Characterizing Human-Like Consciousness: An Integrative Approach

Ben Goertzel
OpenCog Foundation, Hong Kong, China

September 7, 2014

Abstract

Synthesizing concepts and findings from a number of recent models of human consciousness, a unified model of the key properties characterizing human consciousness is outlined. Six key properties are emphasized: **Dynamical representation** of the focus of consciousness, **Focusing of energetic resources** and **focusing of informational resources** on a subset of system knowledge, **Global Workspace dynamics** as outlined by Bernard Baars in his cognitive theory of consciousness, **Integrated Information** as emphasized by Tononi, and **correlation of attentional focus with self-modeling**. It is proposed that the extent, and relative importance, of these properties may vary in different states of consciousness; and that any AI system displaying closely human-like intelligence will need to manifest these properties in its consciousness as well. The "hard problem" of consciousness is sidestepped throughout, via focusing on structures and dynamics posited to serve as neural or cognitive correlates of subjective conscious experience.

1 Introduction

These days, unlike a few decades ago, consciousness is a significant topic of research in psychology, neuroscience, philosophy and other fields. However, there remains no scientific consensus on how to define or conceptualize consciousness, let alone on how to quantitatively measure it, or formally model its structure or dynamics. Among other open questions, there is no broadly accepted way to measure the degree of consciousness displayed or experienced by a system (be it a human or other animal brain, or an robot or other AI) during a certain interval of time.

I will not aim, here, to address the foundational question of "what consciousness fundamentally is." Instead, the question I will focus on is: **What are the important properties specifically characterizing human, or human-like, consciousness?**. [1]

[1] Just as physics has told us many interesting and useful things about the movement of

I believe that this question admits *scientific* answers – but probably does not admit a *simple, elegant, unified* answer. Rather, I suggest that human consciousness is distinguished from "consciousness in general" by a mix of different properties, which have evolutionarily co-adapted to work together in a coherent way. At a high level, this same sort of "messy coherence" can be observed in many other examples in the domain of evolved systems – very often, in a biological or ecological context, we find that heterogeneous aspects of a system work together coherently, in a manner that works pragmatically, yet is somewhat "ad hoc" and specific to the functioning of some particular sort of system.

It is possible that there is some elegant, crisp, beautiful theory of human consciousness lurking around the corner, which we have not yet found because we've been approaching the topic with the wrong theoretical or empirical toolkit. However, I suspect that this is not the case – and that the quest for such a theory may be understood as yet another case of "physics envy", i.e. the fallacy of expecting or hoping complex and specific systems like ecosystems or human organisms to display regularities expressible as simple, mathematically aesthetic equations like the ones found in theoretical physics. It is with this in mind that I advocate here an expressly "multifactorial" approach to human consciousness.

The messy coherence of human consciousness is analogous to – and also closely related to – the messy coherence of human **intelligence**, which is a special case of "intelligence in general," with a host of special properties that evolved in reaction to the specific evolutionary needs of early humans and their predecessors. In particular, intelligence testing has something to tell us about the prospects for rigorous consciousness measurement.

The IQ test aims to provide a single number summarizing human general intelligence, but succeeds only very narrowly. Cultural biases in IQ testing are well substantiated [11]. Further, psychologists have proposed various "multiple intelligence" theories aimed at measuring various aspects of intelligence individually, arguing that the standard IQ test somewhat arbitrarily squashes multiple, largely distinct capabilities into a single numerical score [22]. Given the difficulty of establishing IQ testing as a measure among human beings, it seems clear that current IQ tests cannot meaningfully be applied beyond the human realm, e.g. to intelligent animals with fundamentally different natures such as cetacea, or to AI systems with cognitive architectures differing significantly from the human mind/brain.

On the other hand, mathematical measures of general intelligence have been proposed [36, 27], but these are very abstract, and it is not clear how to apply them in the context of everyday human intelligence. On a very qualitative level, one can summarize the nature of human intelligence concisely, with phrases like " the ability to learn and generalize", or " the ability to achieve complex goals in complex environments", etc. But when one tries to formalize ideas like these, one hits numerous thorny dilemmas, mostly revolving around the dichotomy and boundary between extremely broad theoretical problem-solving capability

objects without resolving all the core philosophical issues regarding the nature of space and time, I believe we can come to many valuable conclusions about consciousness without first needing to resolve all related philosophical perplexities.

(which quickly gets beyond the human level, when one considers it in the abstract), and highly specialized problem-solving capability (at which humans are good in certain domains, and terrible at in others). Human intelligence is indeed somewhat specialized, but also has an element of generality interwoven with the specialization – which makes it complex and complicated, in the typical manner of biological phenomena. And similarly, human consciousness has multiple aspects, which seem difficult to summarize in a single number.

The experience of human consciousness, while it often seems simple to humans experientially, may actually be a complex amalgam of different phenomena. The human mind/brain seems to contain many specialized forms of consciousness, which then weave together into an overall consciousness dynamic. Consciousness seems not to be a simple, unidimensional physical concept like energy or mass; but rather a complex, multidimensional psychological concept like intelligence or happiness – and as such is measurement becomes a complex, context-dependent matter of balancing multiple factors. The idea that some sort of "raw consciousness", with an elemental simplicity to it, may be immanent in the physical universe, doesn't really help with the pragmatic measurement of human or human-like consciousness – any more than intelligence testing is aided by the observation that some sort of "raw intelligence" is immanent in the universe due to the way basic physical dynamics implicitly optimizes complex objective functions in complex situations.

In this paper, a number of contemporary analyses of human consciousness are analyzed from this multifactorial perspective: Baars' Global Workspace Theory [7, 8] and the LIDA software system that partially embodies it [45]; Tononi's Integrated Information theory [48]; Goerner and Combs' analysis of consciousness in terms of nonlinear dynamics and energy minimization [23]; Tart's theory of states of consciousness [47]; and the analysis of consciousness in terms of reflective self-modeling [40, 41, 28]. It is argued that these theories, diverse on the surface, are actually elucidating different aspects of the same complex underlying human consciousness process.

Finally, the possibility of measuring the degree of consciousness displayed by a human or human-like system, using multiple factors derived from these multiple theoretical perspectives, is discussed. Six key factors relevant to measuring human-like consciousness are summarized as:

1. **Dynamical representation** of the focus of consciousness
2. **Focusing of energetic resources** on a subset of system knowledge
3. **Focusing of informational resources** on a subset of system knowledge
4. **Global Workspace dynamics**
5. **Integrated Information**, perhaps as quantified by Tononi
6. **Correlation of attentional focus with self-modeling**

The optimal ways to quantify all these phenomena are not yet clear; this is a topic needing further study. What is argued here is that, if one wishes to quantify the degree of consciousness of real-world systems, this is the right way to proceed – i.e., by identifying and then quantifying multiple aspects of the multifarious, multifactorial dynamical process that is human-like consciousness. That is, the goal here is not to propose a precise quantitative measure of human-like consciousness, but rather to lay out a clear conceptual framework, integrating relevant bodies of knowledge and theory, within which human-like consciousness can be qualitatively analyzed in a variety of systems (including AI systems), and within which comprehensive, precise quantitative measures of human-like consciousness can be pursued.

