Using information theory to measure the emergence of artificial free will in a spiking brain-constrained model of the human cortex

Josh Bourne jbour006@gold.ac.uk
Goldsmiths, University of London, London, SE14 6NW, UK

Fernando Rosas f.rosas@sussex.ac.uk
Department of Informatics, University of Sussex, Brighton, UK; Centre for Psychedelic Research, Department of Brain Science, Imperial College London, London, UK; Centre for Complexity Science, Imperial College London, London, UK; Centre for Eudaimonia and Human Flourishing, University of Oxford, Oxford, UK.

Max Garagnani m.garagnani@gold.ac.uk
Department of Computing, Goldsmiths, University of London, London, SE14 6NW, UK
Department of Philosophy and Humanities, Brain Language Laboratory, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany.

Abstract

Cell Assembly (CA) circuits are known to emerge in neurocomputational models as a result of Hebbian-like learning. Intriguingly, when a brain-like architecture is used, CAs spontaneously "ignite" in absence of any stimulus, and the patterns of network activation occurring during such ignitions closely match those observed in the human brain during non-stimulus driven, endogenous decisions to act [1]. This suggests that sub-threshold reverberation of noise within CA circuits (which drives their ignition) may be a possible mechanism underlying seemingly "free" and volitional (yet possibly pre-consciously determined) action decisions [2]. It is unclear, however, whether such spontaneous CA ignitions are truly an emergent property of the brain-like model, or whether they are somehow "pre-encoded" in the system's features. Can we provide objective evidence supporting (or falsifying) the hypothesis that these ignition events are de facto non pre-determined and can be thus be considered as the network's own endogenous "action decisions"?

To investigate this issue, we used a spiking brain-constrained model of six cortical areas and, after replicating the previously documented CA emergence and spontaneous ignitions in it, we analysed its emergent properties using information theoretic measures. Recent techniques in information theory allow quantifying emergence in complex systems (including the brain) [3]. Here, we applied these measures to test for the presence of emergence during spontaneous, unprovoked CA circuit ignition. Specifically, we analysed the different modalities of emergence associated with cell assembly ignition and lifecycle (downward causation and causal decoupling).

Preliminary results show the highest levels of emergent behaviour (specifically, causal decoupling) during cell assembly ignition, which gradually fade as CA activation dissipates. Such increased levels of causal decoupling observed during (and prior to) CA ignition episodes confirm the presence of an emergent feature in the neural model.

In summary, we present here the application of formal criteria used for determining the presence of emergence in complex systems to a spiking, brain-constrained neurocomputational model of the cortex that can mechanistically explain the neural origins of so-called "free", volitional action decisions. Initial results of the information-theoretical analysis indicate that spontaneous CA circuit ignitions, driven by reverberation of noise within them, truly constitute an emergent feature of the brain-like architecture, suggesting that this phenomenon should be considered as an endogenous (i.e., internally generated, and not pre-determined) feature of the artificial neural network.

References:

[1] Garagnani, M**.** & Pulvermüller, F. (2013) Neuronal correlates of decisions to speak and act: spontaneous emergence and dynamic topographies in a computational model of frontal and temporal areas. *Brain and Language* **127**(1):75–85.

[2] Schurger A, Mylopoulos M, Rosenthal D. Neural antecedents of spontaneous voluntary movement: a new perspective. Trends in Cognitive Sciences. 2016, 20(2), 77-79.

[3] Rosas F.E., Mediano P.A.M., Jensen H.J., Seth A.K., Barrett A.B., Carhart-Harris R.L., et al. (2020) Reconciling emergences: An informationtheoretic approach to identify causal emergence in multivariate data. PLoS Comput Biol 16(12): e1008289.