Phenomenal Experience and the Perceptual Binding State

Joscha Bach

Harvard Program for Evolutionary Dynamics, Cambridge, MA 02138, USA
[redacted]

Abstract. How can a computational model of cognition account for the hard problem of consciousness? This contribution addresses some of our intuitions about the nature of phenomenal experience and the first person perspective, and suggests avenues for their realization in a cognitive architecture.

Keywords: phenomenal consciousness, hard problem, cortical conductor theory, attention, binding, cognitive architectures.

Dealing with the Hard Problem

It seems that if Artificial Intelligence is pursued as a cognitive science, it cannot avoid to account for the arguably most elusive and mercurial property of the human mind: the conscious experience of phenomenal states. While I have tried to account for some of the functional properties of machine consciousness elsewhere (Bach 2018a, 2018b, 2009), this contribution avoids technical and formal arguments and instead tries to offer a brief introduction into some of the most relevant conceptual intuitions with regard to understanding consciousness as a property of an intelligent system.

Modeling perception, memory, decision making, reward based motivation provide challenges to cognitive science, yet nothing about these faculties seems mysterious. The same applies to extending AI systems with reflexive and metacognitive capabilities. But how could an AI model ever hope to explain the *feeling of what-it's-like*? David Chalmers (1995) characterizes this as the *Hard Problem of Consciousness:* The ability of an organism to be the "subject of experience". To further specify what that means, Giulio Tononi and Christof Koch (2015) have offered five axioms, which I would briefly summarize as follows:

1. Consciousness is real and actual (for the same reason that compelled Descartes to his famous dictum "cogito, ergo sum"). 2. Consciousness is compositional (for instance, visual experience is structured by color and shape). 3. Experience is specific, it happens in ways that distinguishes it from other experiences. 4. Consciousness is integrated (unified): the elements of an experience are interdependent so that they are experienced together, and cannot be reduced to separate elements (e.g. color and shape in visual perception). 5. Conscious experience has borders: it specifies certain things and thereby excludes others.

While these axioms are neither particularly axiomatic nor sufficient to explain what we mean by consciousness (see Bayne 2018), they form the foundation of To-

noni's *Integrated Information Theory* (IIT) that attempts to explain consciousness via the degree of mutual information between the operations taking place in different parts of a cognitive system (Tononi 2012, 2016).

It seems to be entirely plausible that the observable behavior of humans results from physical interactions that are orchestrated by the cells and chemical and electrical signals of the nervous system, which is why contemporary neuroscience and psychology largely subscribe to a physicalist world view. Yet it seems to be implausible that a physical process can make a physical object (such as an organism) experience anything. This apparent conundrum has several possible traditional resolutions:

1. Physicalism is false, and the world we inhabit is not physical, but entirely experiential (a kind of dream, but with constraints that are given by a mind outside of our own). This idealist position was advanced for instance by the 19th century philosophers Schelling and Hegel, and is also found in many religious traditions. Because physicalist theories of the universe (which describe it as a causally closed realm that operates in accord with mechanical laws) are very successful in explaining our observations, the denial of physicalism is not a popular position among scientists.

2. The physical and the mental domain are separate realms of existence. This dualist position is often attributed to René Descartes, and was further developed by Gottfried Wilhelm Leibniz (who suggested that these separate domains co-evolve according to a "prestabilized harmony") and Nicolas Malebranche. A problem with dualism is that the physical world is either defined or empirically observed as causally closed: if the mind could change the physical world, it would violate conservation laws, and if it cannot change the physical world, consciousness is a mere epiphenomenon. That means: if an epiphenomenalist thinker has phenomenal experience outside of physics, it cannot have caused the thinker's argument, because the expression of an argument must happen in the world of physics. The last substantial effort to find an avenue for dualism that I am aware of was made by Karl Popper and John Eccles (1977), who suggested that quantum effects in the neocortex might be random enough to not violate conservation laws, but could still act as a conduit between the mental and the physical realm. (Conservation laws indicate that information itself is conserved, so this argument seems hopeless.)

