those physical laws.

The hallmark of the dynamic systems approach herein has been the orientation behavior of the common house fly, *Musca domestica* (Poggio, Reichardt, & Hausen, 1981). Flies as part of their mating behavior tend to orient themselves toward moving objects, which is a seemingly complex system. But, detailed analysis of the flies neural network revealed a very simple system. A motion detector is directly coupled to the torque of the fly, thus when movement in its right visual field is detected, the fly steers to the right, to the left for the left visual field and zero torque is generated when an object is dead ahead. When the fly changes its torque so too does the position of the object in its visual field, therefore as a consequence generating a following behavior.

The consequence of this is that there can be intelligent behavior without symbolic representations. Looking at this from another perspective it can deduced that, like Van Gelder (1995)

[1] Physics here is used as a placeholder term for all the stable laws of the universe, and within this paper it is assumed that there is a possibility to reduce higher order effects to the (sub)atomic domain.

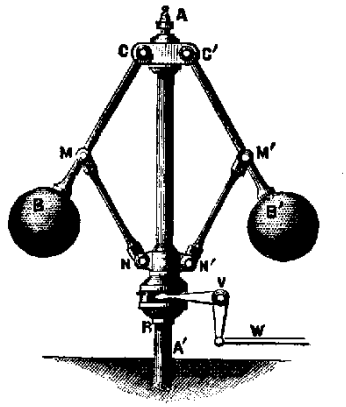

Figure 1. Governor used to regulate steam engines.

in his article *What might cognition be, if not computation?*, that perhaps all cognition is merely the result of the dynamics of physics. He used the example of the steam engine governor (figure 1), an ingenious device to regulate the speed of the engine.

To achieve the desired effect of a stable output speed a device could measure the speed of the rotation, the pressure and so forth and then, through some computational device, algorithmically calculate a desired modulation which should be actuated. But instead, the device exploits the centrifugal force. When the speed of the engine increases the rotation increases and thus, through centrifugation, pushes two weighted balls upward. Attached to the balls are two rods which pull, proportional to the height of the balls, a lever which regulates the engines' intake. The device is ingenious in its simplicity, merely through some basic laws of physics the speed of steam engines could be kept constant, which was crucial in sparking in what is now know as the Industrial Revolution (aprox. 18th and 19th century).

It could be that indeed human cognition, instead of taking in data, processing and then acting out the results of that computational process, our fluent interaction is better to be understood in terms of devices like the governor.

To conclude, it must be learned from the fly, the governor and dynamic systems theory in general, that our interactions with the surrounding world are important for our mental capacities. Behavior is to be understood as a continuum of interactions between the agent and its context, on both the immediate time scale, and on the scale of its evolution.

Underlying the way interaction between the agent and its context happen is the idea that its actions have a certain predictable outcome. Such predicability comes from the notion of causality, which will now be elaborated.

Causality

One of the great marvels of intelligence is that connections between a cause and an effect can be seemingly inferred. This idea that if a certain condition is met it must be that a certain outcome follows, is called causality. Or in different terms, there can be a relation between one event happening after another, in such a way that the first *caused* the latter.

Coming up with systematic way to finding and formalizing those causal relations has been the centerpiece of the Scientific Revolution. Crucial to the revolution has been the often cited father of the scientific method, Francis Bacon (1561-1626). He was one of the first to device the method of inductive reasoning. With inductive reasoning a general causal rule can be inferred form a set of particular observed causal relations.

A simple example is that of a game of pool, when one particular ball hits another we see that the second ball starts moving. If we continue observing this relation, that if a ball hits another the other ball starts moving, we can eventually inductively reason that for any ball that hits another the other will start moving. This relation, when formalized, became the law of conservation of linear momentum. This law has predictive nature, we can assume that all objects in motion follow the law of conservation of linear momentum. So the continued observation of a relation allows for us to predict future relations, which may be completely different.

But does causality actually exist? Does the conservation of linear momentum really constitute a property of the universe? David Hume (1711 - 1776) in his *An Enquiry concerning Human Understanding* argued that the idea of causality is merely a psychological construct, a human tendency. Because all we can observe is simply one fact after another. First we see one ball move and then another, but causality itself cannot be observed. It could be that, instead of the conservation of linear momentum, the balls were made of metal and a trickster was moving them under the table with a magnet.

Furthermore, there is no real necessity for any inductively produced rule to hold in the future. It might be that one moment the conservation of linear momentum does not hold, that the pool balls for example both come to a full stop. It only takes one counter-example to negate all the previous verifications of a rule, a problem which became known as the *Induction problem*. We cannot possibly observe *every* instance of a rule, observations can be flawed, and observations are even restricted by physiology. For example, some rules might be mediated by a special kind of radiation or particle which simply cannot be observed, but would explain some contradicting observations.

From this it seems that the idea of a *necessary connection* (causality) is a psychological phenomenon, simply arising from the observation of many things that seem to happen in conjunction[2] but is not a direct property of reality.

