In Proceedings AISB 2006 Symposium on Integrative Approaches to Machine Consciousness

Towards a Computational Account of Reflexive Consciousness

Murray Shanahan

Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK.
m.shanahan@imperial.ac.uk

Abstract

This paper offers a preliminary sketch for an account of reflexive consciousness based on an implemented architecture that combines a global workspace architecture with an internally closed sensorimotor loop. The proposed account extends the theoretical framework of the already implemented architecture with to two concepts that structure the flow of consciously processed information. First, contextual switches divide the unfolding contents of consciousness into a set of nested episodes, wherein one conscious episode can "refer to" another. Second, the imposition of a focus / fringe structure enables consciousness to encompass material that is merely available to it but not actually present. This combination of reflexivity and fringe may underpin our awareness of our own existence as conscious beings.

1 Introduction

Cognitive theories of consciousness, as the name suggests, posit an intimate link between cognition and consciousness. For example, according to *global workspace theory* (Baars, 1988; 1997; 2002), non-conscious information processing in the human brain is carried out locally within specialist brain processes, while the hallmark of consciously processed information is that it is broadcast (via a "global workspace") and made available to the entire set of these specialists. The upshot is that consciously processed information is cognitively efficacious in ways that non-consciously processed information is not. Specifically, the procession of broadcast global workspace states resembles a serial thread of computation, yet it integrates the results of massively parallel computation, sifting out relevant contributions from the irrelevant (Shanahan & Baars, 2005).

However, one feature of conscious human thought not accounted for by global workspace theory in its basic guise is *reflexivity*, that is to say the capacity for a conscious thought to refer to itself or to other conscious states. (By contrast, so-called higher-order thought (HOT) theories of consciousness take reflexivity as their primary datum (Rosenthal, 1986).) If consciously processed information is, as global workspace theory maintains, cognitively efficacious, then reflexively conscious information processing is even more so – since it enables the thinking subject to reflect on his or her own mental operations, to critique them and improve on them, and to respond to the ongoing situation in ways that depend on a degree of self-knowledge. So the question arises: Can global workspace theory be extended to account for reflexive consciousness?

This question has phenomenological as well as cognitive implications. For if we accept the argument of Shanahan (2005), the very idea of a conscious subject – something it is like something to be, in Nagel's well-known terminology – can be objectively accounted for in terms of a suitably *embodied* instantiation of the global workspace architecture, wherein all the specialist processes are indexically directed towards maintaining the wellbeing and fulfilling the purpose (or "mission") of a single, spatially unified body. By extending global workspace theory to reflexive consciousness, we can bolster this line of argument by showing that a similar treatment is available for a vital aspect of human phenomenology, namely our ability to become conscious of our own existence *as* conscious subjects.

2 Internal Simulation with a Global Workspace

Figure 1. illustrates the operation of the global workspace architecture, which comprises a set of specialist brain processes plus a global workspace. Information processing within the architecture consists of periods of *competition* interleaved with periods of *broadcast*. On the left of the figure, we see the set of specialist processes competing to gain access to the global workspace. Gaining access entails that the winning process (or coalition of processes) gets to broadcast its message, via the global

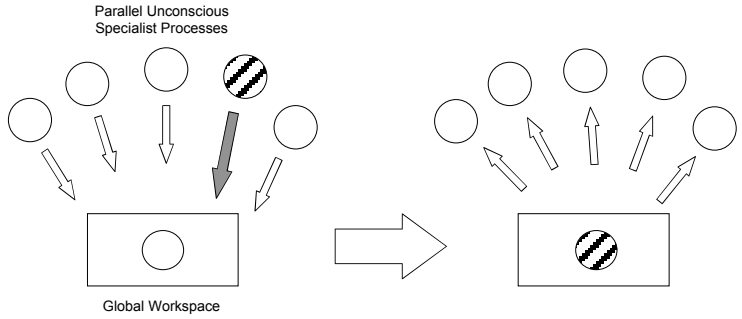

Fig. 1: The Global Workspace Architecture.

workspace, back to the entire set of specialists, as seen on the right of the figure. The global workspace itself is, in essence, nothing more than the infrastructure of a communications network that permits signals generated within localised neuronal populations to influence remote, widespread brain regions. According to global workspace theory, the mammalian brain instantiates such an architecture, and this allows us to draw an empirically falsifiable distinction between consciously and non-consciously processed information. Information processing that is confined to local specialists is necessarily non-conscious, and only broadcast information can be consciously processed.