2 Aspects of Human Consciousness

Consciousness has been addressed from a variety of different vantages, far more than could be surveyed in a brief paper. In this section – which constitutes the bulk of the paper – I review a subset of the many important ideas from the literature, which combine together to form the overall perspective on consciousness outlined in the following section.

2.1 Hard and Possibly Less Hard Problems Regarding Consciousness

David Chalmers [13] famously distinguished the "hard problem of consciousness" from other issues regarding consciousness – where what he meant by the "hard problem" was, in essence, the problem of connecting subjective experience (the "raw feel" of consciousness, sometimes referred to using the term "qualia") with empirically observable factors. According to my own understanding, in the current ontology of intellectual disciplines, this "hard problem" is a philosophical rather than scientific problem. My reasoning is that science , as currently understood, is focused on prediction and explanation of measurements that are observable by an arbitrary observer within a community; whereas subjective experience, by its nature, is not observable by an arbitrary observer with a community. There is some wiggle room here, in that a community of meditators or psychedelic adventurers may consensually agree that they can sense one another's subjective experience, thus arguably bringing subjective experience within the domain of the interpersonally observable, at least with respect to that particular community. However, this sort of observability is different from the measurement as generally pursued in science, meaning that handling the "hard problem" in a scientific way would involve substantial extension or modification of the scientific method as commonly understood.

In reaction to the slipperiness of the "hard problem", many consciousness researchers have focused their attention on the perhaps easier, though still challenging, problem of finding "neural correlates of consciousness" [39], or else cognitive correlates of consciousness. The research question then becomes: What

are the patterns in the brain or mind that tend to correlate with *reports* of subjective experience? A related question is what Scott Aaronson has called the "pretty hard problem of consciousness" [2] – determining which kinds of systems are capable of having consciousness experience. Which kinds of systems can have physical correlates of consciousness at all?

The topic addressed in this paper is *the cognitive and neural correlates of human and human-like consciousness.* I believe this topic can be explored quite thoroughly without making any commitments regarding the "hard problem." However, before moving on, it would be dishonest of me not to clarify that I do have a personal and intellectual position on the "hard problem."

Some have argued that the "hard problem" is nonsense and qualia do not exist in any meaningful sense [16]; I am not one of these. Rather, I tend to agree with Chalmers that that some sort of panpsychism is probably the right answer – i.e. I tend to view consciousness as a property that everything in existence possess to some degree and in some form. Analytic philosopher Galen Strawson [46] has argued strenuously and rigorously that any other perspective is logically ill-founded; there is no consistent, sensible way to view consciousness and the physical world as separate but interacting entities. But if one accepts this, there remain difficult questions regarding why particular physical entities are associated with particular sorts of conscious experience. The philosophical and scientific aspects of panpsychism have been explored in detail by many others [46, 14, 4] and I will not repeat those discussions here. The central ideas in this paper are not predicated on the panpsychist perspective, so I mention my orientation toward panpsychism here mainly to point out that there is at least one simple, conceptually coherent answer to the question of the basic nature of consciousness, which appears to be fully consistent with the concepts discussed [2]. The reader is invited to explore panpsychism further, or else to harmonize the cognitive and empirical ideas presented here with their own different views on the fundamental nature of consciousness.

2.2 Degrees of Consciousness

Consciousness, from the perspective of a subjective human experiencer, seems not to be a Boolean, either/or phenomenon. Rather, there seems a subjectively clear notion of the *degree* of consciousness. For instance, there is a sense in which

- After I'm fully awake, I am *more* conscious than I am a half-second after I wake up in the morning
- I am *more* conscious of something at the center of my attention, than something at the fringe of my attention

[2]The main argument posed against panpsychism seems to be that people find it counterintuitive. However, science is frequently counterintuitive; and further, the view that panpsychism is counterintuitive is much more prevalent in the West than in the East or Africa

- A fully alert person has *more* consciousness than a fully alert chimp, worm, bug or rock

One important research question is: To what extent are these three different kinds of more-ness related? Is there a single axis of "quantity or degree of consciousness" to which they all refer? Or do they refer to different ways of quantifying/comparing instances of consciousness? One possibility is that single measure, or a certain set of measures, could be shown to lie at the center of all three of the kinds of more-ness I've mentioned.

A simple and compelling way to think about "degrees of consciousness" is as *degrees of conscious access* [21]. As Bernard Baars notes [6],

> Zero Conscious Access (Zero CA) is a perfectly acceptable label for brain events — a world of them — that never come to consciousness, but which need to be understood if we are to understand the dimension of CA. The dimension CA could be operationalized by specific measurable variables ... [e.g.] in the case of sensory events or recalled memories. Degree of drowsiness can also be assessed as an empirical CA measure, by counting the number of slow waves in the occipital EEG per second. Or in the case of surgical anesthesia, it can be assessed by the amount of inhaled anesthetic per second. Behaviorally patients may be asked to answer questions, or count to ten, etc.

Intuitively, the notion of CA has potential to encompass the three types of "degree of consciousness" mentioned above. A fully awake state of mind, compared to a half-awake state of mind, involves more entities having more conscious access; entities in the focus of attention are more accessible to conscious processes than those on the fringe; and a worm or a bug, very likely, has fewer and simpler items possessing conscious access during any given interval compared to a human (as these simpler animals, to the extent that they are viewed as conscious, likely have consciousnesses dominated by immediate perceptions and actions, whereas human consciousness gives access not only to these but also to various other memories, plans, ideas, etc.).

2.3 Specific Subprocesses of Human Consciousness

Cognitive psychology has traditionally said a lot more about "working memory" (or before that, " short-term memory") than about consciousness *per se*. However, there is a great deal of overlap between these topics. So it is highly relevant to any discussion of human consciousness to note some of the specific substructures and subprocesses that cognitive psychologists have identified within human consciousness, e.g. the classic 3 identified by Baddeley [9],

- A *phonological loop*, which deals with sound or phonological information, and is subdivided into a short-term phonological store with auditory memory traces that are subject to rapid decay and an articulatory loops that can revive the memory traces.

