Correlations, Theories, and Simulations of Consciousness and Other Emergent Phenomena

Shane Celis

Neuroscience of Consciousness

May 2, 2011

1 Introduction

The study of consciousness has had a resurgence with the search for the Neural Correlates of Consciousness (NCC). One might claim it has made consciousness a scientifically respectable topic. NCC can inform theory but it is essentially theory neutral[3]. Qualitative theories about how consciousness and brains relate abound, but they are often difficult to verify experimentally. Recently, however, theories that with quantitative measures have been proposed, which makes these theories more ammenable to experimental verification. This paper will compare neural correlation approaches to theories of consciouness, suggest a thought experiment to assist in the comparison, and consider the conjecture that consciousness is simulated.

The correlation approach this paper will focus on is Causal Density (CD), and the theory will be Integrated Information Theory (IIT). Each approach provides a quantitative measure of a system. However, what the measures purport to mean differ. CD measures what fraction of nodes in a network are causally significant using Granger causality[21]. Granger causality is a

statistical interpretation of causality. A variable x is said to Granger-cause y when knowledge of x reduces the prediction error of y [21]. It is, however, susceptible to making incorrect inferences due to hidden variables, e.g., it may appear that x Granger-causes y when in fact the hidden variable z Granger-causes both x and y.

IIT provides a quantitative measurement for how much integrated information a system is capable of[22]. This measure is meant to capture the simultaneous integration and differentiation that seems a fundamental part of consciousness. One's experience is integrated, not piecemeal. A visual scene is revealed all at once not in fragments. Yet each experience from one moment to the next is differentiated from the last. IIT measures the capacity of a system to simultaneously integrate and differentiate information. However, it goes further than providing a measure; IIT claims that the integrated information *is* consciousness. It is an identity claim.[1] If this claim is true, it would have some provocative consequences.

The IIT claim means that whenever one finds integrated information, a consciousness of some sort exists and vice versa. It need not have a body, nor senses, nor muscles. Even a humble thermostat might be conscious, perhaps with only a shallow 1 bit of consciousness. Further, it would have ontological consequences[22], namely it asserts that consciousness is as fundamental as mass or charge. That particles exist with mass and charge is accepted as a brute fact. Tononi suggests that consciousness must likewise be accepted as fundamental. Tononi is not alone in this view.

Chalmers supports Tononi's view with his zombie worlds argument[4]. Briefly, it states that one can imagine a world that is physically identical to our own but where no consciousness exists. All behaviour on zombie world is the same but no lights are on; therefore, consciousness is something extra that cannot explained by the purely physicalist paradigm.[2] Chalmers also claims

[1]Identity claims seem philosophically problematic, but Hesslow shows some standard arguments against identity claims of this sort ought not be relied upon[10].

[2]This is not a knock down argument since it relies of conceivability as a proxy for

that consciousness is the best example of a strongly emergent phenomena[5].

Broadly the study of emergence is concerned with how a set of micro-laws and micro-entities can give rise to a set of macro-laws and macro-entities. For instance, consider the behaviour of water an emergent phenomena. At the micro-level, a water molecule has polarity, but it does not have surface tension. At the macro-level, a large collection of water molecules does not have polarity, but it does have surface tension. There is not anything especially tricky about this kind of emergence; Bedau calls it weak emergence[1]. A phenomena is weakly emergent if given the micro-laws and micro-entities, one can simulate them to deduce the macro phenomena. Weakly emergent systems also obey causal fundamentalism: "The macro is the way it is in virtue of how things are at the micro"[12].

A strongly emergent phenomena, however, is a strange beast. It is a case peculiar to science where the macro-level has properties and even causal potency independent of the micro-level. Chalmers claims that consciousness is the best—perhaps only—known example of a strongly emergent phenomena[5]. There are appealing aspects to this claim. It would explain why consciousness sticks out like a sore thumb in the sciences. It could provide free will with causal potency that is independent of physical matter. However, strong emergence is problematic. If independent macro-laws conflict with micro-laws, which one wins? If macro-laws do not conflict, they are superfluous. Strong emergence violates causal fundamentalism and implies downward causation. An example of strong emergence meant to inspire credulity would be to accept that a bee hive has a hive mind which, for lack of a better word, telepathically directs the actions of its workers. A weak emergence perspective would suggest that workers behaviour is locally directed through natural processes that merely appear coordinated as if the hive had a hive mind. No irrefutable evidence has been presented to demonstrate that strong emergence exists. This paper will focus on weakly emergent phenomena which is

possibility. A thing may be conceivable that is physically or logically impossible.

not controversial.