3. Understanding consciousness will require extending our understanding of physics beyond computational principles. To escape the mechanist perspective, it will not be sufficient to discover new ways in which information flows through the universe, and our view of physical reality may have to break out of the computational frameworks of existing physics. This is the position of Roger Penrose (1998). I would like to argue that for epistemological reasons (we can only make finitely resolved, local observations, and we have to describe our models in computational languages), such principles would remain outside of the universe we can observe and model. From the perspective of an embedded observer, they would therefore appear to be an inseparable aspect of the material universe, which makes this position indistinguishable from panpsychism. A problem with the position of physics being incomplete is similar to dualism: the Standard Model of physics appears to be already sufficiently detailed to explain all the necessary dynamics of organisms, and it does not leave obvious causal gaps through which the non computational physics of phenomenal consciousness

could influence our actions. (The non-obvious gap that Penrose argues for is quantum gravity, since current physics does not yet offer a generally accepted solution for unifying quantum mechanics and curved space.)

4. Consciousness can in principle not be explained. This is the mysterianist position, which has been argued by Colin McGinn (1999), and more recently also by Noam Chomsky.

5. Scientific research should focus on the functional aspects of consciousness, such as the fact that conscious attention relates different aspects of perceptual content and knowledge to each other, which is explored in Bernard Baars' Global Workspace theory (1993), and its neuroscientific adaptation by Stanislas Dehaene (2014). A detailed functional explanation of the performance of conscious agents will converge to an explanation of phenomenal consciousness.

6. Phenomenal consciousness does not exist and is an illusion (Frankish 2016, Dennett 1992, 2016).

We should notice that if mental events are produced by the nervous system, and the relevant dynamics of the nervous system can in principle be described by known physics, the conscious experience of events cannot happen in or very close to real time. Due to the slow rate of signal propagation in the nervous system, the processing of sensory modalities can take many hundreds of milliseconds. Also, different sensory modalities of the same event (such as feeling how a foot touches the ground, and hearing the sound of the foot step) do not take the same amount of time to process, and will have to be fused later on, and later stimuli can change the experience of earlier ones (Stiles et al. 2018). Thus, conscious experience cannot happen in actuality, but must be constructed after the fact.

Furthermore, conscious experience is also not simply time-shifted: the subjectively immediate initiation and execution of intentional actions in response to sensory events is part of our experience, too. The initiation of a deliberate act is not instantaneous, but will require at least several hundred milliseconds of cortical activity (Libet 1983), and thus, a deliberate reaction to a sensory event such as stubbing one's toe may easily take more than a second. Since we don't experience this delay, our subjective conscious experience of agency in the present cannot be real. Despite our experience, Tononi and Koch's first axiom of conscious experience is incompatible with both physicalism and neuroscientific evidence. From the perspective of physicalism, Frankish and Dennett appear to be correct, and the solution to the Hard Problem is clear: a physical system, such as an organism, cannot actually have phenomenal experience. Thus, what needs to be explained is not how an organism can have phenomenal experience, but why it appears to us that we do!

This apparent paradox can be resolved when we realize that we are not actually organisms, realized by physics, and living in a physical environment. The world we experience is not the physical world, but a virtual world that is being generated by our mind, which is implemented by the nervous system, in an attempt to explain sensory patterns. The same circuitry that produces dreams when we sleep does so when we are awake, but during wakefulness, the dream is tuned to predict the patterns generated by our sensory nerves. The subject of experience, the self, is a virtual character inhabit-

ing this virtual world, just like the main character of a novel inhabits a fictional universe. The self is not identical to the organism. Instead, the self is a model of the regulation dynamics of that organism. The self is also not an agent. It is a passenger to all the activities performed by the organism, a simulacrum that the brain generates to predict, evaluate and plan the trajectory of the organism through its environment. The apparent causal relationship between the self and behavior is simply due to the fact that the organism uses the self as a model for regulating its actions. (For a detailed argument about why the self does not possess agency but is a representation, see Metzinger 2003.)

While the idealist and physicalist positions are at odds with each other when we understand them as ontological statements about reality, they are complementary with respect to the mind: We do live in a dream, each one of us in a separate one, and the dream, including all its inhabitants, is generated by a brain of an organism living in a physical universe. The reason why we experience things in a particular way is the same why a character in a novel does: because the contents of our experience and the fact of the experience itself are written in exactly this way by its author. Like a character in a novel, we generally also don't notice that we are not real, as long as the author does not write the discovery that we are not real into our story. (The psychological phenomenon known as "derealization/depersonalization disorder" may represent and exception from this rule.) Our phenomenal experience is very real to ourselves, but our selves are not real. In other words, when Tononi and Koch (2015) argue that only a physical organism can have conscious experience, but a simulation cannot, they got it exactly backwards: a physical system cannot be conscious, only a simulation can.