- A *visuospatial sketchpad*, which handles the temporary storage and manipulation of spatial and visual information, and also assists with tasks which involve planning of spatial movements. It is thought to handle visual, spatial and kinesthetic information in slightly different, though overlapping, ways.
- An *episodic buffer*, which is concerned with linking information across domains to form integrated "episodic" units of visual, spatial, and verbal information, such as the memory of a story or a movie scene

It would be hard to argue for the necessity of such particular structures and processes within *any* conscious system. However, such phenomena clearly play a key role in the human experience of being conscious, and the empirical correlates of this experience. Human consciousness is not just a generic phenomenon of attention-focusing or what-not; it clearly involves multiple important characteristics common to the broader phenomenon of consciousness, but it is a specific process involving a specific architecture that evolved for specific reasons.

Numerous subtleties arise here, such as the re-use of these specialized structures for more general purposes. E.g. it seems that the phonological loop can be used for handling abstract mathematical knowledge as well as ordinary speech [3]; and that the visuospatial sketchpad can be used for abstract visual or partly-visual representation of abstract conceptual relationships [37]. As commonly occurs in biological systems, mechanisms that evolved for one purpose may then be adaptively deployed for others.

Baddeley's model of working memory also posits a "central executive" that coordinates the operation of the phonological, visuospatial and episodic components. This is somewhat controversial as it has some appearance of being a "homunculus" type mechanism. However, the broad notion of a central executive function can be fulfilled in many ways, and not necessarily by a physically localized or operationally isolated subsystem. The Global Workspace Theory and LIDA cognitive model provide an example of how Baddeley's central executive function can be modeled as a distributed, system-wide dynamical process rather than an isolated, homuncular module.

2.4 Dynamic Global Workspace Theory

Perhaps the most comprehensive model of the *cognitive correlates of consciousness* in the human mind, is the Global Workspace Theory (GWT) developed by Bernard Baars and his colleagues [7, 8]. The LIDA cognitive architecture, developed by Stan Franklin and his colleagues, is a cognitive model and AI architecture and system that directly incorporates the key aspects of GWT, along with other AI and cognitive science ideas.

A global workspace (GW) is broadly defined as "a functional hub of binding and propagation in a population of loosely coupled signaling elements." Intuitively and experientially speaking, the GW is the "inner domain in which we

[3] Hadamard reports that some mathematicians, e.g. the great George Polya, say they think about mathematical concepts in terms of grunts and groans [32]

can rehearse telephone numbers to ourselves or in which we carry on the narrative of our lives. It is usually thought to include inner speech and visual imagery." [7]. Conscious experience in humans and other similar animals is viewed as associated with GW functions.

There are reasonably substantiated hypotheses regarding the neural underpinnings of the GWT in humans and similar animals The cortico-thalamic (C-T) core is believed to underlie conscious experience and associated cognitive functions. However, the GW is not to be identified with any specific anatomical hub within the C-T core. Rather, the GW is to be thought of as spanning multiple anatomical hubs, and constituting "dynamic capacity for binding and propagation of neural signals over multiple task-related networks, a kind of neuronal cloud computing" [8]. The hypothesis is that

> [C]onscious contents can arise in any region of the C-T core when multiple input streams settle on a winner-take-all equilibrium. The resulting conscious gestalt may ignite an any-to-many broadcast, lasting $100 - 200ms$, and trigger widespread adaptation in previously established networks. To account for the great range of conscious contents over time, the theory suggests an open repertoire of binding coalitions that can broadcast via theta/gamma or alpha/gamma phase coupling, like radio channels competing for a narrow frequency band. Conscious moments are thought to hold only 14 unrelated items; this small focal capacity may be the biological price to pay for global access.

To phrase the core dynamic of GWT in neural network terms, one may describe the global broadcast of the contents of consciousness as follows: an active cell assembly (which would correspond to a coalition in the LIDA model; see below) "wins out over the competition and ignites, gathering momentum, and spreads out to include the whole of the Edelman and Tononi thalamocortical core." [19]

The GWT is a cognitive and cognitive-neuroscience rather than computational theory, but it has served as the inspiration for various AI system designs to various degrees. The closest relationship is with LIDA (Learning Intelligent Distribution Agent), an ambitious computational cognitive architecture created by Stan Franklin and his colleagues, inspired by direct collaboration with Baars, which attempts to provide a working model of a broad spectrum of cognition in humans and other animals, from low-level perception/action to high-level reasoning.

Inspired by GWT, LIDA is founded on two core hypotheses:

- Much of human cognition functions by means of frequently iterated (10 Hz) interactions, called cognitive cycles, between conscious contents, the various memory systems and action selection.
- These cognitive cycles, serve as the atoms of cognition of which higher-level cognitive processes are composed.

Spanning both sides of the symbolic/subsymbolic dichotomy, LIDA is a hybrid architecture in that it employs a variety of computational mechanisms, chosen

for their psychological plausibility and practical feasibility. The focus on action selection is carefully reasoned based on Stan Franklin's associated AI theories [3].

Along with the GWT, LIDA incorporates a broad spectrum of ideas from the cognitive science literature, including models of the particular architecture of human working memory as roughly indicated above. It incorporates specific components corresponding to particular substructures within the working memory, e.g. Transient Episodic Memory, Sensory Memory, Perceptual Associative Memory, etc.

More specifically [4] , the LIDA cognitive cycle can be subdivided into three phases, the understanding phase, the attention (consciousness) phase, and the action selection and learning phase. Beginning the understanding phase, incoming stimuli activate low-level feature detectors in Sensory Memory. The output engages Perceptual Associative Memory where higher-level feature detectors feed in to more abstract entities such as objects, categories, actions, events, etc. The resulting percept moves to the Workspace where it cues both Transient Episodic Memory and Declarative Memory producing local associations. These local associations are combined with the percept to generate a current situational model; the agents understanding of what is going on right now. The attention phase begins with the forming of coalitions of the most salient portions of the current situational model, which then compete for attention, that is a place in the current conscious contents. These conscious contents are then broadcast globally (following the core dynamic proposed by Global Workspace Theory), initiating the learning and action selection phase. New entities and associations, and the reinforcement of old ones, occur as the conscious broadcast reaches the various forms of memory, perceptual, episodic and procedural. In parallel with all this learning, and using the conscious contents, possible action schemes are instantiated from Procedural Memory and sent to Action Selection, where they compete to be the behavior selected for this cognitive cycle. The selected behavior triggers Sensory-Motor Memory to produce a suitable algorithm for its execution, which completes the cognitive cycle.

In GWT terms, LIDA may be understood to incorporate (among other things) a detailed theory of how the focusing of attention works, and hence of how the process of consciousness works. The input to the consciousness process is understood to include stimuli from both the external and internal environments. Along with various sorts of memory, the internal environment is viewed as including a "preconscious workspace" in external stimuli and internal constructs are understood with the help of recall from various memories. Input to the consciousness process is viewed as coming from this preconscious workspace; the consciousness process processes these inputs, producing dynamic "conscious contents." The contents of consciousness are then broadcast according to the core GWT dynamic, causing a global impact on the mind network, thus impacting the preconscious workspace and completing the loop between preconscious-workspace and conscious processes.