Another consequence of IIT, it specifically allows for conscious artefacts. Since it provides a measurement, one could blindly search for a machine that scored a high integrated information. Seth demonstrated that one need not even perform a search; one could analytically construct a type of artificial neural network for whatever value of integrated information one wished[20]. This result does not prove that integrated information is not consciousness, but it does reduce the claim's credibility because no one has any expectation that the constructed neural networks would be conscious to any degree.

2 Thought Experiment

Deciphering how correlation and theory work on the consciousness is difficult in part because consciousness is not well understood. Instead, consider the following thought experiment, which substitutes consciousness with something better understood yet retains some of the same structural problems. Suppose a universe like ours exists where its inhabitants are in a brains-in-vats scenario similar to the portrayal in the movie The Matrix [23]. See the diagram shown in Figure 1. The physical reality the inhabitants experience is simulated by a computer, but the neurological processes are carried out by regular physical processes.[3] This thought experiment does not require that either physics or consciousness is computable.[4] Let us assume for the sake of discussion that the physics is restricted to Newtonian physics whose primary entities are particles.[5]

[3]There are two reasons to use a brain-in-a-vat scenario: 1. One need not assume that consciousness is computable. 2. To clearly separate the harder problem consciousness from the easier problem of physics.

[4]The simulation need only have a computable facsimile of realtime physics such that its inhabitants believe they are in a spatial and temporal world grossly similar to ours but it need not be exactly the same at all scales.

[5]The thought experiment need not restricted to Newtonian physics, but it makes the discussion easier.

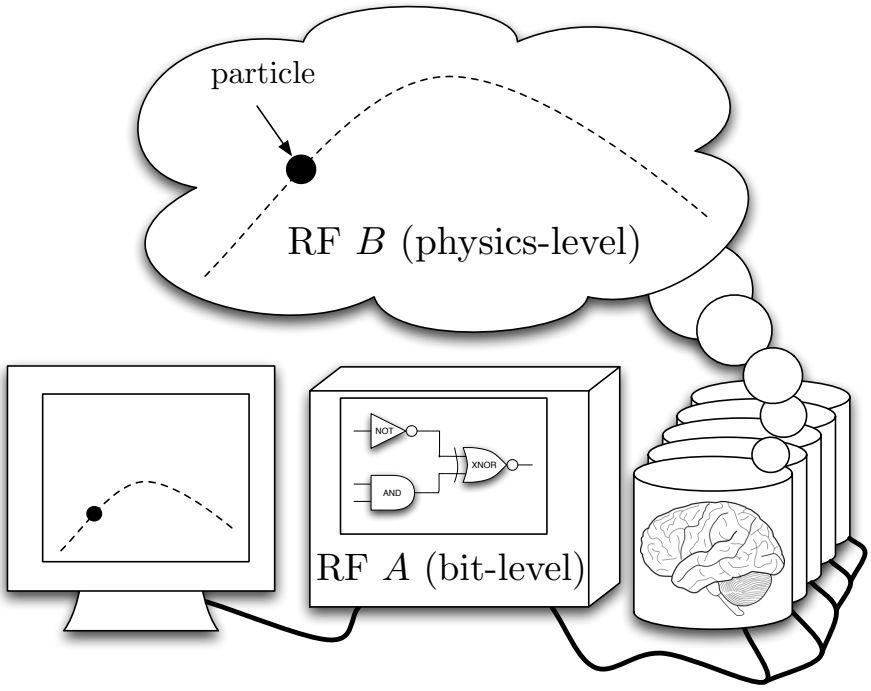

Figure 1: Diagram of the Matrix-like world of the inhabitants. Inhabitants' brains are hooked up to a machine that simulates a physical environment. Reference Frame (RF) A is the bit-level that contains bits, logic gates, and all the causally relevant entities required for a digital computer. RF B is the physics-level that contains particles and all the relevant entities required such that the inhabitants believe they are in a universe similar to ours.

Given this scenario, assume that the inhabitants become partly aware of their condition and have a means of determining the state of the machine that simulates their physical environment. Teams of researchers tackle the problem searching for the Bit Correlates of Physics (BCP). Researchers correlate patterns of bits in the machine with physical activity in their world. But why should those bits produce any physical activity at all? A prominent philsopher argues that one can imagine a universe where the machine performs the same bit manipulations yet no extra physical environment is instantiated;

therefore, physics must be something extra that cannot be entirely explained by the bitalist paradigm, which assumes everything is ultimately constructed out of bits.

