International Journal of Machine Consciousness

THE DESIGNER STANCE TOWARDS SHANAHAN'S DYNAMIC NETWORK THEORY OF THE "CONSCIOUS CONDITION"

Luc Patrick Beaudoin*

Faculty of Education, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada†
LPB@sfu.ca

Received Day Month Year
Revised Day Month Year

Shanahan expounds upon a creative and bold possibility, that concepts from dynamics and networks can be applied to Baar's [1988] global workspace architecture to explain fundamental biological and psychological phenomena. Shanahan proposes that the connective core is critical to cognitive and neuroscience. This review proposes to study the surrounding space of possible designs more systematically, with particular attention to purposive agency requirements and design assumptions which might resist the quantitative reduction Shanahan has attempted.

Keywords: designer-stance; attention; goal processing.

Murray Shanahan has admirably though briefly explained key facts about the mind and brain using network concepts and dynamic theory in a way that is compatible with Baars's [1988] global workspace architecture. His theory integrates phenomena-based, semantics-based and designer-based research ([Sloman, 1993]). His book invites one to ponder the possibility that attentional processes in humans may critically leverage an anatomical and functional network of major information pathways (including a "connective core"). The concept does not depend on the global workspace assumptions, as different theories may also leverage the concept. Shanahan proposes that conscious states involve the temporary control of the connective core by a single coalition of processes, a coalition which temporarily and exclusively leverages the connective core and can thereby communicate with a very large part of the brain (and presumably, mind). Shanahan describes the shifts of control of the global workspace using dynamics theory, wherein the quantitative concept of *attractor* plays a vital role. This singleton, the global workspace, is said to enable coherent and flexible behavior. Shanahan tantalizes the reader with specific central neural pathways that might implement the global workspace in humans.

The book shows us by example how Shanahan would have us speak about conscious vs. unconscious phenomena. The narrative is psychologically compelling. Yet the bulk

* CogZest. PO Box 18127. Port Moody, BC. V3H 4H0. Canada.
† State completely without abbreviations, the affiliation and mailing address, including country. Typeset in 8 pt Times Italic.

of the new design proposals come late in this book (from chapter 4 out of 6). The reader would have been taken even further had the Wittgenstein text (most of chapter 2 being irrelevant in my opinion) been omitted in favour of design-based exploration.

Exploring potentially implementable versions of Baar's global workspace architecture, as Shanahan has begun to do, is very important. My review is written mainly from the "designer-stance". My criticisms may be interpreted as suggestions and encouragement for future research.

While the book aims to explore the space of possible minds, it restricts itself to a very narrow portion of it. Shanahan states that "scaffolding is useless without the right architecture" (p. 177). How can one know (or defend) that one has the "right" architecture without documenting the space of architectures? Moreover, there is no "right" architecture. As Aaron Sloman put it "Humans don't necessarily all have the same architecture, and the very same human will have different architectures at different stages of development, from birth (or earlier) to old age. Nobody knows what all the transitions are." [2011] There might be a specified collection of architectures that could possibly be implemented by humans (or some other genome). The task then is not to find the *right* architecture, but to chart the space of those designs, including their trade-offs.

Exploring the space of possible minds cannot be done using phenomena-based methods alone, given that the realm of possibilities exceeds actualities. While Shanahan valiantly seeks empirical evidence for his musings, and ultimately pitches his contribution as an empirical one, the designer questions have precedence. For example, his assumptions about the brain engaging in simulated interaction with the world need much analytical support to demonstrate that they can possibly be true before data about whether they are true will be relevant—can the frame problem really be resolved in the way he suggests?