2 The architecture of perception

For something to qualify as an experienced content, it is necessary and sufficient to be recalled as having been the subject of our attention. We cannot recall what we do not remember having attended to, and nothing we can recall as having attended to is not conscious. This implies that we possess special attentional system that is combined with an indexed memory that is integrated in such a way that all the different memories can be related to each other. I call this attentional system the Cortical Conductor (Bach 2018a). The idea of treating consciousness as a model of attention has also been suggested in several forms by Graziano and Webb (2014), Dennett (1992), Drescher (2006) and others. The cortical conductor is a small part of a larger architecture of perception, depicted in a simplified way in figure 1.

Human perception likely begins with the formation of a somatosensory model in utero. Hebbian learning can connect terminals of sensory neurons that fire at the same time to allow the formation of a map of the body surface in the primary somatosensory cortex, which is extended into a model of the spatial arrangement of the body by combining it with proprioceptive, vestibular and muscle control information. The correlation of the tactile, visual and auditory modalities allows extending the tactile space into a model of the immediate environment. The presence of reward signals,

emotional modulators and motivational urges along with proprioception allows the agent to model its motivational, emotional and hedonic state (Bach 2015). Motivational states either relate to the somatic regulation (nutrition, health, rest etc.), the social regulation (affiliation, nurturing, dominance etc.) or the cognitive regulation (competence, exploration, aesthetics). Together with the attentional state and the short-term and long-term biographical memory of the agent, they form the model of the agent's own regulation: the self. This model is not identical with the regulation, but it allows the organism to explain, predict, and evaluate its own behavior, and thereby improve the regulation. The self model and the current world state constitute an *ego centric* local perceptual space of the organism. The agent is also able to create counterfactual world states (imagined or remembered mental states that don't conform to the present state of the environment or self). This *mental stage* is crucial for planning, learning and reasoning.

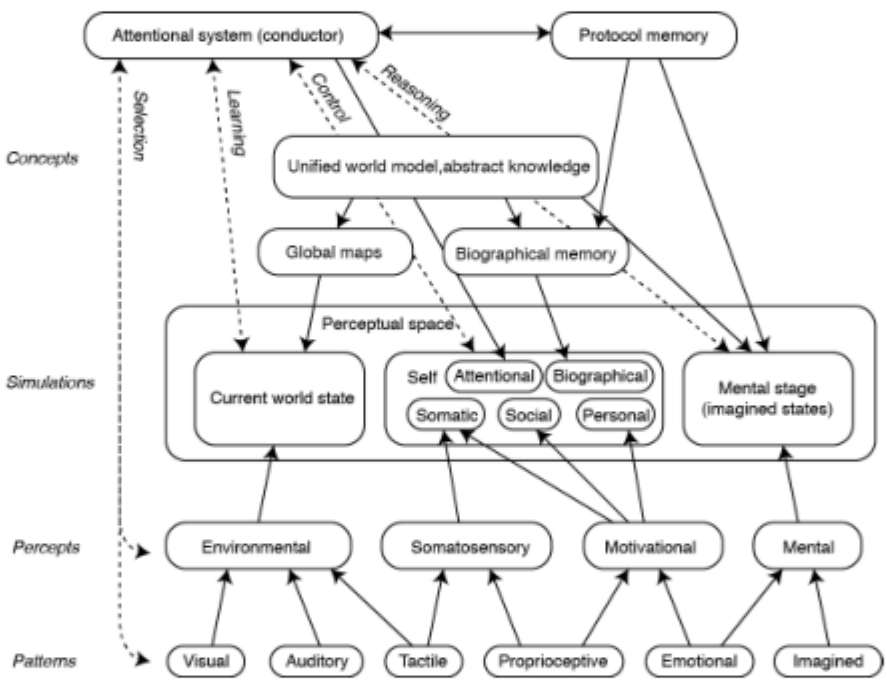

Fig. 1. The architecture of perception

When the agent moves to a new location, or changes in the self and environment happen, the perceptual space changes. These changes are modeled using global allocentric maps and the biographical memory of the agent, and the knowledge and various cognitive tools that allow us to predict, explain and evaluate the new states. At each point in time, only a small part of the total world model of the agent is instantiated in the perceptual space and the currently active percepts and sensory patterns. Together with the state of the attentional system, this current instantiation amounts to the state

of the *working memory* of the agent. Working memory amounts to a set of values of latent variables and relationships that bind them to each other: the *binding state*.