However, for an idea of causality to arise two seemingly disjunct events which are displaced in time must be correl-

[2]Hume also proposed the idea that no actual objects exist, but only the conjunction of certain properties. This was famously illustrated by his thought experiment to think of an object with no properties, which is impossible. While his original texts are still very readable and in public domain (Hume, 1748), Bailey and O'Brien (2006) offer a well written reading guide.

ated. To do this a certain memory of a previous event must be present, otherwise only conjunctions at the exact same perceived moment can be inferred. Memory is necessary for an inference about the future, and it can be argued that this predictive nature of causality is the only evolutionary reason for memory.

Redefining memory

Without the ideas of causality and dynamic system it is tempting to consider memory as a static storage device, commonly using the way computers handle information as a metaphor. This has resulted setting out a definition of memory as a multi-staged process, going from perception to short term memory to a fixed long term memory with various intermediates. This approach is however lacking in usefulness when trying to reconcile the conscious experience of remembering an explicit memory with the process of storing such a memory. Because, when various stages are considered it is tempting to somehow place consciousness as a separate stage, or as a *homunculus* observing the theater of thoughts and memories in the brain.

But, these views neglect that consciousness might be part, or even an emergent property, of the various processes it should conscious about. That is to say the conscious experience of remembering a vivid memory like a first kiss or high school graduation in which lively images, sounds and smells can be relived in great detail, does not somehow require a separate neural correlate, but is actually part of the very process of memory. To follow this line of reasoning one needs to view memory as a continuous process, analogous to a signal in physics.

The dynamic systems approach allows this, indeed putting the discrimination between storage and retrieval on the background. Instead of reasoning that an experience is stored (i.e. consolidated neural terms) and then reactivated by an executive when needed, it can be thought of as a continuum of interactions between the agent and the context.

But consider a world without causal relations, one without a necessary connection between past, present and future. In such a hypothetical world there would be no use for any sort of memory, since no consequences about future states can be made, and spending energy on remembering completely disjunct events would to seem to serve no evolutionary advantage.

So, it can be said that any memory in essence stores the conjunction or coincidences (i.e. correspondence in nature or in time of occurrence) of events, and if a certain coincidence is observed in abundance, a causal rule is inferred which allows us to predict future events. Remembering the past and predicting the future in that sense might be two sides of the same proverbial coin. This gives memory an evolutionary advantage, being able to infer something about the future based on the past is extremely beneficial for reproduction and survival.

So instead of looking at memory as a way to accurately depict previous events, it should be considered a continuously updating mechanism for mimicking the perceived world by storing its coincidences (see figure 2). Following from this notion some criteria must be given, providing a guide in finding a neural correlate.

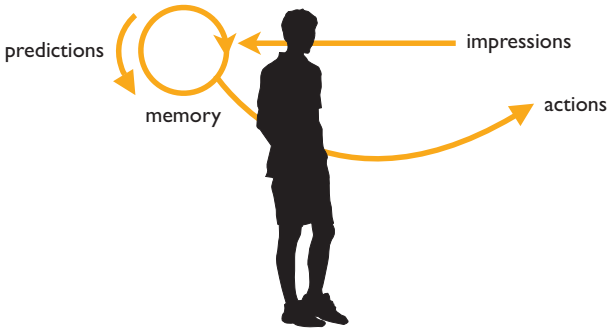

Figure 2. Instead of looking at memory as a static storage device, it should be considered a continuous interaction with the environment allowing to predict future states

Firstly memories are acquired (i.e. learned, they are about something), and are not innate to the brain in the sense that we are born with all our memories. Memory is a constructive process which comes forth from being situated in a world and thus those dynamics apply and one of those rules is causality.

Secondly the acquired memories are somehow placed within the neural networks of our brain, this neural correlate of memory is commonly referred to as a *trace* or *engram*.

Thirdly, memory traces can be reactivated given a certain context, whether it is an external stimulus or an internal narrative, the ability to recall is crucial. It does not mean, however, that explicit memories are recalled only within a certain context (as is the case with associative learning like operant conditioning), but it points to the concept of being "readily available".

This concept has strong relations with the computational idea of *content addressability*. In computer memory following the Von Neumann architecture, information is stored as a certain pattern and then given an address. When this address is called the stored pattern can be retrieved. This is in contrast with content addressable memory where the stored pattern can be retrieved by virtue of its own content, without a specific address, or even with fragments of the content which the stored pattern can then fill in or complete.

And finally memory is organized. Not only in temporal and spatial terms but memories also seem to be clustered into meaningful groups, as shown by the spreading of activation (Collins & Loftus, 1975). This meaningful clustering also happens on the more basal level of perceptual binding, input from multiple sensory modalities is either stored or retrieved as a single event. When a loud bang, for example, coincides often with colorful flashes this eventually might become the memory of fireworks. It are the dynamics of the world that give these coincidences their semantic content.