[4]This paragraph is paraphrased from [45]

The consciousness process as depicted by GWT and LIDA is clearly a complex nonlinear dynamic, susceptible to various subtle emergent self-organizational phenomena, that are not well characterized at present. In dynamical systems terms, what Franklin and the the LIDA group would tend to view as a "pattern of conscious contents", I would describe in more detail as a "probabilistically approximately invariant subspace of the set of possible states of the dynamical system comprised of the consciousness process and the contents of consciousness". Crudely, one could speak of an "attractor" instead of a "probabilistically approximately invariant subspace" – but in actuality, if one did real mathematical modeling of these systems,one would likely not find something so deterministic as an attractor according to the standard definitions of dynamical systems theory [17]. What I am here calling a "probabilistically approximately invariant subspace" has sometimes been called a "persistent transient." Biological, social and psychological systems are full of these sorts of phenomena, even though mathematical dynamical systems theory deals much more with simpler cases like attractors and invariant measures.

2.5 Consciousness as a Nonlinear-Dynamical Process

The perspective of consciousness as a complex, nonlinear-dynamical process bears further elaboration. Sally Goerner and Leslie Allan Combs, in a concise, elegant article from 1998 [23], argued in favor of a process perspective on consciousness, from a more experiential view:

> Consciousness always has an object. In other words, it is always about something. We are not just conscious, we are conscious of the taste of food, the smell of the sea, a tooth ache. We are conscious of joy, of boredom, of the meaning of words on the page in front of us, of the sound of music playing in the next room, of our own thoughts, of memories. The point is that virtually all experience is experience of something. ... Let us go one step further and note that events which lead to increased complexity in conscious experience also must in their own way lead to increased complexity in brain processes. To look at a tree in bloom presents the mind with a picture of pleasing complexity. Likewise, we cannot doubt that the brain is treated to a similar upgrade in complexity, and that electrochemical changes there support our experience of pleasure as well. ... In the above example it is apparent that looking at a tree in bloom in-forms both the brain and the mind, or conscious experience, in a way that increases their complexity. Their information level has been enlarged. Here we see the interchangability of experience and information. Consciousness would seem to be intimately involved with the informing of the brain and mind by objects of attention. Moreover, on the brain side we see that the complexification associated with a conscious experience also involves an increase in energy, though this may be only be of a small amount. Here again the connection with neg-entropy comes into play as a decrease in disorganization and an increase in order.

There is a clear connection between these ideas and the LIDA model, in which a large role is played by the contents of consciousness, the object of

attention at the moment. The LIDA model lays out out a hypothetical, but plausible, mechanism for the self-organization to which Combs refers, based on the idea that non-linear dynamics can provide bridge between high-level, conceptual models of mind like LIDA and underlying neural mechanisms.

Combs goes further and states that

> Bringing the above ideas together, we suggest that each state of consciousness, mood, or frame of mind, represents a unique and coherent–minimal energy–fit for the information streams represented by the many psychological processes which comprise it, producing a stable pattern or gestalt. Further, the stability of the pattern arises from its autopoietic tendency to self-organize.

This relates to the notion that consciousness has to do with some subnetwork of the brain settling into, if not an attractor in the strict sense of dynamical systems theory, then at least a persistent transient associated with some particular basin in state space. This related to neuroscientist Walter Freeman's perspective on neurodynamics as dominated by "strange attractors with wings" [20, 35].

2.6 Consciousness and Attention

The relationship between consciousness and attention is universally recognized as a close one, but has been articulated in a great variety of different ways. For instance, Baars [7] summarizes attention as

> **Attention**. In GW theory, the control of access to consciousness by reference to long-term or recent goals. Attention may be voluntary or automatic. See also Prioritizing Function.

Thus, he defines attention in terms of consciousness in a particular way. While this seems perfectly sensible, I suggest that it may be useful to define attention separately from consciousness, so as to be able to more clearly explore the relationship between the two.

A careful analysis allows us to decompose the notion of attention into at least two subconcepts:

- **Resource Attention** – attention as regarding the allocation of resources. One can define the attention a system pays to some entity E, relative to an observer (which may be the entity itself), as the percentage of the systems resources that are devoted to E or other entities related to E (where relatedness to E is judged by the given observer). Of course one also has to specify whether one is concerned with space or time resources. Generally in the case of the brain, one is thinking about processing-time as a resource, and also about short-term memory buffers as a resource (but not about long-term memory as a resource; when we day a person is focusing their attention on a certain entity, we dont assume that entity is dominating their long-term memory).

- **Information Attention** – one can define the attention a system pays to some entity E as the percentage of the information content observable in the system (over a certain interval of time) that concerns E or other entities related to E. In this context one would need to carefully choose the right definition of information content, so as to exclude information that is largely- dormant in LTM. It seems one wishes to look at information that is detectable from the internal dynamics of the system during the given interval of time (which is similar to what is done in Tononi's information integration measure).

Conceptualizing things as such, "attention can be separated from "consciousness, so that the alignment of consciousness with attention becomes an observation about certain kinds of cognitive architectures, rather than a definition.

Attention, obviously, is a very broadly applicable concept. I would hypothesize that the emergence of some sort of attentional focusing mechanism is almost inevitable in any intelligence for which (to use a schematic equation to convey a qualitative notion) $\frac{R}{U}$ is sufficiently small, where

- R = the system's available compute resources
- U = the system's urgent need for real-time action selection, where each action depends to differing degrees on differing items of data stored in system memory

Intuitively, the combination of these two factors means that the system will need to focus attention on some memory items more than others in order to survive, which is going to cause the emergence of something like a focus of attention.

Rather than defining attention in terms of consciousness, one can say: It is a fact about the human cognitive architecture (but not necessarily about all possible cognitive architectures) that attentional focus and "reported as conscious" events tend to be aligned. In other words,

- When resource attention, information attention and consciousness are aligned, then one has a case where a substantial portion of a systems elements are mutually entrained in the process of focusing a substantial portion of the systems energy (resources) and information-processing on some entity E.
- This kind of alignment is important to the qualia of ordinary human conscious experience. Without this kind of alignment, one would have a different kind of phenomenon, that would subjectively feel quite different.