Exploration of design requires that one adopt the designer-stance. This involves (1) studying the environments or niches of one's theoretical objects (e.g., autonomous agents), (2) exploring requirements that are met by the objects (which are requirements for theories about them), (3) exploring possible designs to meet those requirements and (4) implementing and thoroughly testing some of those designs. Each of these activities, (1) to (4), yields "conceptual artifacts" (theoretical proposals, designs, etc.) One can then (5) analyze relations between these conceptual artifacts yielding yet more conceptual artifacts. It helps the reader for the author to distinguish between core assumptions and peripheral assumptions [Lakatos, 1980], labeling each (see [Cooper, 2007] for the distinction and [Cooper & Shallice, 2000] for an example of its use.) When exploring the space of possible minds, one should explore the effects of adding, removing, and modifying assumptions about the environment, requirements and design. The book's narrative does not facilitate the individuation and manipulation of core and peripheral assumptions.

Shanahan explores requirements when he asks "what is cognition for?" (p. 43). He essentially answers that it is adaptive in the Darwinian sense, and that it accomplishes this end by affecting behavior. How? "it is incorporated into an organism's sensorimotor

loop and thereby perturbs its behaviour." (p. 43) He suggests this role is effected by exploring a space of combinatorially structured affordances. The question "what is cognition for?" assumes that cognition is something discrete and well defined (which it is not). The subsequent question "what are the building blocks of thought?" (p.43) assumes that thought is something specific with components (which it is not.) And his answer leaves out many uses of cognition by humans that have nothing to do with sensorimotor loops—e.g. early astronomers trying to explain the observed motions of planets. A more germane question to launch inquiry into mental architecture is "what are the requirements of autonomous agents?" This is more suited to an engineering methodology and it can help sidestep debates about what "cognition" is.

Many of us who asked that question have answered with many requirements. One major requirement is the ability to generate and pursue one's own goals (see [Simon, 1987], [Beaudoin, 1995], [Boden, 1972]). While Shanahan admits the compelling case for goal-directed activity (p. 54), he does not explore purposive requirements systematically nor does he propose goals as a core theoretical construct. Yet goals have an important effect on what information becomes conscious. If the gardener described on page 72 had been looking for a leaf of a specific size, he might have noticed its shape; but his goal was different and hence he didn't. Even most functionally autonomous goals are neither ontogenetically nor dynamically derived from evolutionary ends. Baars [1988] appositely devotes two entire chapters to goals in relation to "consciousness". Baars states "attention is always controlled by goals, whether voluntary or not" [p. 304]. This requires some qualification, for many shifts of attention are reactive, e.g., a sudden noise can asynchronously disturb a planning activity as can an episodic memory or an obsessive thought, whether or not it is affectively valenced. Moreover, not all motivators are goals—the perception (or memory) of a motivating object can distract attention without particular goal states being created (cf. [Beaudoin, 1995], [Ortony, Clore and Collins 1988]). Still, Baars's chapters do meaningfully draw attention to goals.

While the notion of goals which Baars [1988] outlined is very rich it needs extension. Beaudoin and Sloman [1993; Beaudoin, 1995] set forth detailed requirements for goals and goal processing. An autonomous agent can produce goals that vary in importance, urgency, intensity, insistence, etc. This variability is not merely quantitative—goals evince rich referential and control semantics with dependencies. Further, all kinds of potentially provisional decisions can be made with respect to goals. Shanahan resists proposing data structures with compositional and referential semantics or at least sets up the task to account for such things in dynamic terms (e.g., grammar, complex planning). Shanahan also rejects Baars's notion of contexts (p. 96) which subsumes goals. Is Shanahan suggesting that the deliberation, assessment and selection underlying what at least from an intentional stance is purposive behavior can be implemented using attractors in state space? As argued in [Beaudoin, 1995], purely quantitative specifications of purposive control states (e.g., [Powers, 1973] which also eschews semantic messages) have severe limitations. The question for Shanahan is not whether the mind implements goal states as explicit data structures, but whether dynamics and

networks are sufficient for modeling the rich variety of goal states and processes alluded to above. More generally, in exploring the space of possible minds, it would be helpful to be explicit about the major requirements of autonomous agency that this particular design (or set of designs) does not address.