To illustrate these criteria lets consider the hypothetical example of waking up one morning, and still drowsy, you hit the table where you placed a glass the evening before. Surely the glass falls and breaks. This entire episode is presented

in a multitude of modalities, you see the falling the glass as well hearing the effects of it hitting the ground. To experience this event the input of each of the sensory modalities must be encoded into pattens neural activity and propagated through various neural systems. If it is then assumed that somehow those patterns of neural activity can be recalled at a later point of time, say the subsequent evening, then the event can be relived.

In essence the stored neural activity allows us to relive a previous event, and because of the nature of causality this can have a predictive effect. When we remember the falling of the glass it can be inferred that if its placed on the table it's likely to fall, so behavior can be adjusted accordingly.

Note that, since the world acts according to the laws of physics certain causal rules are inherently present. This is extremely useful, if for example the law of gravity would change within one lifetime it would be difficult to learn to walk. It should be noted that the causal rule of gravity, however, seems so ubiquitous that many of its consequences are encoded into our genes through evolution (e.g. the shape of our body), and do not have to be actively learned through interaction with the environment.

The neural correlates of these criteria will now be elaborated. By no means a complete overview will be given of all the subtleties involved in the formation of memory (see for general overviews Nalbantian, Matthews, & McClelland, 2010; Dayan & Abbott, 2001), however the insights will serve to propose a conjunction between the formation of explicit memories and the experience of consciousness. Furthermore, it will be argued that the mechanism for binding is in principle the same as that for inferring causal relations.

Neural correlates of memory

To find a neural correlate of the formation of memory and consciousness a bottom up, rather than a psychological top down, approach will be employed. To do this the basics of a neural correlate, namely neurons, will be the starting point.

The TLU-paradigm (see box 1) supposes that the existence of an action potential is often a group effort of multiple neurons. Depending on the thresholds and the influence of the connected neurons, it can take the action potentials of a multitude neurons within a millisecond time frame to generate a post-synaptic action potential.

The group effort gives rise to the possibility to consider neurons as a coincidence detecting apparatus, because in essence it requires the coinciding activity of multiple neurons to activate a subsequent one.

Spikes as a correlate for causality

Consider the example of the falling glass. To coherently perceive the unfolding scene the various visual features, sounds, smells and other sensory perceptions must be encoded into neural activity. The neural activity takes the shape of precisely timed action potentials, or spike pattens (also called spike trains, or spike time patterns). Various ways to encode information in spike patterns are possible, for example by changing the average rate at which neurons generate action potentials, called rate coding, or more subtly changing the time between subsequent spikes which is generally considered under the umbrella term of temporal coding (for an overview see Kumar, Rotter, & Aertsen, 2010).

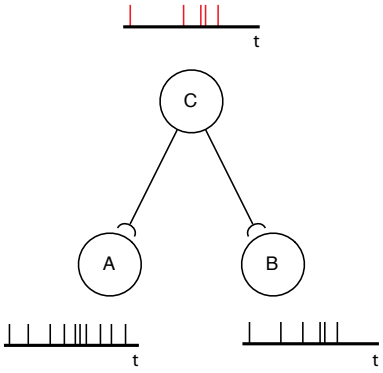

Figure 3. The schematic principle of binding through synchronicity. The neurons A and B provide input to neuron C, illustrated as the bars where each vertical bar corresponds to an action potential. When neurons A and B fire at the same time the threshold of neuron C is exceeded and fires, as depicted by the red vertical bars. The synchronous firing of C can thus be said to detect the coincidence between A and B.

To coherently experience a scene all this neural activity, often in different brain areas, must somehow be integrated into a single perception. This problem is called the *binding problem*. The binding problem refers to a wide array of psychological questions, for example how the different properties (such as shape, color and motion) of an object, represented in separate brain areas, give rise to the idea of a single object or how a location of an object (where, dorsal stream) is combined with the idea of an object (what, ventral stream).

One proposed solution to the binding problem has been binding through synchronicity (see figure 3). Herein the coincidence between the input of various neurons serves to bind activity into a single stream, and this can be seen as synchronous activity (see: Singer, 1999; Gray, 1999; Treisman, 1996).

Synchronous activity can be demonstrated with electroencephalography (EEG, see figure 4) by measuring the Local Field Potential (LFP) which is caused primarily by synchronous post-synaptic activity. Furthermore, various studies have shown Event Related Potentials (ERPs) in association areas of the cortex, showing high levels of local synchronicity after a perceived input or thought (Singer & Gray, 1995). And even decreasing LFP's when a stimulus was not similar (odd) to a previous set of similar stimuli. For example, when a sequence of low pitched tones are subsequently presented and at random points a high frequency tone is inserted a decrease in the LFP can be seen, this effect is called mismatch negativity.

So too, while not directly caused by the concept of binding by confidence since they require more precise understanding of the nonlinear dynamics of neural populations or graph theoretical understanding of the neural networks, neural oscillations provide evidence for the binding hypothesis.

Box 1. Threshold Logic Units

Much can be said about neurons on various levels providing deep and interesting insights in the general function of cognition, disease, aging and much more. Neurons are, however, conceptually most easily described as Threshold Logic Units (TLU's). Input from other neurons is summed, weighed and, when a certain pre-determined threshold is exceeded, output is generated which serves as input toward subsequently other neurons.