A researcher proposes that all bit manipulations, including those within the inhabitants' own computers, are in fact instantiating physical worlds. Shrewder researchers constrain their theory such that only certain bit manipulations instantiate physical worlds. Researchers discover that other machines like theirs exist, some which behave similarly at the bit-level, others which behave quite differently. Questions are posed: Do these machines have an "inner world" like theirs? Can one determine what those worlds are like?

This may seem like a playful caricature of neuroscience research, but there is a serious aim. We have constructed a scrutable scenario that retains some problems present in the neuroscience of consciousness with a simpler subject matter. To summarise, the physical environment is a weakly emergent phenomena that arises due to the activity of the machine. The description of each level is drastically different. The micro-level, or bit-level, is the machine (RF A) which is composed of bits and logic gates. The macro-level, or physics-level, is the physical environment (RF B) which is composed of particles with properties like velocity and position, etc. These different levels are structurally similar to the problem in consciousness of objective and subjective facts. The bit-level is analogous to the neural-level, and the physics-level is analogous to the conscious-level.

Let us consider correlation and theory in this case. How do such tools function in this scenario? BCP can identify how the bit-level correlates with the physics-level, but it cannot reveal why or how one level causes the other. Essentially, it does not explain the correlation; it only describes that a correlation exists. In fact, one may not be able to determine that any causal relationship exists. For instance, if the correlation were known, but there was no way to disrupt the state of the machine, it is difficult to see how a causal relationship could be proven empirically. One might instead prefer to

consider that such domains happen to exhibit a coincidental correlation as implausible as that may seem.[6]

A bit theory of physics should be capable of explaining why certain bit manipulations result in a physical instantation. An attractive theory might propose that certain bit configurations, manipulated in particular ways, were *identical* to physical particles. If one were to follow Tononi's example, this would have similar ontological consequences. The ontological entities at the bit-level would have to admit not just bits and logic gates but particles and positions, etc. This is a strange concession because the bits and logic gates are all that is necessary to provide the complete causal story. The particles and other simulated physical entities have no autonomous causal potency; they are merely along for the ride, an epiphenomena. Suppose instead that the machine simulated a calculator, one must admit numbers and operators as fundamental entities into its ontology. One can see that by this thinking whatever can be simulated must be admitted as fundamental by relying on an identity claim that a micro-level entity *is* a macro-level entity, which seems unreasonable.

What is the right theory to knit together these two disparate sets of phenomena, the bit-level and the physics-level, for this thought experiment? This ought to be a simple problem. After all it is premise of many science fiction novels and movies, but even here it is worthwhile to proceed carefully. One could state that certain bit configurations are identical to particles. One reason not to, mentioned in the preceding paragraph, was because of the ontological consequences. Additionally, there are many ways of simulating physics[7], so one must be more abstract. One could define a particle by its intrinsic properties (position, velocity, mass, charge) and its causal

[6]Just such an idea parallelism, or pre-established harmony, was offered by Leibniz as a solution to the conflict between dualism and determinism with causal closure: the mind and the world correlated with each other perfectly, such that they appeared to interact causally but no such interaction actually occurred between the mind and body[13].

[7]Functionally identical physics simulations may be obtained with different implementations. To borrow a term from philosophy of mind, it has a kind of multiple realisability.

properties (like charges repel, universal attraction). Any system, including one composed of bits, which could reproduce those properties could be said to instantiate a physical environment. Notice that there is no intrinsic relationship between the bit-level and the physics-level since both are causally closed systems. Bits do not entail particles. Nor do particles entail bits. Instead, a sufficiently plastic causal substrate (the bit-level) may be arranged to simulate a new causal substrate (the physics-level). And as evidenced by electronic computers, the reverse is also true. This is the simulation theory of bits and physics that ought to satisfy the inhabitants about how these two levels relate.

This simulation theory is susceptible to the charge that the particles are merely in the eye of the beholder: Essentially, a simulated particle is not a real particle. Searle famously argued that simulation is not instantiation. He wrote, "No one supposes that a computer simulation of a storm will leave us all wet"[18]. There is merit to this charge. The particles are illusory in a sense. The particles do not have any autonomous causal potency. The particles are entirely constituted by the simulation. An explosion of these particles will not harm the computational substrate which simulates them, as Searle rightly notes. However, if one considers the inhabitants of this thought experiment, the physics of a storm could in fact cause its inhabitants to become wet. Or, to be more technical, it would cause its inhabitants to have the sensory experience of wetness. To carefully articulate the meaning, qualifiers are necessary. The particles are illusory with respect to the bit-level, Reference Frame A (RF A). The particles are real, or indistinguishable from real, with respect to the physics-level, RF B.