Although global workspace architecture permits this fundamental distinction to be drawn in a theoretically respectable manner, it still leaves open the question of the content of consciously processed information. But by augmenting the basic global workspace architecture with an internally closed sensorimotor loop (Fig. 2), it is possible to reconcile it with another idea current within the scientific study of consciousness, namely the *simulation hypothesis*, according to which thought is internally simulated interaction with the environment (Cotterill, 1998; Hesslow, 2002; Shanahan, 2006). If the sophisticated mental life of a human being results from the interplay of external stimulation with internally generated activity such as inner speech and mental imagery, then something like the internally closed sensorimotor loop posited by the simulation hypothesis is required to account for it. Moreover, by facilitating the *rehearsal* of trajectories through sensorimotor space, the internal sensorimotor loop helps the individual to anticipate the consequences of their actions and to plan ahead, and thereby fulfils a fundamental cognitive role.

In (Shanahan, 2006), an implemented system is described that reconciles global workspace theory with the simulation hypothesis. The system controls a simple two-wheeled robot with a camera, and enables it to select an action based not only on a set of reactive responses, but also taking into consideration the result of simulating the expected outcomes of its actions using an internal sensorimotor loop, as depicted in Figure 3. Moreover, a global workspace is incorporated into the loop. The procession of states exhibitted by the global workspace, which simulates a possible trajectory through the robot's sensorimotor space, is the outcome of both competition and broadcast : the i^{th} state being broadcast to multiple neuronal populations which then compete to determine the $i{+}1^{\text{th}}$ state. Further details of the system are beyond the scope of this article, and can be found in (Shanahan, 2006).

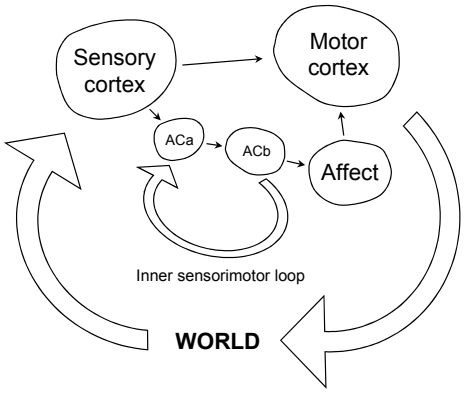

Fig. 2: External and Internal Sensorimotor Loops

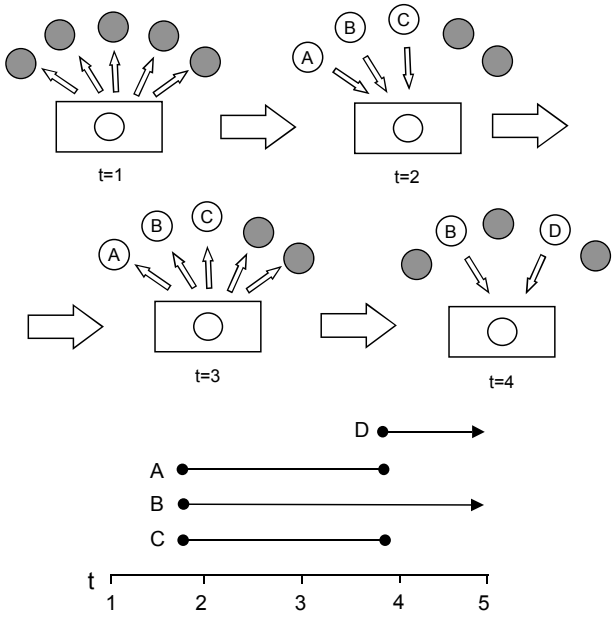

Fig. 3: The Temporal Structure of Consciously Processed Information

3 Context and Temporal Structure

According to the account of reflexive consciousness proposed by this paper, the flow of information through the global workspace is divided into distinct, but possibly nested or overlapping, episodes at various timescales. Beginnings and ends of conscious episodes are triggered by events (contextual switches) – such as entering and leaving a room, or meeting and parting from a friend – which wake up or send to sleep relevant specialist processes, whose job it is to manage the individual's response to situations of that particular type. Figure 3 illustrates the idea. The top of the figure shows snapshots of the global workspace (GW) at four time points. At t=1, all five of the processes depicted are dormant, although they are still receiving information broadcast from GW. This information indicates the occurrence of a distinctive event – a *contextual switch* – and by t=2 this has caused three of the processes (A, B, and C) to become active and begin competing for access to GW. There follows a further period of broadcast (t=3), indicating a new contextual switch. By t=4, this has caused processes A and C to go back to sleep, but has woken up process D.

