

# **Artificial Consciousness, the search for what makes us human**

**WA de Landgraaf**

Stdnr: 1256033

Course: AI 2001, FAAI

Group A, docent Radu Serban

Department of Artificial Intelligence

Vrije Universiteit

De Boelelaan 1081a

1081 HV Amsterdam

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation. This document can be obtained via <http://am.xs4all.nl>

## **Abstract:**

The Society of Mind is an 'implementation' of the philosophical theory of functionalism. Even so, the search for consciousness continues to go on in areas of psychology and philosophy, providing both answers and new questions to be solved. AI as a whole should stop seeing each facet as different competing implementations and should start combining symbolism and connectionism, which would give us the virtues of both and offer more diverse implementations of the theory of functionalism. If AI is to fulfill its promise, only the combination of techniques can lead towards the building of a true Artificial Consciousness.

# **1 - *Introduction***

## ***Background***

For years, man has been mystified by the 3 pounds of mass between his ears. The most complex organ in the human body, the brain is the source of our greatest achievements and most disastrous failures. This is the organ that is man, processes all of the information our body provides and makes him intelligent... and consciousness.

Over the last 40 years, Artificial Intelligence (AI) has become a broad umbrella for a large number of techniques, which have been expanded considerably since the early sixties when AI was first coined. A large number of these techniques, like modeling and knowledge representation, only barely touch the core of what AI was first about: the quest to make a mechanical device intelligent. The search for and replication of intelligence has since kept our community busy, with reasonable successes I should add. However, is it truly (human) intelligence that we want to achieve? Is it intelligence that makes someone human? Or is intelligence just a part of the equation?

My opinion in this matter is that not intelligence makes us human, but that both consciousness and intelligence makes us what we are. For this reason, this

paper will focus on Artificial Consciousness (AC) instead. It is often stated that AC will be a natural extension of AI, however we don't even have a clear view on what consciousness is! We have defined more or less what intelligence is, but this is much more vague for consciousness.

An example to illustrate: Is a dog intelligent? You might answer that, as a dog can fetch sticks, a dog does have some form of intelligence. Or you might say that as a dog isn't able to play a game of chess that a dog certainly isn't (very) intelligent. Both answers just matter in how high you lay the bar when it comes to comparing dog-intelligence with human-intelligence, it is hardly a yes-or-no question, but it is comparable. For intelligence is something you can actively observe and compare.

However: Is a dog consciousness? Is it aware of its own thoughts? As there is no way to compare consciousness, the only example we are sure of that is consciousness is ourselves. For this reason, consciousness is still mostly a subject of philosophy and psychology. The AI community rather ignores this subject all together, it seems. How could we then ever hope to create an Artificial Consciousness?

For this reason, in order to be able to construct an AC we first must define what consciousness is in the human brain. Although there appear to be as many concepts about consciousness as there are philosophers, my goal is to view the most prevailing ones and to compare these with each other.

## ***Problem statement***

In order to attempt to define consciousness, this paper will compare a number of current ideas on what consciousness is, primarily the views of functionalism (Dennett) and physicalism (Block). We won't try to define consciousness exactly, we shouldn't expect an easy answer, but we can take a look at the various different views on the subject in order to concentrate our search and come up with a working definition from these views.

Can aspects of consciousness can be represented using existing AI techniques?

To answer this question, the philosophical views above will then be compared to the various facets of AI (for example expert systems, neural networks, multi-agent systems or genetic algorithms) and be evaluated as to how likely it is for techniques to be used in the forming of an Artificial Consciousness. We will also look into the Society of Mind as a means of implementing both intelligence and consciousness.

During the whole paper, the reader is expected to think along with the broader questions: What is consciousness? Who am I?

## ***2 - Philosophical views***

As we have stated before, there are many theses on consciousness in the philosophy. For this discussion, we will omit the views on consciousness regarding the supernatural (the 'soul') and quantum gravity effects (Penrose-Hameroff), not that they are false but rather that most philosophers argue they are false and that they make computational, artificial, consciousness an impossibility.

What remains are a number of metaphors and theories, of which the most prevailing will be discussed below.

### ***Global Workspace theory***

The Global Workspace theory is based on the metaphor of the mind being a theater, in which the consciousness experience is the illuminated stage in a darkened auditorium. Many people are busy behind the scenes, but only what is currently on the stage is available to the whole audience. As such, consciousness is a form of publicity available to everyone who is watching the actors [[Baars, 1996](#)]. An interesting note is that this metaphor partially comes from Minsky's Society of Mind which, although Minsky is a founder of AI and possibly the most influential researcher in the field, has been assimilated more into philosophy than AI. We shall get back to his work in the next section.