Apart from exploring different requirements, design space exploration also involves specifying different possible designs for the same set of requirements. This is vaguely analogous to the exploration of affordances that Shanahan frequently refers to (e.g., p. 58). For example, given the notion of a connective core, one could explore implementations of complex multi-step external behaviors that involve preparatory "low bandwidth" communications of coalitions which currently do not have full control of the connective core but might sensibly need to in short order. Assuming the global workspace principle is powerful, what would happen if one assumed that the modules themselves instantiated a global workspace architecture? Could attractors work within as well as between processes? How can one characterize the space of different architectures in which the connective core plays a major role but which are not constrained by all the other global workspace architecture assumptions? Many other design variants could be explored given the books subtitle.

Shanahan could counter that it is up to other labs to propose rival theories. However, the author of a theory is in a unique position not only to deepen his theory but to extend its scope, which may involve providing direction for teams of researchers working on variants of the general architecture. And it is important to anticipate the strongest arguments against one's opponents in order to improve one's theory or refute the others. This would include not only very different theories, but also potential similar explanations and designs.

One of the most important assumptions in cognitive science, assumed as a fact by most, is that attentional resources are limited (in this case, the global workspace). Yet, as Baars [1988, p. 110] notes, the question is rarely asked: why is there such a limited capacity? This is not a question that can be answered empirically, nor should it merely be assumed to be an evolutionary happenstance or a contingent fact about the brain, though it ultimately *might* be. Indeed, from the designer-stance, one of the most tantalizing and critical questions about the global workspace architecture, and indeed any theory that assumes central processing or internal communication limitations, is the very engineering (as opposed to empirical) reasons behind the bottleneck. Baars [1988] and Shanahan both directly address the question and their answers are highly significant scientific conjectures. In [Beaudoin, 1995, ch. 4], I describe several other reasons. Yet, we should not assume that we have a full answer to this critical question. Discovering other reasons (if they exist) or refuting it in design space would significantly enrich our understanding of the space of possible minds. Some researchers should attempt to build models that address the same agency requirements yet violate the limited attention assumption.

Shanahan expounds on fundamental concepts related to his theory, including those of process and concept, which require qualification. Shanahan courageously likens the individuation of processes to the formation and dissolution of crowds in a department

store. I say "courageously" because it does make one wonder whether processes have a meaningful role to play at all. If mental processes are as fluid as this, then something like dynamic theory may indeed be required. In his words "process should not be thought of as a neatly bounded computational unit with clearly defined inputs and outputs" (p. 98) and "Generally speaking, then, processes are resident in the brain at multiple scales, and their boundaries are fluid, indistinct, and overlapping." (p. 98) Yet he also claims that a process is an identifiable, separate and pluggable entity (p. 124, 170). It will be quite challenging for AI researchers to model processes in this way, as one naturally attempts to design specific mechanisms and processors, but the exploration is a worthy one. In describing processes as anatomically distributed (e.g., "coalitions of processes") Shanahan implicitly excludes the possibility of a single process flowing over the connective core—this may prove to be too limiting.

Shanahan's argument that concepts sould be considered as skill is weak. "to acquire a concept is to master a systematic set of skills, and to acquire an abstract concept is to layer a systematic set of thinking and talking skills on top of a foundation of more basic sensorimotor skills" (p. 59). Fortunately, this view does not lead to disasters in this particular book, but for its progeny I will explain some of what is wrong with it, since this widely held and tempting but erroneous view has many important ramifications for cognitive science and education. Scholars who champion the slogans of "embodiment", "situativity" and "connectionism" are apt to reject the objectivity of concepts perhaps because they believe it entails a correspondence view of mental states (where possibly propositionally represented content in the head matches content in the world). Yet, ironically, the mental skills approach to concept acquisition is the prevalent approach in education which implicitly embraces a correspondence view. Bereiter [2002] proposes that mastering a concept ought to be conceived as a potentially rich relation between the knower and an object of knowledge (e.g., a "conceptual artifact"), a relation which supports intelligent action. Bereiter's conception has no correspondence implication, is theoretically germane, and if applied properly could help to correct common egregious North America educational practices.