As Figure 3 (top) shows, competition for access to GW is restricted to the currently active or "awake" set of specialist processes, and the set of active processes can be thought of as reflecting the current *context*, a conception which is broadly in line with the notion of context prominently deployed by Baars (1988) in his original presentation of global workspace theory. Each distinct conscious episode, bracketed by a pair of contextual switches, falls under the jurisdiction of a particular process, a process that should be relevant in the current context. Intuitively, temporal context is a richly structured, hierarchical concept. The context of a lunchtime falls within the larger context of a day, while the context of a conversation can overlap the context of a lunchtime. Similarly, conscious episodes, which are associated with temporal contexts, can be nested or overlapping. However, it should be noted that diagrams such as Figure 3 (bottom) only show the set of processes that have the *potential* to contribute to the unfolding content of the global workspace at any given time point. For example, although process B is *active* at time t=3 in Figure 3, this does not entail that it has won (full) *access* to GW at time t=3. This means that (focal) consciousness typically does not contribute to a conscious episode for its entire duration, but only at those times when the corresponding process gains access to GW. On the other hand, as we'll see in the next section, any active process competing for access can contribute to *fringe* consciousness.

Allowing specialist processes to wake up and go to sleep in response to contextual cues gives them a simple form of internal state (on or off), and therefore allows them to respond to information in a way that is sensitive to past events. But from the standpoint of the present paper, the most important consequence of this demarcation of conscious episodes is

that it allows one such episode to "refer to" another. This could occur either when the referring episode of conscious thought falls entirely within the episode it is referring to (Fig. 4, left), or when the referred-to episode of conscious thought and the referring episode of conscious thought both occur within a third, enclosing conscious episode and the former occurs before the latter (Fig. 4, right). In either case, the referred-to episode might be the an ongoing experience, the recollection of an experience from the distant past (long-term memory) or the recent past (working memory), or part of an ongoing rational or creative process involving inner rehearsal. A typical referring (reflexively conscious) episode might offer some judgement on the (non-reflexively conscious) episode it is referring to, such as "that was unpleasant" or (for a reasoning process) "that hasn't got me any further".

4 Focus and Fringe

The above characterisation of a reflexively conscious thought as a conscious episode that "refers to" another conscious episode is all very well. But it leaves open many questions, including that of the mechanism by which this reference is achieved. So to flesh out our account of reflexivity, something further is required. According to the present treatment, in addition to the temporal structure described above, the flow of consciously processed information has a focus / fringe structure (Mangan, 1993; 2001). The fringe contains hints of material that has the potential to be brought into focal consciousness if required. As Mangan (2001) puts it, "The fringe creates a non-sensory feeling of imminence which implies the existence of far more than consciousness actually presents at any given moment. ... This is the fundamental trick that lets consciousness finesse its severely limited capacity ...".

The contention of this paper is that this is indeed a "fundamental trick", a means to enhance the cognitive efficacy of conscious information processing in many ways. Of especial interest here is the fact that, at any given time, while focal consciousness is contributing to one conscious episode, broadcasting information supplied by the corresponding active process, the fringe can simultaneously retain the trace of *another* co-occurring conscious episode, governed by a different active process. To see this, consider Figure 4 (right). Suppose that at time t=3 active process Z has won access to GW, and is therefore supplying the current content of focal consciousness. At the same time, although process Y is not enjoying (full) access to GW, it is still active, and can therefore influence fringe consciousness.

We have the outline, here, of a mechanism by which one conscious episode can refer to another, wherein the referring episode is in focal consciousness while fringe consciousness retains a trace of the referred-to episode. But to see how this might be realised more concretely we need to zoom in and examine the evolving contents of GW at a finer timescale. In the computer model described in (Shanahan, 2006), GW was implemented as an attractor network. During execution, GW exhibited periods of stability (broadcast) during which it settled into an attractor, punctuated by periods of rapid change (competition) during which it got nudged out of a previously stable attractor and taken into a new one. During the periods of competition, it was sometimes observed that faint hints of competing attractors would become temporarily overlaid on GW's current attractor, each trying to take over.