3 Simulation and Causation

One idea that motivates the preceding section is that given some preconditions, simulation is indistinguishable from reality. In contradiction to Searle,

simulation can instantiate causal system with some caveats. One caveat is that the simulation must be causally isomorphic, i.e., it must preserve the same causal properties of the system. Many simulations used in science are not causally isomorphic, most are merely quantitatively isomorphic. Further, if a system is causally closed, then it is indeterminable from within that system whether any deeper causal substrate exists.[8] This indeterminability has ontological consequences. For example, assume the inhabitants are physicalists who accept that only physical properties exist in their universe. Once they learn their physics environment is simulated, physicalism no longer holds. They could become bitalists, but there is no assurance that they have reached the deepest causal substrate because it is indeterminable from within. To rescue ontology from being unknowable, one could claim a causally closed simulation has its own ontology.[9]

4 Simulation and Consciousness

Returning to neuroscience, my conjecture is that consciousness is simulated. This is not a new idea[15]. Even researchers outside of neuroscience have articulated it[6]. Merker gives an evolutionary account where consciousness "takes the form of a synthetic, stabilized, and coherent neural simulation of the animals body in relation to its surrounding space"[14].

Before going further, it is worthwhile to identify prior uses of the word simulation in the neuroscience literature that do not relate to the conjecture in this paper. Hesslow proposes the internal stimulation of sensory and motor systems, i.e. simulation of such systems, produce an "inner world" [11]. Note that this does not reveal how such a simulation produces phenomenal experience; although, it is claimed to be an explanation. Gordon proposes self simulation as a means of predicting other people's behaviour, but makes

[8]This only holds if the causal isomorphism is continuously preserved.

[9]Or a causally closed simulation merely behaves as though it has its own ontology.

no claims about phenomenal experience [9]. Neither of these uses of the word simulation represent the claim made in this paper.

4.1 Eliminativism

If consciousness is simulated, it provides a new way of looking at eliminativism, which is the claim that qualia is illusory[7]. Perhaps qualia is illusory in the same way that particles are illusory with respect to the bit-level in the thought experiment, i.e., qualia is an illusion with respect to the physical substrate it is implemented on, but qualia is real, or indistinguishable from real, within the simulation of consciousness.

4.2 Epiphenomenalism

One consequence of this conjecture is consciousness seems epiphenomenal, i.e. mental events would merely be along for the ride with no autonomous causal potency.[10] Not only that, epiphenomenalism would be rampant: anything that may be simulated could in principle be epiphenomenal. One used to worry that mental events might be mere epiphenomena, but now one has no assurance that physical events are not epiphenomena as well. However, if the simulated system has causal closure, then that causal substrate could in principle be the fundamental level.

Consider the meaninglessness of epiphenomenalism if any fundamental causal substrate may in principle be substituted by some other synthetic causal substrate. Imagine scientists determine that our universe is in fact simulated on a computer as suggested by The Simulation Argument[2]. Would one then suggest our physical world is a mere epiphenomena? Perhaps, one would and even be partly correct in the assertion. Would bombs be any less

[10]I argue that epiphenomenalism does not imply that mental events are causally impotent; it implies that mental events are causally potent but causally reducible to physical events.

destructive because they were epiphenomenal? Epiphenomenalism, in that case, is an irrelevant metaphysical issue because the universe need not be any different in principle if its fundamental causal substrate is implemented by our physics or some other synthetic substrate beneath our physics.

The up side of this conjecture is that it side steps the issues with claiming an identity relationship. Instead, it connects the neural-level to the conscious-level with a causal relationship. The causal relationship is peculiar but it does not seem to be that much more baffling than the one given by the thought experiment. It may stimulate ideas for research in deducing higher level causal substrates from lower levels.

The down side of this conjecture is that it does not provide any details on how consciousness is achieved by the brain. The conjecture essentially does some philosophical tidying up by providing a framework for understanding how anything could ever be conscious within a physicalist paradigm. Or more broadly, how a set of new ontological and causally potent entities can emerge in a system.