In the Global Workspace theory, consciousness is associated with a 'global

broadcasting system' that transmits information throughout the brain. In order for the theory to work, we can describe three different entities: expert processors, a Global Workspace and contexts. The (unconscious) expert processors can be seen as human experts which are specialized in one function, and can be as small as a single neuron or encompass a whole area in the brain (like Broca's or Wernicke's area). The Global Workspace itself can be seen as the podium. Experts work together and compete with each other in order to reach the podium, where they can broadcast global messages to the entire brain. The third element of the GW theory is the context. These are the entities behind the scenes of the theater metaphor, the set of experts that guide, shape, constrain and facilitate the messages being sent from the global workspace. They provide the context of the messages being sent based, for example, on previous conscious events. Do note that none of these components on themselves are consciousness; only together do they provide the effect of consciousness.

As Dennett notes, there is a 'convergence coming from quite different quarters on a version of the global neuronal workspace model' [Dennett, 2001]. Not everyone agrees however, as we will see later on.

## ***Fame in the Brain, Functionalism***

“Functionalism is the view that mental events are the same as functional states; and a functional state is a state a physical system is in, when it has a set of sensory and other inputs, together with a set of potential behavioral outputs.” [Sanger, 2004]

Even though the GW theory in itself is reasonably accepted, different philosophers have different theories when it comes to the details. Dennett, for instance, rather uses the metaphor of different parts in the brain having various amounts of fame, and emphasizes that there isn't any such thing as top-down attentional control:

“We should be careful to take the term 'top-down' too literally. Since there is no single organizational summit to the brain, it means only

that such attentional amplification is not just modulated 'bottom-up' by features internal to the processing stream in which it rides, but also by sideways influences, from competitive, cooperative, collateral activities whose emergent net result is what we may lump together and call top-down influence" [\[Dennett, 2001\]](#)

As such, he continues, the brain works like a democracy, the 'top' is distributed among the contexts expert components behind the stage.

Dennett often relates consciousness to fame (cerebral celebrity), but he agrees that this is a slightly ambiguous term. As such, he prefers the term clout, political influence, that each part of the brain has. When certain parts of the brain compete for control on the 'stage', or global workspace, the one with the greatest clout wins, until there is another part which has collected an even greater clout or until the first part has lost his clout.

Dennett has always been vigorously against the metaphor of a central stage where an actor can broadcast his message, it's more of a debate between various parts of the mind, as such there isn't a central part of the brain where consciousness is achieved; on the contrary, different parts of the brain gain clout in this democratic, nearly anarchic, manner among connected areas (the more influential parts of the brain that agree with 'your' grab for attention, the more likely you are able to speak towards the other, possibly disagreeing, parts of the brain in the arena) . There is no central attention-giving process, there are only attention-getting processes. As such, he gives us two questions:

1 - How is this fame in the brain achieved?

2 - And then what happens?

Dennett has proposed an answer for the first question (competition and support from other parts of the brain), but the second question is what he coins the Hard Question. For what happens after a certain part in the brain has achieved enough fame, how is this focused attention on one part of the brain suddenly available throughout the entire brain? Some philosophers rather want to avoid the question as Dennett quotes:

“If, in short, there is a community of computers living in my head,

there had also better be somebody who is in charge; and by God, it had better be me." (Fodor, 1998, p207)

Dennett then continues with a metaphor that Real fame is not the cause of all the normal aftermath; it is the normal aftermath. As such, he argues that again there isn't anyone behind the controls: once real fame is achieved, global accessibility is achieved, and consciousness is achieved. The Hard Question is nothing more than the will of philosophers that there is a Subject in charge. As the whole is the Subject, nobody is in charge, and we come to the conclusion that there isn't a Hard Question. 'One of the hardest tasks thus facing those who would explain consciousness is recognizing when some feature has already been explained [...] and hence does not need to be explained again'. "Global accessibility IS consciousness" [\[Dennett, 2001\]](#)

## ***Phenomenality and Physicalism***

"Physicalism (or monoism) is the theory that every state of consciousness, every mental state, can be mapped to a brain state, and vice versa." [\[Sanger, 2004\]](#)

Physicalists see consciousness in a different, but related, perspective, compared with functionalism. Consciousness and awareness, according to Block [\[Block, 2001\]](#), are used in different ways and are used to explain different phenomenon. They are ambiguous terms, as he puts it.