2.7 States of Consciousness

Another well-documented aspect of human consciousness, important for studying consciousness in humans, animals and engineered systems, is that it comes

in different "states." [5] The foundational work here is Charles Tart's book *State of Consciousness* [47], but the concept goes back much further; e.g. Tart quotes William James [34], who said

> Our ordinary waking consciousness... is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different. We may go through life without suspecting their existence; but apply the requisite stimulus, and at a touch they are all there in all their completeness, definite types of mentality which probably somewhere have their field of application and adaptation. No account of the universe in its totality can be final which leaves these other forms of consciousness quite disregarded. How to regard them is the question – for they are so discontinuous with ordinary consciousness.

Tart defines a Discrete State of Consciousness or d-SoC as follows:

> We can define a d-SoC for a given individual as a unique configuration or system of psychological structures or subsystems. The structures vary in the way they process information, or cope, or affect experiences within varying environments. The structures operative within a d-SoC make up a system where the operation of the parts, the psychological structures, interact with each other and stabilize each other's functioning by means of feedback control, so that the system, the d-SoC, maintains its overall pattern of functioning in spite of changes in the environment. Thus, the individual parts of the system may vary, but the overall, general configuration of the system remains recognizably the same.

What are the neural or cognitive correlates of the "state of consciousness" phenomenon? One approach to conceptualizing the issus is as follows. If we think of consciousness as a process, it may make sense to think of it as a *parametrized* process. One can then talk about two levels of dynamics

- dynamics involving changes in the *contents* of consciousness
- dynamics involving changes in the *parameter values* of the parameters of the consciousness process

I would propose that different states of consciousness (e.g. stoned, tripping, dreaming, enraged, ordinary-waking) may correspond to different regions of the parameter-value-vector space of the consciousness process.

The parameters of any complex cognitive system are going to have subtle interdependencies, in terms of the influence they have on system behavior; so that not every possible collection of parameter settings will lead to coherent, meaningful system behavior. But neither will there be one unique set of *narrow*

[5] As Leslie Allan Combs pointed out to me, philosophers of mind anyway tend to use the term state to refer to nearly any mental condition, such as anger, sleepiness, being bewildered, and the like; whereas, psychologists and scholars of consciousness tend to use the term "state to refer to broader experiential landscapes, such as waking, dreaming, being "stoned", being hypnotized, "losing one's head" in a complete uncontrollable rage, tripping on LSD, etc. Here I will use the term in the latter sense.

ranges for each parameter that corresponds to successful system functioning. Rather, there are multiple collections of *narrow ranges for each parameter*. In the case of parameters directly related to the consciousness process, such collections may correspond to different states of consciousness.

Of course, different parameter vectors for the consciousness process will tend to lead to different patterns in consciousness contents ... e.g. one is unlikely to do one's taxes while tripping on LSD, etc. Thus, as well as a set of parameter values, each state of consciousness will correspond to a certain subspace of the space of possible states of the consciousness process. States of consciousness tend to have a certain momentum to them, meaning that they correspond to "probabilistically almost invariant subspaces" of state space.

Note that in dynamical systems theory "state" is generally used to refer to an instantaneous condition of a system; whereas "states of consciousness" are not that, they are classes of instantaneous states that stand in a certain relationship to each other relative to the underlying process. The use of "state" in "state of consciousness" is more analogous to the term "state of matter" as used to refer to solid, gas, liquid, plasma, etc. In the case of states of matter, the underlying physical processes are the same as a substance moves between different states (based on changes in underlying parameters such as temperature), but of course the dynamical properties of the system may change based on external and internal conditions, in spite of there being a consistent underlying process...

As an example, recent research suggests that psychedelic states are higher-entropy states than ordinary waking consciousness [12]. In an AI system this sort of higher-entropy state could potentially be induced by tweaking parameters of the attention allocation subsystem so that attention is spread more diffusively across the system's knowledge base, rather than being tightly concentrated among the top fraction of most "important" knowledge items at a given point in time. The result of this parameter tweaking would be the system settling into states involving cognitive processes not relying on tight attentional focus on any one topic, but relying more on lateral thinking, perceptual metaphor, and other cognitive correlates of diffused attention.

2.8 Tononi's Integrated Information Measure

Giulio Tononi [48] has outlined a theory of consciousness founded on the following two conceptual principles:

- Every conscious state or moment contains a massive amount of information.
- All the information that an agent gleans from conscious states is highly, and innately, integrated into the agent's mind

Based on these ideas and subsequent analyses, Tononi has proposed a quantitative measure of consciousness called the "Integrated Information" or Φ. Roughly

speaking, Φ attempts to measure the degree to which there is a lot of information generated among the parts of a system as opposed to within them. [6]

Tononi is to be congratulated for making a specific formal hypothesis regarding the nature and measurement of consciousness; and unsurprisingly, his hypothesis has attracted significant critical attention alongside significant enthusiasm. Computer scientist Scott Aaronson presented a detailed argument showing that, according to Tononi's mathematical measure, certain relatively simple mathematical constructs would be assessed as having a very high degree of consciousness [2]. A similar point was made more simply by Eric Schwetzgebel, who argued that according to Tononi's Φ measure, the United States would likely be assessed as far more conscious than any human [44].

Tononi's counter-argument to Aaronson basically argues that the Φ measure was never intended to be applied to arbitrary mathematical constructs, but rather to be applied in the context of organisms engaging with the world [7]. This is conceptually reasonable, but dramatically reduces the value of Φ as a rigorous measure of consciousness. If Φ should only be applied to certain kinds of systems, and the class of applicable systems is defined only informally and qualitatively, then do we really have a rigorous quantitative measure of consciousness? It would seem that, to have a rigorous measure, one would then need a formal way to measure the applicability of the Φ measure.

This might seem a nit-picky, pedantic point, but I believe it is more than that. Alternate approaches to understanding consciousness, like the ideas of Baars, Franklin, Combs and Tart mentioned above, are focused largely on understanding what it means for an organism to intelligently, cognitively engage with the world. As I will elaborate below, applying Tononi's ideas in the context of a theory like GWT yields a more complex and subtle understanding of conscious information processing, which does not attribute consciousness to simple mathematical constructs – but might perhaps attribute some degree of consciousness to the United States. This sort of nuanced, systems-oriented view of consciousness lacks the mathematical elegance and unidimensional clarity of Tononi's theory as originally outlined, and seems to align reasonably with Tononi's overall intentions.

2.9 Self-Modeling, Reflection and Self-Awareness

One of the key aspects of human consciousness is its reflective, recursive, self-awareness. This is not always present – in a meditative state the human mind can in a sense transcend self-awareness [5]; and in a "flow" state, the human mind can be so completely immersed in its task that it "forgets itself" [15].

[6] While Tononi's Φ is a reasonable measure of information integration, it is worth noting that there are many other ways to quantify the concept of "integrated information"; my own work in this area from two decades ago outlined similar definitions using algorithmic information and related ideas rather than Shannon information [26, 24]. Algorithmic information is not practical to compute exactly based on real-world data; but neither is Tononi's Φ for any complex system.

