FACULTY OF SOCIAL SCIENCES
UNIVERSITY OF COPENHAGEN

Feeling of being

Consciousness as memes and vectors

Analysing consciousness bottom-up and top-down

Jens Egholm Pedersen <xtp778@alumni.ku.dk>

Supervisor
Claudia Carrara-Augustenborg <claudia.augustenborg@psy.ku.dk>

1 Introduction

Consciousness remains an elusive concept despite extensive scrutiny from many traditions (Van Gulick 2017). Philosophy of mind, neuroscience, psychology and - recently - computer science have been prodding the concept from a plenitude of angles. [1] This endeavour is paramount to understanding the human condition, but raises unavoidable and intricate existential questions (Amodei et al. 2016; Soares and Fallenstein 2016).

Since the 19th century developments within neuroscience are rapidly increasing our understanding of the cognitive processes that partake in the construction of consciousness (Atkinson, S.C. Thomas, and Cleeremans 2000).[2]

The advances and sheer amount of details neuroscience and neuroimaging techniques led Atkinson, S.C. Thomas, and Cleeremans (2000) to call for more detailed theories that identifies *neural correlates of consciousness (NCC)*. That call was answered by a plethora of theories that attempted to infer frameworks, on which novel understandings of consciousness could be based.[3]

Simultaneous to the advances in neuroscience, computational theory increased its efforts within artificial intelligence (AI) (Nilsson 2009), and has matured to a point where many of the neurophysiological properties can be replicated in silicon (Tononi 2004; Schmidhuber 2015; Walter, Röhrbein, and Knoll 2015).[4] The computational prowess of modern digital systems has been shown to grow with a staggering exponential speed (Moore 1965) - - a development that has proven to hold since 1965 and that shows no intention of halting (Moravec 1998; Kurzweil 2001). If the complexity of the brain can be assumed to be finite, this growth will inevitably catch up with the biological equivalent.[5]

[1] The Stanford Encyclopedia of Philosophy offers an overview of many of the disciplines and approaches involved in the quest to explain consciousness so far (Van Gulick 2017).

[2] Some philosophers require consciousness to include metaphysical properties (Van Gulick 2017) (dualism). This essay eludes the question by focusing on falsifiable and positivistic theories, in the hope that they can bring us closer to the truth - whether that entails dualism or reductionism.

[3] There are far too many relevant and interesting papers to list here, but to just mention a few influential examples, see Baars 2005; Block 2007; Crick and Koch 2003; DAMASIO 2003; Dehaene and Naccache 2001; Graziano 2013; Kouider et al. 2010; Tononi 2004; Zeki 2008; Schmidhuber 2015; Nilsson 2009.

[4] Other technologies show promising advances in forming computational substrates such as molecular biology and quantum computing, but have yet to reach the complexity of digital electronic computers.

[5] This is arguably already the case as demonstrated in this piece of software that simulates an

This last point is important because it tells us that computations can drastically aid the understanding of consciousness, provided that the paradigm of neuroscience and computer science find common footing through NCCs. Such a denominator is required because the very definition of computation depends on a formally defined input before any meaningful output can be given (Nilsson 2009; Schmidhuber 2015) (see glossary on page 11).

Despite the extensive research into NCC there is of yet no 'smoking gun' proving the exact link from neural systems to consciousness (Van Gulick 2017; Hohwy 2009; Dennett 2017). Many interesting candidates and perspectives that already accounts for much clinical evidence exist however.

This essay attempts to follow the advice of Hohwy (2009), who advocates a broader approach to finding NCC. Armed with Richard Dawkins' concept of *meme*s (Dennett 2017) and Geoffrey Hintons idea of *thought vectors* (Goh 2017), alternative conceptualisations of NCCs are briefly scrutinized through the lenses of two contemporary theoretical frameworks for consciousness: global workspace theory by Baars (2005) and the information integration theory by Tononi (2004).

Before proceeding, the reader should be aware that consciousness in this essay exclusively relates to the *hard* problem of inner experience as coined by Chalmers 1995, also known as phenomenological consciousness (see glossary on page 11).