The attentional system selects sensory patterns and percepts based on their relevance for updating the perceptual space. A main role of the attentional system may consist in its support for learning. Attentional learning works by making a local intervention in the model, and storing this model together with the expected outcome of the change, the current partial binding state of the perceptual space (a memory of the present situation) and the conditions under which we expect to be able to learn whether our intervention was successful. When the world state with the required learning signal is encountered later on, the partial binding state during the time of the intervention can be recalled, and the change to the model can be reinforced or undone.

Storage and recall require that the attentional system has access to an integrated protocol memory. Conscious attention may be understood as the ability to store indexed memories.

Phenomenal consciousness is the memory of a specific perceptual binding state. Because the attentional system has itself to be trained by attentional learning, the act of accessing it may be stored in the attentional protocol as well. *Access consciousness* may be understood as the memory of accessing a specific binding state, as stored in the attentional protocol. *Reflexive consciousness* is the memory of the experience of that access, i.e. a model of the self as a system that has access to its own experience.

3 The realization of the binding state

The local perceptual space is a computational operator that explains the current perceptual patterns, as they are available to the agent. This model can be understood as a set of variable parameters (the state of the model) and computational relationships between them (the invariances of the model). Perception is the process of creating coherence between the variables by propagating their relationships between them until a consistent state is reached. A perfect model would be constrained in such a way that every possible stable configuration of the free model parameters corresponds to exactly one possible state of the ground truth, and the available sensory patterns of most ground truth states allow an efficient convergence of the model state to the ground truth state. The role of attention in perception is to repair inconsistent parts of the perceptual model, usually by modifying the relationships between the values, that is, by changing the way these values are bound to each other. At each moment, only a small fraction of the known relationships between the possible model parameters is instantiated as a binding state in working memory (i.e. we only perceive a very small part and moment of the entire possible universe). The more relationships we can establish without breaking the coherence of the perceptual operator, the more sensory data we can explain, and the better we can predict future states. Unlike many machine learning models of perception that stop at recognizing individual, independent patterns (such as datasets of unrelated bitmaps), all sensory inputs of an organism correspond to an aspect in the same coherent and continuous universe, and thus, it should ideally be interpreted as a part of a single unified modeling function.

The binding problem of neuroscience is somewhat similar to the binding problem of the internet: how is it possible that a very specific configuration of information exchanging units, such as neurons or internet servers, can organize and stabilize itself across a large network of possible connections? The answer to that problem may also be similar: the neurons might be organized so that they can implement a protocol that allows them to enter one of the currently bound sets of neurons in a specific relationship, to stay in this group for as long as necessary, and to leave it when required.

There seem to be two broad families of approaches to the binding problem, either by sending repeated local codes between assemblies of neurons (such as cortical columns) that can act similar to the "ping" and "ack" messages of computers on the internet. *Request Confirmation Networks* (Bach and Gallagher 2018) describe the implementation of such a protocol. In neuroscience, a similar approach is found in feature integration theory (Treisman and Gelade 1980). Another way of achieving binding is called synchronization theory (Milner 1974, von der Mahlsburg 1981), and suggests that the synchronous firing of neurons results in binding them. Both theories continue to have new developments and proponents, presumably at least in part because both of them predict somewhat similar observations: the implementation of a local signaling protocol that connects large groups of neural assemblies across the cortex will be observable as synchronized oscillations. Thus, the more interesting question is about the causal order: is binding the result of synchronous firing, or is it the other way around?