Biologically, when a certain amount of ion-channels have opened enough to trigger a cascade of opening voltage-depended channels, an action potential will flow over the neuronal membrane towards the end its axon. There the change in membrane potential will cause vesicles with neurotransmitters to be released at the synaptic junction. At the receiving end of the synaptic junction, usually a dendritic spine of another neuron, the released neurotransmitters will cause conformational changes in its channels, which give the possibility of another action potential or cause changes in the epigenetics of the neuron through second-messenger systems.

The membrane dynamics are complicated and nonlinear and many factors contribute to the overall shape of the output function: types of ion-channels (including various sub-types), density of these channels, placement, shape of the dentritic tree and its spines, existence of myelin, etc.

Ultimately epigenetics determine much of the neurons behavior. Precise understanding of the expressed genes and their methylation patterns, the influence of RNA-i's and other co-factors has revealed that the inner world of a neuron is complicated. The basics of the TLU-paradigm however remain unchallenged.

Gamma band oscillations (30 – 100 Hz), for example, have often been implicated in the binding of sensory modalities. Furthermore, gamma and theta (4 – 7 Hz) band oscillations have been found in the hippocampus and enthorinal cortex during memory formation and retrieval (Klimesch et al., 2010; Axmacher et al., 2006). Both in support of a theory of binding through synchronicity and a role in memory formation.

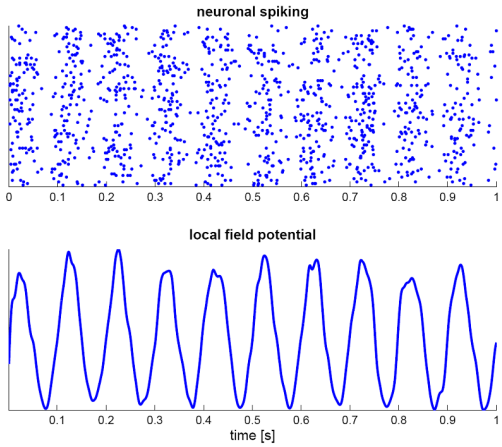

Figure 4. "Simulation of neural oscillations at 10 Hz. Upper panel shows spiking of individual neurons (with each dot representing an individual action potential within the population of neurons), and the lower panel the Local Field Potential reflecting their summed activity" (Wikipedia, 2011).

To summarize, the binding which allows us to perceive objects represented as activity in various brain areas can be solved using the principle of neural coincidence detection, which gives rise to synchronous activity that can be demonstrated with EEG.

But if highly synchronous activity happened without any form of temporal organization its representative power would be severely limited. To infer causality the activity must be somehow displaced in time. By changing the arrival time of neural activity, such as the case with delay lines, coincidences through time can be inferred (see figure 5).

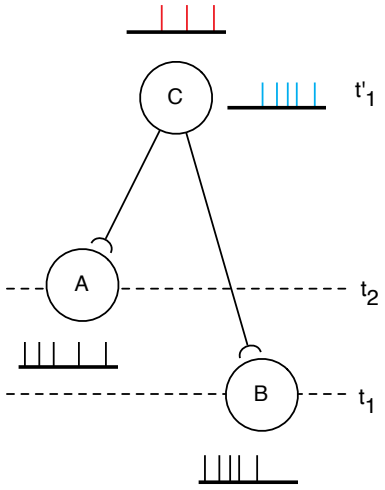

Figure 5. Causal inference through coincidence detection. Red spikes depict the firing of neuron C, blue the delayed activity of neuron B. Note that the length of the lines do not necessarily represent the length of the axon, but serve to depict the time necessary for a spike to reach neuron C; various morphological variations as well as nonlinear sub-threshold membrane dynamics can play a role.

Thus delays within the connections between the individual neural ensembles are necessary for the precisely timed spike train intervals to encode for causality and feedforward connections might introduce those delays by virtue of speed of neural transmission, sub-threshold membrane dynamics and various other mechanisms.

The last neuron in the chain might then actuate a certain behavior specific to a certain coincidence or causal rule. Memory in such feed-forward connections is, however, limited by the topological length of the chain.

Note that while it is a useful method of visualizing the neurons in the chain as depicted in figures 5 and 3 as fixed, in reality the activity can be largely independent of the topology. In other words, if the neural network is thought of as a network of roads and the spikes as cars then it is useful to note that for the cars to travel in a certain order or reach a certain destination many possible routes can be taken.

Recurrent topologies as memory

In order to introduce a form of memory, to keep patterns active, the concept of recurrence can be explored.

Recurrence refers to the idea that neurons are either directly or indirectly connected to themselves (figure 6). This allows them to activate themselves, possibly with a delay. For example, neuron A might activate neuron B which in turn activates neuron A again or a single neuron might activate itself, which causes a timed loop of action potentials, keeping patterns of synchronous activity locked.