2 Memes and thought vectors

Assuming an evolutionary viewpoint, any theory of consciousness will have to explain its added value to evolutional fitness[6] as well as its evolutionary path to achieve that outcome (Dennett 2017). Dennett calls these the *what for* and *how come* questions, and employs them rigorously on his journey to explain the concept of consciousness. It requires several layers of indirection to grasp the evolutionary path to the complex concept of the mind, especially in order to explain how it could arise in its present form, given the harsh

entire worm of the species C.elegans: `http://openworm.org`. Although the worm does not possess advanced cognitive abilities, this is a proof that biological organisms are not outside the reach of silicon wafers.

[6]Referred by some as *teleofunctionalism* as an opposition to functionalism, where pain, for instance, would not be seen as a function on the same level as a limb.

and ruthless principle of natural selection. For this purpose Dennett invites Dawkins' idea of a meme: "a kind of way of behaving ... that can be copied, transmitted, remembered, taught, shunned, denounced, brandished ..." (Dennett 2017). Memes fill the gap between the age of simple organisms and complex *talking* lifeforms because they are able to answer both the *what for* and *how come* of how language can evolve gradually.

Instead of requiring a fully functional uttered word to achieve any form of meaning in an evolutionary context, memes can explain why thoughts and concepts gained attraction in the race of life: they offer the aid of abstraction that drastically increases the ability to survive (Dennett 2017). Briefly speaking a meme is a tool for the brain, just like a hammer is a tool for the body. And like a hammer, a meme will exist in many variations. Like genetic mutations a large share of those will be useless, or maybe even harmful, but a fraction will prove to increase the fitness of the organism carrying the meme, thus ensuring its own survival[7]

According to the theory of memes an adult brain is then infected with countless memes with one common denominator: long history of keeping the adult (and the adults lineage) alive. Consciousness exists amidst all this chaos not as a product, but as a by-product to the communication between and within the memes themselves (Dennett 2017).

From the perspective of computer science a meme is an algorithm: a stepwise solution to a problem, that can take variation and eventualities into consideration (Nilsson 2009; Russell and Norvig 2002). Algorithms in computer science are formalised and described to great lengths, and have shown to be able to execute in many different computing substrates (Nilsson 2009). All algorithms have inputs in one form or another that completely determine the behaviour of the code. Such a list of inputs is also known as a vector and can be of arbitrary length.

The recent developments in machine learning is largely relying on vectors to input, transport and output data (Russell and Norvig 2002). Such complex learning models are comprised of many smaller algorithms who, also, depend on vectors. If one were to run a complicated learning model and freeze it in

[7]Interestingly enough this makes memes able to compete for fitness and survival, giving them some form of autonomy.

time, all the vectors could be captured as a 'snapshot' of the model in time. In other words the model can be represented as a large number of vectors at any given point in time, only to be resumed, repeated or halted at the whim of the creator (Nilsson 2009).

This 'snapshot' idea has led the deep-learning researcher Geoffrey Hinton at Google to contemplate a similar vector for the human mind that would be able to fully capture the state of a consciousness at a point in time(Goh 2017). Hinton did not further elaborate the idea, but the technique to take 'snapshots' of algorithm is widespread within the machine learning community (Russell and Norvig 2002).

Considering memes as algorithms reduces them to a "black box" which takes some input and returns some behaviour. This could, as well as any algorithm, be captured using vectors, allowing the memes to become the embodiment of the thought vector of the brain. In that light memes are formalized abstractions for the brain, capable of executing specified desired behaviour in a highly complex world. Further, this behaviour does not simply happen out of nowhere, but can be rigorously described in mathematical terms by capturing all the "*thought vectors*" of each individual meme and sub-meme.

3 The search for NCC

Guided by years of academic training and tradition, numerous disciplines have approached the problem of consciousness top-down by, in abstract terms, describing frameworks (like the global workspace theory) or bottom-up by establishing concepts that can scale to cover abstract ideas like consciousness (Van Gulick 2017).

The problem with both approaches is that they require a leap of faith in at least one direction. Bottom-up approaches are forced to mobilise immense methodological frameworks to establish the connection to something as complicated at phenomenology and qualia (Van Gulick 2017). Top-down perspectives needs to ratify their ideas towards the growing corpus of clinical data using some sort of measure that relates to the higher-order concept, essentially looking for a problem that fits their solution (Van Gulick 2017).

4 Testing memes

The likelihood of solving the puzzle from one direction alone is slim (Van Gulick 2017; Hohwy 2009).