This suggests the possibility that fringe consciousness might be realised as a rapid series of *faintly pulsing attractors*, each of which becomes transiently overlaid on the current attractor, but none of which yet has enough influence to dominate GW completely. (The dynamics here is reminiscent of Bressler & Kelso's (2001) notion of *metastability*.) Because these brief attractor pulses occur in GW, they are broadcast, and can therefore contribute to the flow of conscious information, as global workspace theory requires. Now we can appeal to *temporal synchrony*, as postulated by various authors as a solution to the binding problem (von der Malsburg, 1999), to realise reference between conscious episodes. The process currently supplying the content of focal consciousness – that is to say, the process associated with the referring episode – simply has to wait for the attractor corresponding to the process associated with the referred-to episode to pulse in

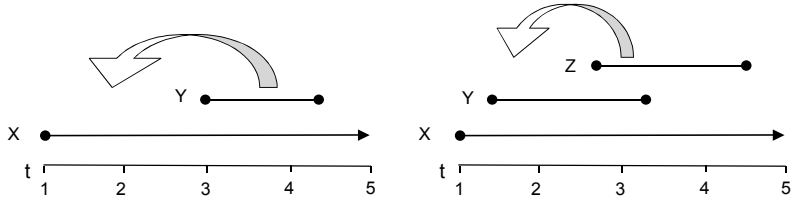

Fig. 4: Reflexively Conscious Episodes

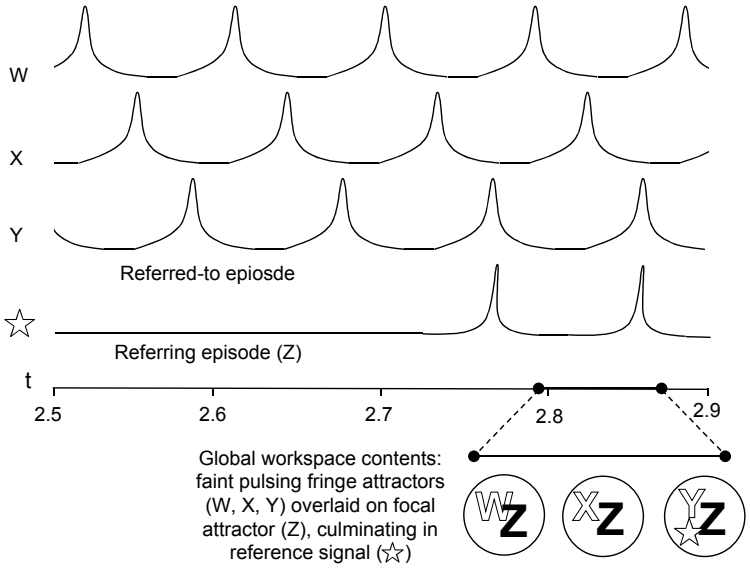

Fig.5: Focus-Fringe Reference by Temporal Synchrony

GW, when it can, so to speak, signal "THAT ONE" to GW (Fig. 5). Since this signal will be broadcast at the same time as the attractor pulse of the referred-to episode, the required reference will be secured.

5 Fringe-Borne Self-Awareness

According to the simulation hypothesis, conscious thought is simulated interaction with the environment. This entails that insofar as a conscious experience relates to anything other than an immediately present stimulus, the information processing that underpins it, as well as implicating the broadcast mechanism of the global workspace, must recruit a higher-order, internally closed sensorimotor loop (Fig. 2). This is the case for both the recall of a past conscious episode and the conscious rehearsal (or imagination) of a trajectory through sensorimotor space, where the latter conception encompasses inner speech, mental imagery, and so on.

Now, the fundamental role posited for the fringe is to augment the flow of consciously processed information with an awareness of the many possible ways that the content of the GW could unfold from its present state, without having to supply detailed information about any one of those possibilities. For example, our awareness of the three-dimensionality of a solid object can be cashed out in terms of a fringe awareness of a host of sensorimotor possibilities, such as moving around to view the back of the object, or picking it up and rotating it to see a different facet.

In the context of an internal sensorimotor loop, the fringe carries an awareness of the tree of possiblilities for conscious recall or rehearsal that branches out from the GW's current state (Fig. 6). Now suppose that, using the mechanism outlined in the previous section, one (reflexively) conscious episode Z refers to another conscious episode Y with the thought "that didn't work because P" (in the case of recall) or "that wouldn't work because P" (in the case of rehearsal). Then, thanks to the broadcast of this message, the entire set of specialist, unconscious processes will be offered the challenge of finding a potential variation of Y in which P is not the case. The vast majority of these specialists will be irrelevant to Y. But any that are successful in finding a potentially useful variation will be able to promote, via the fringe, the possibility of rehearsing it properly. This shows how reflexive consciousness can marshall massively parallel resources to further increase the cognitive power of (non-reflexive) conscious information processing, which is itself more cognitively efficacious than non-conscious information processing.