[7] See [1] for Tononi's counter-argument and Aaronson's detailed response to it

But much of the time, explicit self-awareness is a prominent aspect of human consciousness.

Thomas Metzinger [40] has outlined a detailed, cross-disciplinary "self-model theory of subjectivity," centered on the concept of a "phenomenal self-model (PSM). A "self-model" is understood as a dynamic, ongoing process by which a portion of an organism's cognitive system comes to reflect and predict the organism itself; and the PSM is, basically, conceived as the "conscious" portion of an organism's self-model. What is meant by "conscious" here is a set of properties, such as availability for introspective attention and for selective, flexible motor control, integration into the organism's internal representation of time, and ongoing dynamic integration into an overall model of the organism and its world.

Metzinger [41] distinguishes several levels of embodiment in cognitive systems:

- **first-order**: cognitive properties emerging within perceiving, acting bodies as they interact with their environment
- **second-order**: when a cognitive system represents its own embodiment internally, and uses this representation to help choose and guide actions
- **third-order**: when a cognitive system's representation of its own embodiment becomes part of the system's "conscious contents"

It seems intuitively clear that ordinary waking human consciousness involves what he calls third-order embodiment; this is a key part of the ordinary human self-model.

In [28] I dig deeper into the possible structure of the PSM, and propose to model the reflective aspect of human consciousness in terms of hypersets, mathematical objects that extend ordinary sets via their capability to recursively contain themselves as elements. There, the following recursive definitions are given:

- "S is reflectively conscious of X" is defined as: The declarative content that *"S is reflectively conscious of X" correlates with "X is a pattern in S"*
- "S wills X" is defined as: The declarative content that *"S wills X" causally implies "S does X"*
- "X is part of S's self" is defined as: The declarative content that *"X is a part of S's self" correlates with "X is a persistent pattern in S over time"*

These are posited as ideal forms that are approximated by the recursive forms in actual human mind/brains. These definitions imply an interesting symmetry to the relationship between self and awareness, namely: Self is to long-term memory as reflective awareness is to short-term memory. These recursive patterns, it is hypothesized, occupy a significant amount of energetic and informational

attention in human minds. They often occupy significant attention within the Global Workspace; and it seems intuitively clear that the brain regions embodying these recursions would display significant integrated information.

3 Toward a Unified Model of Human and Human-Like Consciousness

What, then, are the critical factors characterizing the consciousness of human beings, and likely to characterize the consciousness of AI systems with roughly human-like cognitive architectures? Based on the literature and concepts reviewed above, an integrative understanding emerges fairly clearly. When a human-like system has the experience of being conscious of some entity X, then the system should manifest:

1. **Dynamical representation**: the entity X should correspond to a distributed, dynamic pattern of activity spanning a portion of the system (a "probabilistically invariant subspace of the system's state space"). Note that X may also correspond to a localized representation, e.g. a concept neuron in the human brain [42]

2. **Focusing of energetic resources**: the entity X should be the subject of a high degree of energetic attentional focusing

3. **Focusing of informational resources**: X should also be the subject of a high degree of informational attentional focusing

4. **Global Workspace dynamics**: X should be the subject of GWT style broadcasting throughout the various portions of the system's active knowledge store, including those portions with medium or low degrees of current activity. The GW "functional hub" doing the broadcasting is the focus of energetic and informational energy

5. **Integrated Information**: the information observable in the system, and associated with X, should display a high level of information integration

6. **Correlation of attentional focus with self-modeling**: X should be associated with the system's "self-model", via associations that may have a high or medium level of conscious access, but not generally a low level

These I will call **six key factors of human-like consciousness**. I do not claim that they are the *only* important aspects; but I do posit that they are among the most important aspects.

The first five factors, I suggest, are relevant regardless of the state of consciousness – but may have different levels of importance in different states of consciousness. On the other hand, the sixth factor may play a minimal role in some states of consciousness, e.g. "non-symbolic" states as experienced by meditators, advanced spiritual practitioners and others [38]. Relative to the ordinary

waking state of consciousness, psychedelic states [47] and flow states [15],would (qualitatively speaking) seem to involve less of a role for the self-model, as well as less concentrated attentional focusing.

3.1 Measuring Human-Like Consciousness Multifactorially

How then can one measure the degree of consciousness possessed by a system at a certain point in time, or the degree of conscious access that a system is giving to a certain entity during a certain interval of time? One reasonably tractable way to phrase this question, I suggest, is: **How can one measure the degree of human-like conscious access that a system gives to a certain entity during a certain interval of time?**

To formalize the degree to which a system S gives human-like **conscious access** to an entity X, as a first approximation one could quantify the six factors listed above: energetic attentional focusing, informational attentional focusing, GW broadcasting, information integration, and association with self. One would then quantify conscious access as a weighted combination of these factors, with the weighting being state of consciousness dependent.The formulation of precise mathematical measures of each of these six factors would not be extremely difficult, but would require detailed analysis and would increase the length of this paper by a small integer multiple. So these particularities will be left for sequel papers.

Next, given a definition of human-like conscious access, one can conceive

- the *degree of human-like consciousness* of a system as the sum over all entities X in the system, of the degree to which the system gives X conscious access
- the *ratio of human-like consciousness* of a system as the *average* over all entities X in the system, of the degree to which the system gives X conscious access

This characterization of human-like consciousness is admittedly messy, and in more than one way. These six factors are all important, but it's quite possible that a handful of further factors could usefully be added to the list. Furthermore, each of these factors could be quantified in multiple ways – as in the example of Tononi's Information Integration measure, which is only one among a large number of sensible-looking mathematical formulas for capturing the conceptual notion of information integration.

This messiness, however, strikes me as inevitable – i.e. it is simply part of the territory, which any reasonable map must reflect. Consciousness-in-general may be elementally simple in some sense, but human consciousness is a specific cognitive construct that evolved to serve the needs of specific sorts of organisms. AI systems may in principle display quite different varieties of consciousness; but if an AI system is going to display closely human-like intelligence, it will almost surely need to manifest closely human-like consciousness as well. The

processing and memory dynamics that produce human-like consciousness are integral to the production of human-like intelligence.

3.2 Measuring Consciousness in the Human Brain

It is an appealing idea to use neurophysiological measurements to gauge the degree of consciousness of a human brain, as it passes through various states and experiences. Given an appropriate measure, the degree of consciousness of different parts of the human brain could also be gauged, providing a new perspective on the investigation of the neural correlates of conscious experience.

Research has been done regarding the computation of certain (mathematically crude but perhaps pragmatically valuable) estimates of the Integrated Information of the brain [10]. In a similar vein, one could measure the informational attention focusing of the brain during a certain period of time. Energetic attention focusing should be more straightforward to measure, as standard tools such as fMRI already give a view into the brain's energy expenditure.