Hohwy (2009) argues that further progress in the search for NCC requires new approaches to the study of consciousness. He suggests that future work "targets the presumably causal, mechanistic interplay between content processing and overall conscious state" (Hohwy 2009, p. 436).

Memes started in practice many million years ago as simple concepts to help the survivability of simpler organisms, but ended as complicated models and generalisations to assist in navigating and sense- making in a fast-paced world of constant abstract communication (Dennett 2017). Drawing upon this, memes can to a large extend be seen as the building blocks for our minds. Without abstractions to understand collections of items and words humans would never function in a modern society.

According to Dennett memes pervades our lives and defines our very language and, by extension, our mental models with which we see the world (Dennett 2017). If that is indeed the case it should be examined closer from both perspectives: top-down and bottom-up.

Baars (2005) and Tononi (2004) each posited their own influential theories of consciousness, representing top-down and bottom-up approaches respectively. In the following both theories will be assessed in the light of memes and thought vectors to test its applicability in both domains.

Global workspace theory

Baars (2005) suggests that consciousness exists in a *workspace* of the brain, whose contents "activate widespread regions in brain" (Baars 2005, p.52). Baars uses examples from neuroscience on sensory consciousness, working memory, attention and coma to explain how they all fit with the idea of one large canvas that concerts many complicated modules to form a unified notion of self.

Baars emphasises that the global workspace enables multiple networks to communicate with each other in order to solve tasks in concert (Baars 2005).

Within his framework then, is an understanding of vertical complexity, ranging from simpler brain modules to higher-order conglomerates. This is in line with Dawkins' idea of memes as a multi-layered and recursive construct, where memes can encompass other memes (Dennett 2017).

Baars writes that his global workspace theory "may be thought of as a theater of mental functioning", where only one spot on the stage is active at a time (Baars 2005, p. 46). The sensory cortices can be activated internally to perform unconscious processing which, when the signal is strong enough, triggers a conscious 'audience'. Together with S. Franklin, this has even been implemented as a computer model (Baars 2005), which aligns with the computational idea to consciousness of the thought vector approach.

The global workspace theory emphasises computation and empirical evidence to a degree that it is hard to shoot down, while at the same time being too diffuse in its terms to provide any straightforward methods to falsify its theories. It remains to be seen, for example, in what way the decentralised information is gathered in the workspace and what effect that has on consciousness.

Based on the clinical information for the global workspace theory, it is hard to find significant differences between it and the memes. Both theories operate on a high level, although Baars seem to elude the specific questions when it comes to the practicalities of putting his theory to the test.

Information integration theory

Tononi proposes an alternate theory on what consciousness is and how it occurs. He attacks it by attempting to solve the problem of how to measure consciousness (Tononi 2004). Tononi starts out by defining a measure for how well information is integrated in a network (Φ) [8] (Tononi 2004).
He then proceeds to equate this information integration measure with consciousness and proves its saliency by testing it with several empirical observations.

Tononi writes out the equation for Φ and thus formalises his theory math-

[8] By *integrated* he refers to how effective information can be relayed and made mutually available in a network (Tononi 2004)

ematically. This aligns with the computational paradigm where even complicated concepts can be reduced to its discrete constituents. However such a single measure for the 'experience' of consciousness resonates badly with the multifaceted view of memes. If our consciousness were to consist of many (competing) memes that constantly fight for the right to survive, consciousness itself would equal more than just the sum of its integrated parts.

The approach to memes would also be sceptical to the notion that consciousness scales with the amount of information. Rather, consciousness would scale with the efficacy of the memes to construct the side-effect that consciousness (according to the theory of memes) is. Both perspectives seems to be able to explain the neurological evidence mentioned in Tononi (2004). To mention a few from the memes point of view, split-brain patients will inevitably experience competing signals since memes per definition are competing, and the evidence pointing to consciousness in a distributed thalamocortical network (instead of a single cortical area) can easily be explained by memes being inherently decentralised (although higher functions is assumed to happen in the youngest parts of the brain) (Dennett 2017; Tononi 2004).

The basic building block of Tononi's theory (information) is consistent with the thought-vector approach to memes where the information is simply just algorithmic instructions. However, Tononi never explicitly defines what comprises information in his article, so the might be unseen discrepancies. On the higher level the theories start to diverge, perhaps because Tononi does not describe the concepts to great detail. If the information integration theory is taken to the extreme as presented in the article, the two theories will likely diverge due to the inherent methodological collectivism in (recurrent) memes.