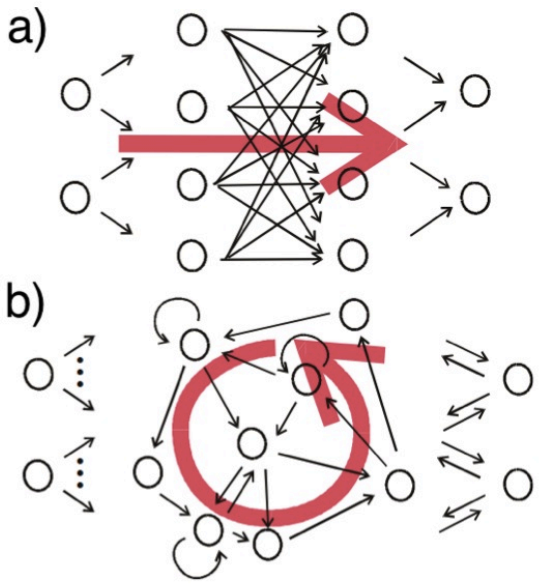

Figure 6. Two common types of neural network graphs. The network a) illustrates a feedforward pattern. Notice the ordering of layers. Network b) illustrates a typical recurrent network, such as a Hopfield (1982) network. (from Jaeger (2001)).

Therefore recurrent networks allow for a type of memory. They can store causal relations of the outside world by virtue of delays in synaptic transmission as well as their network organization.

It should be noted that like feedforward connections there is no need for any single neuron in these networks to be associated with only one trace of synchronous activity. Neither should any one trace be encoded in only one network. It seems possible that there are many different traces of synchronous activity within the neural network for a single causal relation, and new experiences (i.e. novel encoded sensor input) do not necessarily interfere with old traces. Indeed, neural networks are a highly distributed form of representing the causal rules and coincidences from the sensory input.

Mathematical descriptions of recurrent networks that are trained on certain input patterns by modeling Hebbian learning, such as Hopfield (1982) networks have shown to be capable of pattern completion and noise reduction in a content addressable fashion with multiple stable patterns.

More recent computational advances have also shown that sparsely distributed recurrent networks in the field of reservoir computing, such as Echo State Neural Networks (Jaeger, 2007) and Liquid State Machines (Maass et al., 2002, 2003), can be computationally trained to mimic complex nonlinear signals. In essence keeping the fed patterns active in a large randomly and recurrently connected reservoir. These complex nonlinear signals in turn can be interpreted as a mathematical description of real world causality, thus the ability to mimic them closely resembles that of memory.

To summarize, precisely timed synchronicity between spike trains (Kumar et al., 2010) in neural networks seems to be capable of this by binding multiple traces together, such as those from multiple sensor modalities. And networks with delays can encode for the causal temporal dynamics of the world. Activity in the recurrent connections seems to allow for memory. Precise timing of the delays in the networks create a representation, an echo, of the dynamics of the outside world: a neural representation of physical coincidence.

Plasticity as stabilizing

Research in the neurobiology of memory showed a secondary more long term effect in conjunction with the effects of spike activity in the neural networks. Synaptic plasticity like Long Term Potentiation (LTP) and Long Term Depression (LTD) improve or decrease the ability of the pre-synaptic cell to communicate with the post-synaptic cell bidirectionally. These processes are relatively slow, requiring synthesis of new proteins by gene expression. The effects are, however, long lasting and have been implied to play a major role in the formation of long term memory. This role is supported by the strong presence of LTP in brain regions commonly associated with the formation of long term memory, like the hippocampus and various neocortical areas (for an overview of neocortical plasticity see Feldman (2009)).

These processes have become neurobiological evidence for Hebbian learning, which states that neurons that fire in conjunction become associated ("neurons that fire together wire together"). Conceptually this idea was developed further to the concept of auto-association, stating that repeated effects of Hebbian learning form a pattern of association, called an engram.

Studies of the effects of varying the timing of weak and strong input from the enthorinal and the dental gyrus, however, found that the synaptic modification also depended on the temporal order of the inputs. Potentiation occurred when a weak input preceded the strong input by less than 20 ms, and a reversed order let to depression (originally Levy and Steward (1983), but see Caporale and Dan (2008) for a recent review).

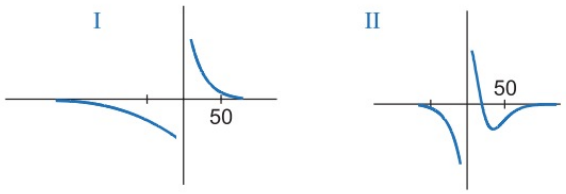

Figure 7. An example of STPD in an excitatory-excitatory connection, x-axes shows relative timing between pre-synaptic and post-synaptic stimulation in milliseconds, y-axes the change in sensitivity

Various types of plasticity dependent on the order arrival order of action potentials were found, which collectively became known as Spike Time Dependent Plasticity (STPD, see figure 7). STPD underlies the importance of the precise timing and order of spikes on a millisecond scale.