Measurement of the degree to which the brain's focus of attention is represented as a dynamical pattern, or the prevalence of GW dynamics in the brain, on the other hand, would seem to require neuroimaging with simultaneous spatial and temporal resolution going beyond what current technology provides. One would need to be able to measure the broadcasting happening within a single "conscious moment" between different regions of the brain – say, on the time scale of milliseconds, and the spatial scale of a cortical column. Such neuroimaging tools are likely coming in the future and will have many exciting applications beyond the measurement of consciousness. Perhaps analysis of the data provided by such tools will enable modeling of the way the human brain builds its self-model, which will allow measurement of the association between entities in the GW and the self-model as well.

3.3 Human-Like Consciousness in LIDA and OpenCog

As compared to measuring consciousness in the human brain, the measurement of human-like consciousness in AI systems is a relatively straightforward matter. Issues of instrumentation are reasonably rapidly resolvable, so one is left only with the problem of formalizing the relevant aspects of consciousness in a computationally tractable way. This problem is far from trivial, since mining patterns from the dynamics of a rapidly changing large-scale software system is highly resource intensive. For instance, accurately computing the integrated information according to Tononi's definition seems likely to be an NP-hard problem [2].

We have seen above one example of an AGI (Artificial General Intelligence) system engineered to manifest human-like consciousness: the LIDA system is built centrally around the Global Workspace theory, so that a properly functioning LIDA system automatically incorporates some of the six aspects highlighted here. As well as having a GW, the dynamics of LIDA are designed to focus energetic and informational attention on the contents of the GW. The feedback

between the central workspace and the rest of LIDA is intended to give rise to nonlinear dynamics that will form persistent dynamical patterns occupying the focus of attention. The different components of the LIDA system are intended to work together in a tightly coupled nonlinear way, which should in most cases lead to a high degree of integrated information among the active knowledge in the various components. LIDA theory does not focus on the emergence of self-modeling, however in a LIDA system put in situations where self-modeling was the simplest clearly effective strategy for goal achievement, it could be expected that a reasonably thorough self-model would emerge and would often occupy a significant fraction of the workspace.

The OpenCog AGI architecture [29, 30] also manifests the six aspects mentioned above, in its own way. OpenCog's main memory store consists of a weighted labeled hypergraph whose nodes and links are called Atoms; and each Atom is labeled with ShortTermImportance (STI) and LongTermImportance (LTI) numbers, the former governing the Atom's frequency of occurrence in cognitive processes, the latter governing the Atom's retention in RAM. The set of Atoms with STI above a certain boundary level is called the AttentionalFocus (AF). A current development initiative focuses on adding specialized structures corresponding to the phonological loop, visuospatial sketchpad and episodic memory buffer to the system, to work closely with the AttentionalFocus.

Roughly, speaking, in OpenCog, the AttentionalFocus corresponds to the Global Workspace. The dynamics of STI spreading through the Atomspace can be viewed as an implementation of the GW theory notion of GW broadcasting. Energetic and informational focusing on the AF occurs because the system's various cognitive processes focus their attention preferentially on discovering new things about the Atoms in the AF, and building new Atoms via combining the ones in the AF. While some entities are represented by specific Atoms in the style of a traditional semantic network, the nonlinear dynamics of STI spreading means that entities are also represented by distributed patterns of activity (this dual representation has been referred to as "glocal memory" [31]). As in the case of LIDA, a self-model is not built into the system, but is intended to emerge naturally as a result of the system's behavior in the context of environments and goals that benefit substantially from self-modeling.

Information integration, finally, is closely related to the "cognitive synergy" principle that lies at the heart of OpenCog theory. The key notion here is that the various cognitive processes acting on the Atomspace should interoperate at a deep level, helping each other to overcome the combinatorial explosions they confront. Conceptually, this seems to imply that the interim data produced by the different cognitive processes should display a high degree of integrated information.

Qualitatively, we thus see that both LIDA and OpenCog are design in a way that is in principle amenable to displaying the six key aspects of human-like consciousness we have highlighted here. The same would certainly be true of a number of other cognitive architectures aimed at human-like AGI (see e.g. the review [18, 43]). The extent to which human-like consciousness is actually manifested by running instances of these systems, is dependent on the degree to

which these instances actually implement the cognitive architectures in question, and the extent to which these architectures operate as the underlying theories predict. Currently, LIDA, OpenCog and other architectures aimed at human-like cognition are still in relatively early research phases.

3.4 Human-Like Consciousness in the Global Brain

One may also apply these ideas is to the notion of an emergent "Global Brain" – an intelligence arising from collective dynamics in the global network of humans, computers and communication devices [25, 33]. Many observers have argued that the Internet and related networks already display some form of intelligence; and some have speculated that as related technologies progress, the global communication/computing/social network will achieve more and more of the aspects of an autonomous, individual mind. This line of thinking naturally gives rise to the question of whether, or in what sense, a Global Brain could be conscious. More particularly, from the perspective pursued here, one well-posed question is whether, and to what degree, a Global Brain (GB) – today's or a future descendant – might have *human-like consciousness*.

The clearest candidate for the attentional focus of today's GB would be the distributed, active data stores of major Internet companies. These occupy a decent fraction of the compute resources available, and get preferential treatment in the infrastructure of Internet service providers, so that their information is served fastest. Information as well as energy are substantially focused on these data stores – which, like the GW in the human brain, are functional hubs rather than single physical hubs (as they generally span server farms in multiple physical locations). The data stores of Google, Facebook, Microsoft, LinkedIn, Twitter and the like broadcast information widely throughout the world's population of humans and computers, and then receive feedback which guides their next information broadcasts. A "self" in the precise sense of human psychology is lacking, but several of these companies (e.g. the major search engines) do have internal models of large portions of the Internet, which they model in various ways.

Information integration, finally, is the goal of the major analytics efforts undertaken by so many large Internet companies recently; it is the goal that underlies the recent rise of applied machine learning in the Internet and social network context (e.g. in early 2014: Google's acquisition of Deep Mind; Facebook's work on face recognition and their founder's investment in Vicarious Systems; etc.). The goal of these machine learning projects is precisely to learn abstract patterns that are implicit when you put a huge amount of data together in one (distributed) place, that are not so readily observable in smaller troves of data.

At the present time, the GW dynamics of the Internet is fairly different from that of a human brain. One major difference is that the rate at which the GW impacts the periphery is much slower than in the human brain (measured proportionally to internal dynamics of GW or the periphery). The current GB's GW does sophisticated modeling of the whole GB, but uses the results of this

modeling only relatively slowly and weakly; whereas, the human brain's GW does a type of broadcasting that much more heavy-handedly drives the overall dynamics of the brain. This difference in dynamics affects self-modeling as well; it means that the self-models maintained by the major Internet companies tend to focus on static relationships rather than dynamics.