Having skills is often (though not always, e.g., with respect to the concept of democracy) an important but incomplete part of this relation. According to Bereiter, understanding a concept often entails a subset of the following: having statable knowledge (also known as "declarative knowledge"), implicit knowledge (including the kinds of simulation abilities Shanahan seeks to explain), episodic knowledge (e.g., of events when the concept applied), impressionistic knowledge (e.g., feelings and attitudes towards its objects), some interest in the class of object (e.g., normally someone who understands a tool very well is interested in using and/or talking about it), etc. There are of course many types of conceptual artifacts (e.g., theories, concepts, rules, axioms). And there are different types of concepts, e.g., basic-level concepts (such as "dog", "tree") and higher-order concepts (such as gravity). Shanahan is correct in saying that it does not matter whether concepts exist independently of the mind (in the Platonic sense); however, pragmatically, many ideas do have objective properties (we can refute them,

extend them, draw inferences from them, etc.) (See [Popper & Eccles, 1972].) This is not just metaphysical driftwood. There are many concepts whose understanding we cannot comprehend without reference to the objective conceptual artifact in relation to which the knower acts. An adequate analysis of 'having a concept', such as that of [Bereiter, 2002], entails many requirements for a theory of concept acquisition well beyond the systematic skill precepts (including "imagination") towards which Shanahan is attracted.

I had hoped that by 2010 rhetorical attacks on the "classical AI" straw man would have ended. They began over 20 calendar years ago—much longer if measured in terms of AI researcher person years. Authors should instead contrast their work, in detail, with specific theories including some that are similar to their own. Avoiding a reduction to dichotomy is particularly important for those whose aim is to explore the space of possible minds [Sloman, 1984].

This book has provided important new concepts about what might underlie distinctions between conscious and unconscious states.

Acknowledgments

Thanks to Aaron Sloman and Carol Woodworth for commenting on drafts of this document.

References

Baars, B. J. [1988]. *A cognitive theory of consciousness.* (Cambridge University Press, New York, NY).

Beaudoin, L. P., & Sloman, A. [1993]. A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for Artificial Intelligence* (Birmingham, England), pp. 229-238.

Beaudoin, L. P. [1994]. *Goal processing in autonomous agents.* Ph.D. thesis (Cognitive Science). School of Computer Science, University of Birmingham, (Birmingham, UK).

Bereiter, C. [2002]. Education and Mind in the Knowledge Age (544 p.). (Routledge: N.Y.)

Boden, M. A. [1972]. *Purposive explanation in psychology.* (Harvard University Press, Cambridge, MA).

Cooper, R. P. [2007]. The role of falsification in the development of cognitive architectures: Insights from a Lakatosian analysis. *Cognitive Science*, 31, 509-533.

Lakatos, I. [1980]. The Methodology of Scientific Research Programmes: Philosophical Papers (Vol. 1). (Cambridge University Press, Cambridge, MA).

Ortony, A., Clore, G. L., & Collins, A. [1988]. *The cognitive structure of emotions*. (Cambridge University Press, Cambridge, MA)

Popper, K. R., & Eccles, J. C. [1977]. *The self and its brain - an argument for interactionism* (p. 597). (Routledge & Kegan Paul, London).

Powers, W. T. [1973]. *Behavior: The Control of Perception*. (Aldine Publishing Co., Chicago)

Simon, H. A. [1967]. Motivational and emotional controls of cognition. Psychological Review, 74(1), 29-39.

Sloman, A. [2011]. Personal communication.

Sloman, A. [1993]. Prospects for AI as the general science of intelligence. In Aaron Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for Artificial Intelligence*. (Birmingham, England), pp. 1-10.

Sloman, Aaron. [1984]. The structure of the space of possible minds. In S. Torrance (Ed.), *The*

Mind and the Machine: philosophical aspects of Artificial Intelligence. (Chichester, England), (pp. 35-42) .