It seems that STPD allows synchronous activity that occurs often, such as the case with observer causal relations or the binding of specific properties, to stabilize or consolidate. So coincidences or causal relations bound by feed-forward patterns that occur often are favored by, STDP, but so too are coincidences that happen with the phase-looped locked activation in recurrent connections, creating a background, or echo, of favored activity.

What seems to be happening at a more meta-level was postulated by the connectionist Parallel Distributed Processing paradigm (Rumelhart & McClelland, 1986). Gradual changes in the parallel distributed processing of the brain allows for pattern completion and future prediction. These gradual changes happen in the direction of the perceived input, that is our inner neural representation of the dynamics of the world becomes more similar to what we perceive

A formal model of this gradual updating is known as the Temporal Context Model (TCM) (Howard M.W. & Kahana M.J., 2002; Howard et al., 2005). Inputs that are similar to the *inner context* are more likely to activate and gradually update that *inner context* in the direction of that input. TCM is a form of mental drifting that explains why more recent things are easier to remember than older things.

In summary, the world behaves according to specific causal dynamics. These causal dynamics, coincidences, are encoded into specific traces of synchronous activity which due to recurrent connections form reminiscent background activity, like an echo. When new sensory input arrives this is overlaid on top of the background activity, filtering input that fails to exceed thresholds but also completes patterns. This pattern completion not only happens on a semantic level, but also on a temporal (causal) level, allowing inferences about the future to be made.

Due to the precise timing of spike trains in these traces some connections are strengthened or weakened due to the effects of STDP. This strengthening and weakening happens in a gradual manner, slowly updating the background activity to represent that of the outside world. This process inherently favors coincidences that occur often, not only in the real world but also those that coincide with the background activity. It allows for a more robust way of completing patterns (spreading of activation) and predicting the future.

This means parting from the idea that memories are somehow stabilized and stored away for later use. Instead, they are an integrate part of thought and far more dynamic and versatile. The notion of consolidation might therefore be a misnomer, memories might actually never truly stabilize but are constantly adjusted. Indeed the neural activity in our brain is both memory (consolidation) and thought (retrieval) at the same time, mimicking input from the real world. It allows for vast pattern completion and both spatial and temporal reasoning, for example inferring that glass that falls might break.

But how does this constitute the experience of reliving a memory, or indeed consciousness?

Memory & consciousness

Suppose that by means of a very specific time-machine your brain from moments in the past can be transplanted in its entirety into you right now. This would mean that all the activity in your previous brain is now your current activity, including sensory input. It would be possible that for the short moment while new sensory input wasn't picked up, which would cause a realization that some the expected (predicted) state of the outside world was incorrect, you would consciously feel exactly like that previous moment. Now, indeed this moment would be very brief, since encoding new sensory percepts happen on millisecond scale and the dynamics are out of sync with the synchronous activity that neural network expected (causing a reduction in synchronicity).

Now lets suppose that instead of your entire brain only the parts that process sensory information were transplanted, and that somehow new sensory input was either suspended or not attended to. It would be entirely reasonable to assume that you would perceive, in all its sensory modalities, that exact moment your brain was transplanted from. The transplanted neural activity would be matched against the background activity reminiscent in the recurrent networks that was favored by STPD and for that moment you would actually be conscious of a previous moment.

These two thought experiments illustrate only an instantaneous transplant, but we can instead take another route. Assume that of a one moment transplant, neural activity is somehow streamed from the past. So lets say we suspend all brain activity of a willing participant right now and start streaming activity from his or her past. Instead of an image now a stream, analogous to signal, is consciously experienced by the participant. Now when the stream is stopped, neural activity by virtue of the recurrent connections and spontaneous activity, will cause causal predictions to be made. The neural network will walk a path towards a stable prediction of a future state. This might be experienced then as a mind wandering off, remembering and predicting future moments.

But it is possible for this to happen without intervention of a brain time-transplant, because it is exactly what the (recurrent) neural networks do. All the time sensory information is overlaid on the already present activation, finding coincidences and strengthening those connections, and keep some connections active in recurrent activation. But in absence of sensory input, which might be modulated by serotonergic system driven attention, the brain is able to re-induce past activity, the mind so to speak relives.

Recent research in Resting State Networks revealed what has become known the Default Mode Network (Greicius, Krasnow, Reiss, & Menon, 2003; Sheline et al., 2009; Greicius, Supekar, Menon, & Dougherty, 2009) a set connected brain regions often correlated with consciousness and attention (Medial Temporal Lobe, Medial Prefrontal Cortex and Posterior Cingulate Cortex) that become more active when the brain is not attending a particular task. Abnormal activity in the Default Mode Network is now even being explored as a possible diagnostic for the depression and other mental illnesses (Buckner, Andrews-Hanna, & Schacter, 2008; Broyd et al., 2009).

In absence of input, or in abundance of a modulatory attention effect (effectively "tuning out" sensory perception), the mind will start wander into what it has perceived might perceive. Evolutionary depression has even been explained as a way for the mind to attend for extended periods of time to problems that otherwise are too hard to solve (Andrews, 2007).