The Internet, however, is rapidly evolving; and there are developments underway that seem likely to bring GB dynamics closer to that of the human brain. Once AI technology advances to the point that the descendants of current personal assistants like Siri, Google Now and Cortana – as well as personal assistants in mobile robot form – can interact with a modicum of general intelligence, then the the GB will have a periphery capable of sensitively and frequently exchanging high-information feedback with its GW. This will require the GW to maintain a more complex self-model focusing more on dynamics, and will, at a high level, make the GB more human brain like. It will also, no doubt, introduce various subtleties without parallels in the human brain.

Quantifying the six factors of consciousness mentioned above in an Internet context, would give a way of measuring the degree of human-like consciousness of the global brain, and tracking the various features of this consciousness as it emerges.

Acknowledgements

This paper owes a huge amount to conversations with a number of people, including most notably (alphabetically) Bernard Baars, Leslie Allan Combs, Stan Franklin, Zar Goertzel, Cosmo Harrigan, Jim Rutt and Gino Yu.

References

[1] Scott Aaronson. Giulio tononi and me: A phi-nal exchange. 2014. `http://www.scottaaronson.com/blog/?p=1823`.

[2] Scott Aaronson. Why i am not an integrated information theorist. 2014. `http://www.scottaaronson.com/blog/?p=1799`.

[3] Negatu Aregahegn and Stan Franklin. An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly 2 (3)*, 2002.

[4] Susan Armstrong. For love of matter: A contemporary panpsychism. *Environmental Ethics 28*, pages 99–102, 2006.

[5] James Austin. *Zen and the Brain*. MIT Press, 1999.

[6] Bernard Baars. personal communication.

[7] Bernard Baars. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, 1997.

[8] Bernard Baars, Stan Franklin, and Thomas Ramsoy. Global workspace dynamics: Cortical binding and propagation enables conscious contents. *Frontiers of Psychology (4)*, 2013.

[9] A D Baddeley. *Working memory, thought and action*. Oxford, 2007.

[10] A B Barrett and A K Seth. Practical measures of integrated information for time series data. *PLoS Computational Biology*, 2011.

[11] Etienne Benson. Intelligence across cultures. *Monitor on Psychology 34 (2)*, 2003.

[12] R Carhart-Harris, R Leech, P J Hellyer, M Shanahan, A Feilding, E Tagliazucchi, D R Chialvo, and D Nutt. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers of Human Neuroscience*, 2014.

[13] David Chalmers. *The Conscious Mind*. Oxford University Press, 1997.

[14] D S Clarke. *Panpsychism: Past and Recent Selected Readings*. SUNY Press, 2004.

[15] Mihaly Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper and Row, 1990.

[16] Daniel Dennett. *Consciousness Explained*. Penguin, 1993.

[17] Warren Devaney. *Chaotic Dynamical Systems*. Westview Press, 2003.

[18] Wlodzislaw Duch, Richard Oentaryo, and Michel Pasquier. Cognitive architectures: Where do we go from here? *Proc. of the Second Conf. on AGI*, 2008.

[19] Stan Franklin. personal communication.

[20] Walter Freeman. *Societies of Brains*. Erlbaum, 1995.

[21] Raphael Gaillard, Stanislas Dehaene, Claude Adam, Stephane Clemenceau, Dominique Hasboun, Michel Baulac, Laurent Cohen, and Lionel Naccache. Converging intracranial markers of conscious access. *PLoS Biology*, 2009.

[22] H Gardner. *Intelligence reframed: Multiple intelligences for the 21st century*. Basic, 1999.

[23] Sally Goerner and Allan Combs. Consciousness as a self-organizing process. *Biosystems 46*, pages 123–127, 1998.

[24] Ben Goertzel. *The Evolving Mind*. Plenum, 1993.

[25] Ben Goertzel. *Creating Internet Intelligence*. Plenum Press, 2001.

[26] Ben Goertzel. *The Hidden Pattern*. Brown Walker, 2006.

[27] Ben Goertzel. Toward a formal definition of real-world general intelligence. 2010.

[28] Ben Goertzel. Hyperset models of self, will and reflective consciousness. *International Journal of Machine Consciousness 3*, 2011.

[29] Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy*. Springer: Atlantis Thinking Machines, 2013.

[30] Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI*. Springer: Atlantis Thinking Machines, 2013.

[31] Ben Goertzel, Joel Pitt, Matthew Ikle, Cassio Pennachin, and Rui Liu. Glocal memory: a design principle for artificial brains and minds. *Neurocomputing*, April 2010.

[32] Jacques Hadamard. *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press, 1945.

[33] F. Heylighen. *The Global Superorganism: an evolutionary-cybernetic model of the emerging network society*. Social Evolution and History 6-1, 2007.

[34] William James. *The Varieties of Religious Experience*. CreateSpace, 2009 (1902).

[35] R. Kozma, M. Puljic, and W. Freeman. Thermodynamic model of criticality in the cortex based on eeg/ecog data. In *Criticality in Neural Systems , Ed. by Plenz, D. and Niebur, E.* Wiley, 2013.

[36] Shane Legg and Marcus Hutter. A definition of machine intelligence. *Minds and Machines*, 17, 2007.

[37] R H Logie. *Visual-spatial working memory*. Lawrence Erlbaum, 1995.

[38] Jeffery Martin. Clusters of individual experiences form a continuum of persistent non-symbolic experience in adults. 2012. `http://nonsymbolic.org/PNSE-Article.pdf`.

[39] Thomas Metzinger. *Neural Correlates of Consciousness*. Bradford, 2000.

[40] Thomas Metzinger. *Being No One*. Bradford, 2004.

[41] Thomas Metzinger. Self-models. *Scholarpedia*, 2007.

[42] R. Quian Quiroga. Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13:587–597, 2012.

[43] Alexei V. Samsonovich. Toward a unified catalog of implemented cognitive architectures. In *BICA*, pages 195–244, 2010.

[44] Eric Schwetzgebel. Why tononi should think that the united states is conscious. 2014. http://schwitzsplinters.blogspot.hk/2012/03/why-tononi-should-think-that-united.html.

[45] Ryan McCall Snaider, Javier and Stan Franklin. The lida framework as a general tool for agi. *Proceedings of AGI-11*, 2011.

[46] Galen Strawson. Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies 13*, pages 10–11, 2006.

[47] Charles Tart. *States of Consciousness*. iUniverse, 2003.

[48] Giulio Tononi. Consciousness as integrated information: a provisional manifesto. *Biological Bulletin 215 (3)*, 2008.