Block illustrates his own opinion using the paradox of recent psychological findings about consciousness. Conscious perception of faces for example can be pinpointed quite exactly (FFP), as can perception of places (PPA). These areas aren't activated by stimuli that aren't faces or places, so the neural basis of consciousness can't be 'localized in one set of cells, but rather in the very areas that do perceptual analysis'. As such, visual consciousness would occur where it is perceived, it can be identified in the ventral stream as being responsible for visual consciousness. However, this ventral stream is also activated when the subject isn't consciously aware of a perception. Visual

extinction, visual neglect are terms in the psychology for when patients either don't see all the stimuli being presented ('favoring' one half of vision over the other, stimuli to the other side are extinguished) or when patients totally neglect one half of every object. This is in spite of the ventral (visual) stream being activated. Thus visual consciousness isn't located in the ventral stream at all! Block then continues about neurological findings between consciousness and unconsciousness perception. It appears that there is an equal amount of semantic priming on both sides of the brain, regardless of the fact that the subject is aware of a perception or not. Activation strength of stimuli thus can't make this difference, thus there must be something in addition that does.

The search is on for the remaining part of the puzzle, termed X. There are two views: either the ventral stream needs to provide a token, which is binded to the perception in order to make the rest of the mind aware of the perception, or consciousness is a matter of 'neural synchrony at fine timescales'. There are neurological findings (using the more time-scale accurate ERP technique compared to fMRI) that the latter is the case, thus consciousness could be related to an orchestra: only if every part of the orchestra works together and synchronous can a musical piece be performed without errors [my metaphor]. As such, only if every part of the brain fires at exactly the right time can there be an awareness throughout the whole brain (the information is globally accessible).

Block links consciousness to phenomality: 'what it is like to have an experience'. Accessibility is not the same as phenomality, and the X that causes one might not be the same that causes the other. He uses the example of suddenly noticing a sound while realizing that it has been going on for some time. There is thus phenomality without broadcasting through the global workspace, without global accessibility.

Summarizing, the major notion of physicalists is that they believe that they believe that consciousness is related to a physical or biological property that realizes it within humans, and humans alone. Global broadcasting alone can't be the only cause for consciousness, it is enabled by phenomality. Saying that

'X= global broadcasting' alone is a 'substantive claim but not one anyone can claim to know to be true'[[Block, 2001](#)].

### **3 - *Society of Mind***

'Society of Mind' is a book written by one of the founders of AI, Marvin Minsky, and has led to a number of changes in the field of AI.

The mind as Minsky sees it is composed of many, many functions, each being executed by a single agent, which reacts upon incoming information (from multiple agents) and sends its output to other (multiple) agents. Agents can work together, and thus form larger agencies that are uniform from the outside. These agencies can then in turn be seen as unitary agents. Agents can either be active or inactive, and process information in parallel regardless of the state of other.

There are a number of different primitive agents that make up the larger agencies:

- K-lines, which simply turn on other agents. These are the most common form of agents, but also are a powerful mechanism.
- Nemes, which invoke 'representations of things, and are mostly produced by learning from experience' [[Singh, 2003](#)]. Examples of these are polynemes and micronemes.

Nomes, which control how representations are manipulated. These are analogous to control mechanisms, which control 'how the representations are processed and manipulated'. Comparing these to a true society, these would be the rulers over the Nemes. Examples are Isonomes, Pronomes and Paranomes:

- Isonomes send messages to a number of agencies that should do the same kind of cognitive operation.
- Pronomes 'are isonomes that control the use of short-term memory representations'. They link to a large number of representations, describing what is related to a specific 'thing'.

- Paranomes are used to coordinate multiple representations. The example that Singh gives is that of a location paranome.

Minsky then continues on the use of Frames for representing knowledge.

Frames are constructed from a combination of the above agents, and are used to represent information. Each frame has a number of slots, which properties and other related 'things' are attached to. These then could be built from pronomes that control the use of the slots. Transframes, the central form of knowledge representation, represent events and everything related to the event. There are more related frame-types, but the overall picture of knowledge representation should be clear.

The difficulties arise when these large agencies (or societies of agents) have to communicate to each other, the global accessibility from the former theories. Minsky has the various agencies only communicate frames with those in its surroundings, but thanks to the use of paranomes, communication is far less used than would be thought, only links to the frames themselves are connected and disconnected. (Para)nomes already contain links to related entities, so instead of sending messages, links are made directly to the representation itself. For reasoning, Minsky offers a number of possible alternatives (a GPS-like difference engine, and Censors/Suppressors, a method for unlearning or suppressing 'negative expertise'). Another key subject for the Society of Mind is the growing of the network and the related topic of learning.