But is this not exactly the sought conjunction? Is the idea of being able to consciously remember a previous moment not exactly analogous to a reminiscent recurrent activation which tries to infer causal relations by synchronicity? When parting from the idea that somehow a separate stage of consciousness and that memory serves a need to fixate the past, it seems plausible that the two inherently conjoined. It requires a certain leap of faith to find a connection between many seemingly disjunct phenomena, but one that might give satisfying answers.

Discussion

There is a reason that in textbooks on cognitive neuroscience when reviewing memory there is almost no mention of consciousness. Some textbooks fail to address the question of consciousness in its entirety, because consciousness is hard. Not hard in the sense that its inherently a difficult concept. But, in the sense that it is difficult, even for contemporary science, to make a case for something that most people feel very strongly about. As Dennett (2007) once put it "everybody is an expert on consciousness". So in part this article was about finding a way to subtly work in a definition of consciousness from a subject less stigmatized, namely memory.

Memory was chosen because it offers a wide array of neuroscientific data, and still offers a very direct way of talking about consciousness. We know what it is to remember.

But during the process of writing the article it became overwhelmingly clear that not all the data could be convincingly reconciled. How neural oscillations are caused and what their precise relation with recurrence is, is still a matter of debate, and it is notoriously hard to reason from single neurons upwards to the anatomical level.

Some crucial steps therefore had to be left out for simplicity sake, for example it has been shown that when memories are retrieved the same brain regions necessary for encoding the memory are activated. A visual memory seems to reside in the visual cortex. Then to create a coherent scene thalamo-cortical oscillations have been proposed.

Furthermore, the hippocampal area has been proposed as a fast system for memory formation. Providing links to older traces and incoming information from the association areas (through the entorhinal cortex) in the neocortex which allows patterns to be completed based on existing traces. Gradually these links seem to cause reorganization in the neocortical areas in such a way that the trace becomes independent of the linkage from the hippocampal areas.

This neocortical-hippocampal system of memory formation has been called the Complementary Learning System, (for a review see Simons & Spiers, 2003; Ron Sun, 2008). And to minimize interference between non-overlapping episodes the hippocampus has a tendency to assign separated patterns to episodes in the hippocampal CA3-subfield. This pattern-separation is also greatly facilitated by strong inhibition from the Dentate Gyrus (DG). The DG has almost no overlapping representations and to ensure a consistency, new neurons are constantly added through neurogenesis.

But how this anatomical data can be readily translated to the discussed material was ultimately beyond the scope of this article.

Empirically there is still the question how to test this proposed conjunction. Advances in optogenetics and synthetic biology might one day serve to actually perform the type of brain time-transplant, but to this day remains science-fiction. More plausible ways are *in vitro* modulation of the spike trains to see how precisely this can alter behavior with and without STDP, predicting that disruptions of precise spike time activity without STPD might result in more catastrophic changes.

In humans experiments with Transcranial Magnetic Stimulation (TMS) might serve to interfere with synchronistic in cortical areas, which combined with a test for consciousness might serve as circumstantial evidence.

It can even be speculated that epileptics or subjects to TMS would suffer from memory loss and changes in conscious experience without the slower learning mechanisms of STPD.

Furthermore parting from the idea that memory is static also opens up the possibility to see memory as modulatory for attention and even conscious perception itself. Recent research into the perception of visual motion indeed shows that memory can skew the perception of a scene (Kang, Hong, Blake, & Woodman, 2011), and the proposed conjunction predicts that these effects are even more widespread.

And Artificial Intelligence might be able to one day recre-

ate the hardware necessary for the type of memory employed by neurons, which might give hints towards an artificial consciousness.

In the article *As we may think* Bush (1945) not only envisioned hypertext (the core of the internet), but also envisioned science to be more than mere information gathering. He warned that taking the time to think and explain is still important, but that scientists are less and less able to do so. To quote from the article: *"The investigator is staggered by the findings and conclusions of thousands of other workers–conclusions which he cannot find time to grasp, much less to remember, as they appear"* and more than 6 decades later this might be more relevant than ever.

Therefore ultimately the hope is that this article was able to combine some of the recent advances, and set out a novel line of reasoning by taking the time to take a step back and look at the truly staggering amount of acquired data and ideas. Such that in the future experiments can be designed to investigate the connection between memory and consciousness as part of the same neural process.

Acknowledgements

K. Jansen and I. Slingerland for their additional proofreading and support. Dr. F. Cnossen for allowing me to do this project autonomously and her faith in my work.

Licence

References

Andrews, P. W. (2007). The functional design of depression's influence on attention: A preliminary test of alternative control-process mechanisms. *Evolutionary Psychology*, *5*(3), 584–604.

Axmacher, N., Mormann, F., Fernández, G., Elger, C. E., & Fell, J. (2006, August). Memory formation by neuronal synchronization. *Brain research reviews*, *52*(1), 170–82.

Bailey, A., & O'Brien, D. (2006). *Hume's 'Enquiry Concerning Human Understanding': A Reader's Guide*. Continuum.