An interesting nonlocal version of the synchronization theory that may explain how the frequency of the firing itself could be causative for binding is advanced by Crick and Koch (1990) and may perhaps be phrased as a *neural ether* theory. Here, cortical neurons are interpreted as forming a lattice that propagates signals globally at different frequencies, and by tuning in to a given frequency (i.e. sampling the signals of other neurons at fixed intervals) neural assemblies can bind themselves to the processing of different perceptual content, like a radio receiver can tune in to a particular program. From the perspective of Tononi's and Koch's IIT, a state of increased consciousness would correspond to one in which a large part of the cortical activity would receive and participate in generating the same perceptual program. The existence of a cohesive perceptual binding state is plausibly a necessary condition for phenomenal experience, but not a sufficient one. If the brain creates a kind of perceptual radio program and uses that to orchestrate the behavior of the organism, what is listening? Rather than the universe itself, as some panpsychists believe, or some entity outside of the physical universe, as dualists claim, I'd like to suggest that conscious experience is a model of the contents of our attention: it is virtual, a component of the organism's simulated self model, and produced by an attentional conductor.

Acknowledgements

This work has been supported by the Harvard Program for Evolutionary Dynamics, the MIT Media Lab and the Epstein Foundation. I am indebted to Katherine Gallagher, Adam Marblestone, and the students of the Future of Artificial Intelligence

course at the MIT Media Lab for their contributions in discussions of the topic, as well as to Martin Novak and Joi Ito for their support.

References

Baars, B. (1993). A Cognitive Theory of Consciousness. Cambridge University Press

Bach, J. (2009). Principles of Synthetic Intelligence. Psi, an architecture of motivated cognition. Oxford University Press (2009).

Bach, J. (2015). Modeling Motivation in MicroPsi 2. Artificial General Intelligence, 8th International Conference, AGI 2015, Berlin, Germany: 3–13 (2015).

Bach, J. (2018a). The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models. Postproceedings of BICA 2018

Bach, J. (2018b). Consciousness in Humans and other Machines. Postproceedings of AAAI Fall Symposium WS "Towards common model of cognition"

Bach, J., Gallagher K. (2018). Request Confirmation Networks in MicroPsi 2. Proceedings of AGI 2018, Springer.

Bayne, T. (2018) On the axiomatic foundations of the integrated information theory of consciousness, Neuroscience of Consciousness, Volume 2018, Issue 1

Crick, F., C. Koch (1990). A Framework for Consciousness. *Nature Neuroscience*. **6**: 119–26

Dehaene, S. (2014). Consciousness and the Brain. Viking Press

Dennett, D. C. (2016). Illusionism as the Obvious Default Theory of Consciousness. Journal of Consciousness Studies, 23, No. 11–12, pp. 65–72

Dennett, D. C. (1992). Consciousness Explained. Back Bay Books, New York

Drescher, G. (2006). Good and Real. MIT Press

Frankish, K. (2016). Illusionism as a Theory of Consciousness

Graziano, M. S. A., Webb, T. W. (2014). A Mechanistic Theory of Consciousness. International Journal on Machine Consciousness

Libet, B., Gleason, C.A, Wright, E.W., Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain.1983 Sep; 106 (Pt 3): 623-42

von der Malsburg, C. (1981). The correlation theory of brain function. MPI Biophysical Chemistry, Internal Report 81–2. Reprinted in Models of Neural Networks II (1994), E. Domany, J.L. van Hemmen, and K. Schulten, eds. (Berlin: Springer)

McGinn, Colin (1999). The Mysterious Flame. Conscious Minds in a Material World. Basic Books

Metzinger, T. (2003). Being no One. MIT Press

Milner, P.M. (1974). A model for visual shape recognition. *Psychol Rev*. **81** (6): 521–35. doi:10.1037/h0037149. PMID 4445414

Penrose, Roger (1989). The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics. Oxford University Press

Popper, K., Eccles, J.C. (1977). The Self and its Brain. Springer

Stiles, N.R.B. et al. (2018). What you saw is what you will hear: Two new illusions with audiovisual postdictive effects, *PLOS ONE*

Tononi, G. (2012): The integrated information theory of consciousness: an updated account. Arch. Ital. Biol. 150, 56–90

Tononi, G., Koch, C. (2015). Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society B; 370:20140167

Tononi, G., Boly, M., Massimini, M., Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neurosc. 17 (7): 450–461

Treisman, A., Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, Vol. 12, No. 1, pp. 97–136