"What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle." [[Minsky, 1986](#)]

## ***4 - Comparison of views and links to AI techniques***

The main difference between physicalism on one side and functionalism on the other, is that global accessibility is viewed in a different perspective. For the

functionalists, "Global accessibility IS consciousness", broadcasting in the global workspace is all there is to it, while for physicalists accessibility is a side-effect of consciousness, but it surely isn't consciousness itself. Physicalists emphasize the fact that consciousness is a property that only is possible in humans, that it 'requires a specific biological realization'. Functionalism on the other hand is implementation-independent. It isn't surprising to see the similarities between functionalism and Society of Mind: in fact, Society of Mind could even be seen as an implementation of functionalism, yet it hasn't filled in the details. I think that this is the major problem why functionalism isn't generally accepted: It's too easy, it only fills in the why and how on an overall scale. Psychologists still will search for the holy grail of consciousness regardless: scientists have a natural tendency to try to find out why and how things tick.

Society of Mind has gone further than AI has gone today: it proposes a marriage between connectionist and symbolic methods. Minsky proposed both frames and agents, that together would provide the building blocks of the society. The debate between connectionists and symbolists however continues to this day. Neural networks or semantic knowledge representation? Society of Mind uses both, but at different levels, a revolutionary idea at the time that hasn't received a lot of backing or gained a large following. Connectionists have limited themselves to fixed stand-alone architectures, instead of seeing separate networks as building blocks for larger problems. The symbolists on the other hand are restricting themselves to only knowledge and not looking at matters like (common sense, general purpose) reasoning, adapting and learning.

The field of Multi-Agent Systems is a relatively new field, and one that is heavily influenced by Society of Mind. Yet, this field has hardly begun to implement the "richly heterogeneous architecture that uses multiple representations and methods, has special support for reasoning about commonsense domains, [...] that can make use of an initial endowment of commonsense knowledge, that can reflect upon and improve its own

behavior and develops in stages, and that is based upon not any one architectural arrangement of agents but rather is capable of reconfiguring itself to produce many distinct 'ways to think'" [Singh, 2003]

Instead, it appears to have been focused on a number of fixed architectures, instead of deepening the details laid out by Minsky (Reasoning, growing and adapting networks).

What shall we then think of physicalism, can we use this in our view of the Society of Mind? We must refute the notion that consciousness is something strictly biological; even if timing or something similar proves to be of the up most importance in reaching artificial consciousness, nothing is impossible to compute or simulate. We therefore would lean towards the notion of Dennett: 'Handsome is as handsome does, that matter matters only because of what matter can do'. Even so, we shouldn't throw out physicalism all together: consciousness still could be a mere 'phenomenon that implements global accessibility. The matter of activation without awareness shows us that the line between consciousness and unconsciousness is very thin. Especially the findings of introspectional attention could very much answer the question if consciousness truly is global accessibility or can be localized to various perceptual regions of the brain. Only then can we answer the question of what consciousness is and attempt to simulate this artificially, without doubts to what we are actually simulating.

## **5 - Conclusion**

The philosophical theory of Functionalism, and certain properties of Physicalism, are very much compatible with current techniques for creating AI systems. As expected however, consciousness still isn't defined. Is it an intrinsic property of the brain, or can it be localized? Are we using neural synchrony for awareness, or does timing not have to matter? The AI community can't prove physicalism, but what the AI community can do is attempt to prove Functionalism.

What is clear is that the Society of Mind theory has yet to be fully exploited by the AI community. Further in depth review of some of the proposed choices, or the creation of new ones and a comparison of attempted SoM implementations could lead to a deeper understanding of the inherent complexity of such societies. Also, it is clear that both connectionistic and symbolistic methods would be necessary in order to create such models. The search for consciousness and the search for artificial consciousness are both great undertakings, but in the end there is no other way to derive that what makes us human.

## Bibliography

Baars, 1996

B. J. Baars (1996), Cognitive views of consciousness: What are the facts? How can we explain them?, *The Science of Consciousness: Psychological, Neuropsychological, and Clinical Reviews*, volume 79. London, Routledge.

Block, 2001

N. Block (2001), Paradox and Cross Purposes in Recent Work on Consciousness, *Cognition*, April 2001, volume 79, 1-2

Dennett, 2001

D. Dennett (2001), Are we Explaining Consciousness Yet?, *Cognition*, April 2001, volume 79, 221-237

Minsky, 1986

M. Minsky (1986) *The Society of Mind*, Simon and Schuster, New York.

Sanger, 2004

L. Sanger (2004) Physicalism, Wikipedia:  
<http://en.wikipedia.org/wiki/Physicalism>

Singh, 2003

P. Singh (2003), Examining the Society of Mind, To appear in the journal *Computing and Informatics*, from:  
<http://web.media.mit.edu/~push/ExaminingSOM.html>