To round off the account, let's develop further the parallel between fringe-borne spatial awareness (of solid objects, for example), and the fringe-borne awareness of the unfolding content of the global workspace itself. According to the present account, the conscious awareness of the three-dimensionality of nearby objects or of the space through which the body can move consists of hints in the fringe of a systematically organised set of possible trajectories through sensorimotor space. These hints are systematically organised in the sense that they conform to various constraints, which include the reversibility of certain actions (eg: moving forwards then backwards gets you back where you started) and the cyclic character of certain trajectories (eg: turning an object

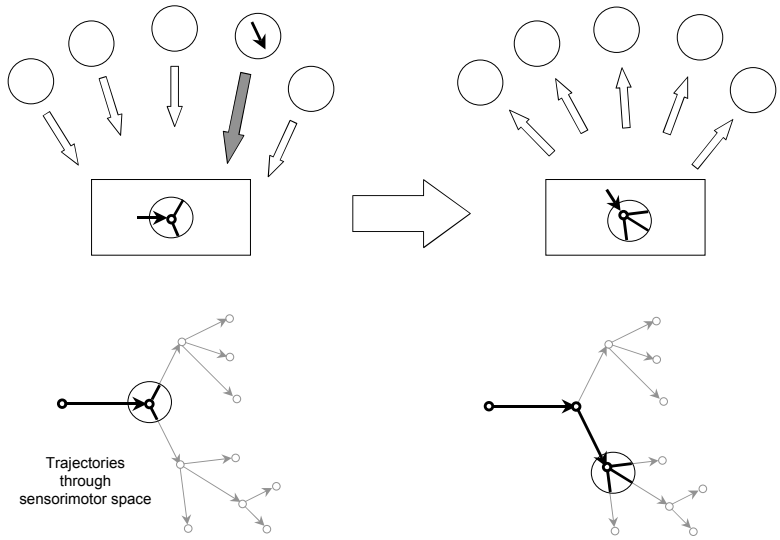

Fig. 4: Fringe-Borne Awareness of Possible Sensorimotor Trajectories

through 360º takes it back to its initial configuration).

In a similar vein, the fringe may sustain our awareness of the personhood of both ourselves and of others, hinting at material available for conscious rehearsal that pertains to our or their bodies, biographies, likes and dislikes, beliefs, desires, and intentions, skills and abilities, and so on. In the present context, the portion of this fringe-borne material of most interest relates to the way the content of the individual's consciousness unfolds. As the fringe-borne awareness of an object's solidity implies awareness of a systematic set of spatial constraints, so the fringe-borne awareness of personhood implies awareness of a systematic set of constraints on consciousness, such as its unity, its identity over time, and its indexical relationship to the body. Furthermore, in the same way that spatial constraints govern conscious thinking about solid objects, so these phenomenological constraints govern reflexively conscious thought. Insofar as we become consciously aware of ourselves as conscious beings, perhaps we do so thanks to our capacity to entertain reflexive thoughts combined with a fringe-borne awareness of the laws governing the way conscious thought unfolds.

References

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

Baars, B.J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press.

Baars, B.J. (2002). The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Science* 6 (1), 47–52.

Bressler, S.L. & Kelso, J.A.S. (2001). Cortical Coordination Dynamics and Cognition. *Trends in Cognitive Science 5* (1), 26–36.

Cotterill, R. (1998). *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.

Hesslow, G. (2002). Conscious Thought as Simulation of Behaviour and Perception. *Trends in Cognitive Science* 6 (6), 242–247.

Mangan, B. (1993). Taking Phenomenology Seriously: The "Fringe" and its Implications for Cognitive Research. *Consciousness and Cognition* 2 (2), 89–108.

Mangan, B. (2001). Sensation's Ghost: The Non-Sensory "Fringe" of Consciousness. *PSYCHE* 7 (18), http://psyche.cs.monash.edu.au/v7/psyche-7-18-mangan.html.

Rosenthal, D. (1986). Two Concepts of Consciousness. *Philosophical Studies* 49 (3), 329–359.

Shanahan, M.P. & Baars, B.J. (2005). Applying Global Workspace Theory to the Frame Problem. *Cognition* 98 (2), 157–176.

Shanahan, M.P. (2005). Global Access, Embodiment, and the Conscious Subject. *Journal of Consciousness Studies* 12 (12), 46–66.

Shanahan, M.P. (2006). A Cognitive Architecture that Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition*, in press.

Von der Malsburg, C. (1999). The What and Why of Binding: A Modeler's Perspective. *Neuron* 25, 95–104.