5 Discussion

Merging the idea of memes with the idea of a thought vector allows for a method to approach the idea of higher-level thought, without losing any of the formalism required to test the theory in practice. This became apparent during the brief examination of Tononi (2004) and Baars (2005), because the theory offered valid explanations (alternative or aligning) in both cases. Instead of defining a top-down or bottom-up concept (like Φ) and scaling it to

explain clinical evidences, memes bridge the gap between the "non-thinking" and "thinking" substrates by offering a series of stepping stones.

Rather than dismissing the points of Tononi, memes offers counterproposals that are significantly easier to test in practice. By sticking to a model where memes must be autonomous entities with an algorithmic formalisation, memes have an advantage to many other theories in the literature that have proven near impossible to verify or falsify (Van Gulick 2017).

The weakness of the theory lies in the vagueness of memes. While it seems more feasible to formalise the definition of a meme than any other higher-level cognitive theory, it still remains to be seen how such a definition will look like. It is also not proven whether the current mathematical cool is capable of expressing the memes to a satisfying degree (what is algorithmic steps for a brain?). Similar to the 'minimally sufficient' condition for consciousness to occur (see glossary on NCC) a neural correlation of memes (NCM) is urgently required. Although the problem is arguably not as big as finding NCC.

6 Conclusion

This essay merged the idea of memes with the notion of thought vectors and briefly applied the concepts to two influential theories of consciousness: global workspace theory and integrated information theory.

The analysis only scratched the surface of a century old and abstruse subject, but provided nonetheless a reason for optimism. It was shown that 1) memes can explain parts of the neurological evidence from the literature, that 2) memes were largely compatible with the views presented by both Baars (2005) and Tononi (2004) with some added nuances and finally that 3) memes offer both a top-down and bottom-up perspective.

The last point is especially interesting because our most powerful computing substrate (silicon transistors) are excellent bottom-up machines, but are ineffective at abstracting to a level where top-down processing is feasible. Memes allow computational substrate to bootstrap this high-level way of *comprehension*, and may be a way to empower computational machines with the ability to finally adopt new perspectives and augment our capability for top-down thinking - and eventually our consciousness.

A raw and unguided tour de force into AI, void of any form of human consciousness or ethics, is hardly desirable and the advancements in deep learning further stresses the search for a proven theory of conscious and NCCs.

Glossary

artificial intelligence Artificial intelligence (AI) covers the broad discipline in computer science that is concerned with replicating intelligent behaviour in computational systems. The exact definition is controversial for historical reasons (Nilsson 2009). . 2

bottom-up Bottom-up approaches in this article refer to the combination of many smaller concepts to form a greater whole. This approach is typical for the natural sciences. An example of such a bottom-up approach to understanding consciousness is Tononi's idea of an information integration measure (Tononi 2004). 5, 6, 9, 12

computation Computation refers to any process (in any substrate) that can deduce new information based on old information. In this is manifested as computing instructions. 2

consciousness Consciousness pertains to the feeling of being alive and attentive. This circular definition covers over the fact that consciousness is an old and multifaceted idea that covers many complicated concepts (Van Gulick 2017). In this essay consciousness will exclusively relate to the *hard* problem of inner experience as coined by Chalmers 1995, also known as phenomenological consciousness.. 2

meme *Meme* is a shortened form of the ancient Greek *mimeme* meaning 'imitated thing' and was coined by Richard Dawkins. A meme refers to a idea or a *way of behaving* that can be "copied, transmitted, remembered, taught, shunned, brandished, ridiculed, parodied, censored, hallowed" (Dennett 2017). 3–6, 8

NCC Neural patterns or condition that is minimally sufficient for a conscious thought to occur. See (Atkinson, S.C. Thomas, and Cleeremans 2000; Hohwy 2009). 2, 3, 6, 9, 10

thought vector A thought vector is a list of numbers (vector) that describes the attributes of a state within a neural network (Goh 2017) at a specific point in time. A thought vector thus captures the *thought* of a network at a single instant. One can imagine how this can be applied to larger neural systems, like mammal brains, to 'capture' a mental state. 3

top-down This essay employs top-down as a higher-order approach to a solution or approach to a problem. An example of a top-down approach to understanding consciousness is the global workspace theory by (Baars 2005) or the framework presented by Francis Crick and Christof Koch (Crick and Koch 2003). While both contain elements of neurobiology (bottom-up) they are explicitly trying to offer an explanation on what consciousness is. 5, 6, 9

Bibliography

Amodei, Dario et al. (2016). "Concrete Problems in AI Safety". In: *CoRR* abs/1606.06565. arXiv: 1606.06565. URL: http://arxiv.org/abs/1606.06565.