The thrust of this paper may sound like a recapitulation of the ideas that motivated Artificial Intelligence (AI): the brain is the hardware that simulates the mind which is the software, and so long as one found the right program, one could produce an intelligent artefact on a digital computer[17]. However, AI focused on functionalism, which might satisfy a behaviourist but not someone who was interested in consciousness. AI also focused on computation as the key to understanding intelligence. Instead, this paper is treating computation merely as a well understood plastic causal substrate. While it would be nice if consciousness were computable, this conjecture is not committed to such a view.

The direction of research this conjecture suggests is, given the micro-level of a causal system, determine what, if any, macro-level causal system might exist. One could begin with some toy systems with two separate causally closed systems. Assume one has a bit system that simulates either

a calculator or a particle system. Can one deduce whether the system is simulating a calculator or a particle system? Both the calculator and particle system can be implemented in many different but functionally compatible ways, which makes the task more difficult. Can one analyse the bit system without knowing what to expect and derive what kinds of macro-level entities it might contain? Answering either question seems quite difficult and may not be possible in general.[11] Complexity science with its investigation of emergent behaviour seems well poised to attempt tackling such a question, and measures like causal density[21], discovering a domain alphabet[16], or measuring emergence[19] seem like promising ways to begin such an analysis. Systems that are not causally closed will be more difficult analyse, but they would also be more instructive because the brain is not a causally closed system.

We have found the substrate that constitutes our perceptions, feelings, and dreams, but we do not yet know how such things might emerge from our brains. Correlation and theory were discussed as they relate to neuroscience and more generally with the help of a thought experiment. The thought experiment also retained problems analogous to the separate domains of objective and subjective facts. Two conjectures were posited: simulation can instantiate, and consciousness is simulated. Ultimately, this paper suggests that analysing the causal substrate in the brain may reveal higher causal substrates, including the substrate that we might identify as consciousness. We can count ourselves lucky that we know there is something to look for within our heads. For once we have acquired the scientific tools to find it, who knows how many hidden worlds might lie at our feet.

References

[1] M Bedau. Draft for principia: comments welcome downward causation and the autonomy of weak emergence. *Citeseer*, Jan 2007.

[11] See Dennett's Two Black Boxes thought experiment. [8, pp. 412]

[2] N Bostrom. Are we living in a computer simulation? *The Philosophical Quarterly*, Jan 2003.

[3] D.J Chalmers. What is a neural correlate of consciousness. *Neural correlates of consciousness: Empirical and conceptual questions*, pages 17–40, 2000.

[4] D.J Chalmers. Consciousness and its place in nature. 2002.

[5] D.J Chalmers. Strong and weak emergence. *The re-emergence of emergence. The emergentist hypothesis from science to religion*, pages 244–256, 2006.

[6] Richard Dawkins. The selfish gene. page 360, Jan 2006.

[7] Daniel C. Dennett. Consciousness explained. page 511, Jan 1993.

[8] Daniel Clement Dennett. Darwin's dangerous idea: Evolution and the meanings of life. page 586, Jan 1996.

[9] R Gordon. Folk psychology as mental simulation. *citeulike.org*, Jan 2004.

[10] G Hesslow. Will neuroscience explain consciousness? *Journal of theoretical biology*, 171(1):29–39, 1994.

[11] Germund Hesslow. Conscious thought as simulation of behaviour and perception. *Trends Cogn Sci (Regul Ed)*, 6(6):242–247, Jun 2002.

[12] F Jackson.... In defense of explanatory ecumenism. *Economics and Philosophy*, Jan 1992.

[13] Gottfried Wilhelm Leibniz and George R. Montgomery. Discourse on metaphysics and the monadology. page 92, Jan 2008.

[14] B Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, Jan 2005.

[15] A Revonsuo. Consciousness, dreams and virtual realities. *Philosophical Psychology*, Jan 1995.

[16] M.D Schmidt and H Lipson. Discovering a domain alphabet. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1083–1090, 2009.

[17] J Searle. Is the brain a digital computer? *Proceedings and Addresses of the American ...*, Jan 1990.

[18] John R. Searle. Minds, brains, and science. page 107, Jan 1984.

[19] A Seth. Measuring emergence via nonlinear granger causality. *Artif. Life*, pages 545–552, 2008.

[20] A Seth, E Izhikevich, and G Reeke.... Theories and measures of consciousness: an extended framework. *Proceedings of the ...*, Jan 2006.

[21] A.K Seth. Causal connectivity of evolved neural networks during behavior. *Network: Computation in Neural Systems*, 16(1):35–54, 2005.

[22] G Tononi. Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, 215(3):216, 2008.

[23] Andy Wachowski and Lana Wachowski. The matrix. 1999.