Broyd, S. J., Demanuele, C., Debener, S., Helps, S. K., James, C. J., & Sonuga-Barke, E. J. S. (2009, March). Default-mode brain dysfunction in mental disorders: a systematic review. *Neuroscience and biobehavioral reviews*, *33*(3), 279–96.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008, March). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38.

Bush, V. (1945). As we may think. *The Atlantic*, *176*.

Caporale, N., & Dan, Y. (2008, January). Spike timing-dependent plasticity: a Hebbian learning rule. *Annual review of neuroscience*, *31*, 25–46.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6).

Creath, R. (2011). Logical empiricism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2011 ed.). http://plato.stanford.edu/entries/logical-empiricism/.

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

Crick, F., & Koch, C. (2003, February). A framework for consciousness. *Nature neuroscience*, *6*(2), 119–26.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience* (Vol. I; M. MIT Press Cambridge, Ed.). MIT Press.

Dennett, D. (2007). http://www.ted.com/talks/dan_dennett_on_our_consciousness.html. Technology, Entertainment, Design. ([Online; accessed 17-July-2011])

Edelman, G. M., Gally, J. a., & Baars, B. J. (2011). Biology of Consciousness. *Frontiers in Psychology*, *2*(January), 1–7.

Feldman, D. E. (2009, January). Synaptic mechanisms for plasticity in neocortex. *Annual review of neuroscience*, *32*, 33–55.

Frankland, P. W., & Bontempi, B. (2005, February). The organization of recent and remote memories. *Nature reviews. Neuroscience*, *6*(2), 119–30.

Fries, P. (2009, January). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual review of neuroscience*, *32*, 209–24.

Gray, C. (1999). The Temporal Correlation Hypothesis of Visual Feature Integration : Still Alive and Well. *Neuron*, *24*, 31–47.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003, January). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(1), 253–8.

Greicius, M. D., Supekar, K., Menon, V., & Dougherty, R. F. (2009, January). Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cerebral cortex (New York, N.Y. : 1991)*, *19*(1), 72–8.

Herrmann, C. S., Munk, M. H. J., & Engel, A. K. (2004, August). Cognitive functions of gamma-band activity: memory match and utilization. *Trends in cognitive sciences*, *8*(8), 347–55.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005, January). The temporal context model in

spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological review*, *112*(1), 75–116.

Howard M.W., & Kahana M.J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 31.

Hume, D. (1748). *An enquiry concerning human understanding*. http://www.gutenberg.org/ebooks/9662. Project Gutenberg. ([Online; accessed 17-July-2011])

Jaeger, H. (2001). *The "echo state" approach to analysing and training recurrent neural networks* (No. GMD Report 148). GMD Forschungszentrum Informationstechnik, Sankt Augustin.

Jaeger, H. (2007). *Echo state network* (Vol. 2) (No. 9).

Kang, M.-S., Hong, S. W., Blake, R., & Woodman, G. F. (2011, June). Visual working memory contaminates perception. *Psychonomic bulletin & review*. Available from http://www.ncbi.nlm.nih.gov/pubmed/21713369

Klimesch, W., Freunberger, R., & Sauseng, P. (2010, June). Oscillatory mechanisms of process binding in memory. *Neuroscience and biobehavioral reviews*, *34*(7), 1002–14.

Kumar, A., Rotter, S., & Aertsen, A. (2010, September). Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. *Nature Reviews Neuroscience*, *11*(9), 615–627.

Levy, W., & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*(4), 791–797.

Maass, W., Natschl, T., & Markram, H. (2003). Computational Models for Generic Cortical Microcircuits. *Computational neuroscience*, 1–26.

Maass, W., Natschläger, T., & Markram, H. (2002, November). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural computation*, *14*(11), 2531–60.

Nalbantian, S., Matthews, P. M., & McClelland, J. L. (2010). *The Memory Process: Neuroscientific and Humanistic Perspectives*. The MIT Press.

Poggio, T., Reichardt, W., & Hausen, K. (1981). A neuronal circuitry for relative movement discrimination by the visual system of the fly. *Naturwissenschaften*, *68*(9), 443–446.

Ron Sun. (2008). *The Cambridge Handbook of Computational Psychology (Cambridge Handbooks in Psychology)*. Cambridge University Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1; D. E. Rumelhart & J. L. McClelland, Eds.) (No. 2). MIT Press.

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., et al. (2009, February). The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 1942–7.

Simons, J. S., & Spiers, H. J. (2003, August). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature reviews. Neuroscience*, *4*(8), 637–48.

Singer, W. (1999). Neuronal Synchrony : A Versatile Code for the Definition of Relations. *Neuron*, *24*, 49–65.

Singer, W., & Gray, C. M. (1995, January). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, *18*, 555–86.

Treisman, a. (1996, April). The binding problem. *Current opinion in neurobiology*, *6*(2), 171–8.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Wikipedia. (2011). *Neural oscillation — Wikipedia, the free encyclopedia.* Available from http://en.wikipedia.org/wiki/Neural_oscillations ([Online; accessed 22-Januari-2011])