Atkinson, Anthony, Michael S.C. Thomas, and Axel Cleeremans (Nov. 2000). "Consciousness: Mapping the theoretical landscape". In: 4, pp. 372–382.

Baars, Bernard (Feb. 2005). "Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience". In: 150, pp. 45–53.

Block, Ned (2007). "Consciousness, accessibility, and the mesh between psychology and neuroscience". In: *Behavioral and Brain Sciences* 30.5-6, pp. 481–499. DOI: 10.1017/S0140525X07002786.

Chalmers, David J. (1995). *Facing Up to the Problem of Consciousness*. URL: http://cogprints.org/316/.

Crick, Francis and Christof Koch (Feb. 2003). In: 6.

DAMASIO, ANTONIO (2003). "Feelings of Emotion and the Self". In: *Annals of the New York Academy of Sciences* 1001.1, pp. 253–261. ISSN: 1749-6632. DOI: 10.1196/annals.1279.014. URL: http://dx.doi.org/10.1196/annals.1279.014.

Dehaene, Stanislas and Lionel Naccache (May 2001). "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework". In: 79, pp. 1–37.

Dennett, Daniel C. (2017). *From bacteria to Bach and back*. ISBN: 978-0-393-24207-2.

Goh, Gabriel (2017). *Decoding the Thought Vector*. URL: https://gabgoh.github.io/ThoughtVectors/ (visited on 12/27/2017).

Graziano, Michael S. A. (2013). *Consciousness and the Social Brain*. ISBN: 978-0190263195.

Hohwy, Jakob (2009). "The Neural Correlates of Consciousness: New Experimental Approaches Needed?" In: *Consciousness and Cognition* 18.2, pp. 428–438.

Kouider, Sid et al. (July 2010). "How Rich is Consciousness? The Partial Awareness Hypothesis". In: 14, pp. 301–7.

Kurzweil, Ray (Mar. 2001). *The Law of Accelerating Returns*. URL: http://www.kurzweilai.net/the-law-of-accelerating-returns.

Moore, G. E. (Apr. 1965). "Cramming More Components onto Integrated Circuits". In: *Electronics* 38.8, pp. 114–117. ISSN: 0018-9219. DOI: 10.1109/jproc.1998.658762. URL: http://dx.doi.org/10.1109/jproc.1998.658762.

Moravec, Hans (1998). "When will computer hardware match the human brain". In: *Journal of Transhumanism* 1.

Nilsson, Nils J. (2009). *The Quest for Artificial Intelligence*. 1st. New York, NY, USA: Cambridge University Press. ISBN: 0521122937, 9780521122931.

Russell, Stuart J. and Peter Norvig (Dec. 2002). *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall. ISBN: 0137903952. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20%5C&path=ASIN/0137903952.

Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61. Published online 2014; based on TR arXiv:1404.7828 [cs.NE], pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

Soares, Nate and Benya Fallenstein (2016). "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda". In: *The Technological Singularity*.

Tononi, Giulio (Nov. 2004). "An information integration theory of consciousness". In: *BMC Neuroscience* 5.1, p. 42. ISSN: 1471-2202. DOI: 10.1186/1471-2202-5-42. URL: https://doi.org/10.1186/1471-2202-5-42.

Van Gulick, Robert (2017). "Consciousness". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2017. Metaphysics Research Lab, Stanford University.

Walter, Florian, Florian Röhrbein, and Alois Knoll (2015). "Neuromorphic implementations of neurobiological learning algorithms for spiking neural networks". In: *Neural Networks* 72.Supplement C. Neurobiologically Inspired Robotics: Enhanced Autonomy through Neuromorphic Cognition, pp. 152–167. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2015.07.004. URL: http://www.sciencedirect.com/science/article/pii/S0893608015001410.

Zeki, Semir (Feb. 2008). "The disunity of consciousness". In: 168, pp. 11–8.