FACULTY OF SOCIAL SCIENCES
UNIVERSITY OF COPENHAGEN

Feeling of being

Dimensionality of consciousness

On measuring experience

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

Abstract
Modern developments ...

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

Chapter 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 highly 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)* (Atkinson, S.C. Thomas, and Cleeremans 2000). 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

[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.

equivalent.[5]

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 also glossary on page 4).

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).

This essay attempts to approach this problem from a new angle. Armed with Richard Dawkins' idea on *meme*s (Dennett 2017) and Geoffrey Hintons idea of *thought vectors* (Goh 2017),

Before proceeding, the reader should be aware of this essay consciousness will exclusively relate to the *hard* problem of inner experience as coined by Chalmers 1995, also known as phenomenological consciousness.

1.1 Vectors and dimensionality

what for how come

1.2 Alternative approaches

Common for each contribution is a fundamental desire to deepen the understanding of both consciousness as a concept, and the principles partaking in the creation of the concept. Guided by years of academic training and tradition, each discipline have approached this top-down by, in abstract terms, describing or bottom-up (Van Gulick 2017), : either constructing abstract frameworks (Block 2007; Kouider et al. 2010) or

A raw and unguided tour de force into AI void of any form of human consciousness or ethics is hardly desirable, and, putting the moral existential and moral dilemmas aside, these advancements further stresses the search for NCCs.

top-down *and* bottom-up (Dehaene and Naccache 2001; Baars 2005)

1.3 Convergent theory

Some researchers even call for new experimental approaches, because the current

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

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). 3, 5

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

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

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.. 3

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.

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.