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

Josh Bourne[1], Fernando Rosas[2,3,4,5], Max Garagnani[1,6]

[1]Computing Department, Goldsmiths, University of London (UK); [2]Department of Informatics, University of Sussex; [3]Department of Brain Science, Imperial College London; [4]Centre for Complexity Science, Imperial College London; [5]Centre for Eudaimonia and Human Flourishing, University of Oxford (UK); [7]Department of Philosophy and Humanities, Freie Universität Berlin (Germany).

Introduction

Cell Assembly (CA) circuits emerge in neurocomputational models as a result of Hebbian-like learning. When a brain-like architecture is used (see Fig. 1 below), 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 (Garagnani & Pulvermüller, 2013).

It is unclear, however, whether such spontaneous CA ignitions are truly an emergent property of the brain-like model, or whether they are "pre-encoded" in the system's features.

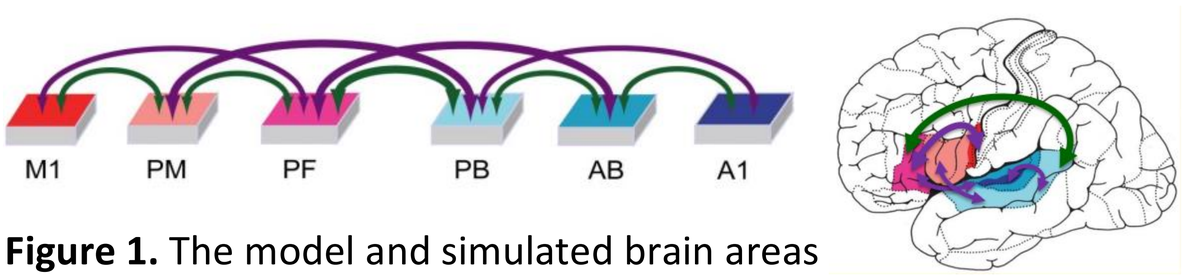

Figure 1. The model and simulated brain areas

Research Questions / Aims

This project used information theory to analyse the properties of a spiking version of the above network during spontaneous CA circuit ignition episodes in it, addressing the following questions:

- Do the seemingly unpredictable CA ignitions constitute an emergent phenomenon of the system?
- If emergence is present, then of what kind, and how is this influenced by CA ignition and lifecycle when compared to baseline activity?

Methods

1. We replicated the phenomenon of spontaneous CA ignition using the spiking version of the model (Garagnani *et al*, 2017)
2. We recorded CA circuit activity during 90 spontaneous ignition episodes and used the "Practical criteria" approximation technique, (Rosas et al 2020) to formally analyse the emergent properties of the system's dynamics
3. CA cells' activity was compared to the activity of a control group of randomly chosen cells of the same size (n=90) as the CA considered. Independent samples t-tests were used to evaluate the emergent properties across the two groups.

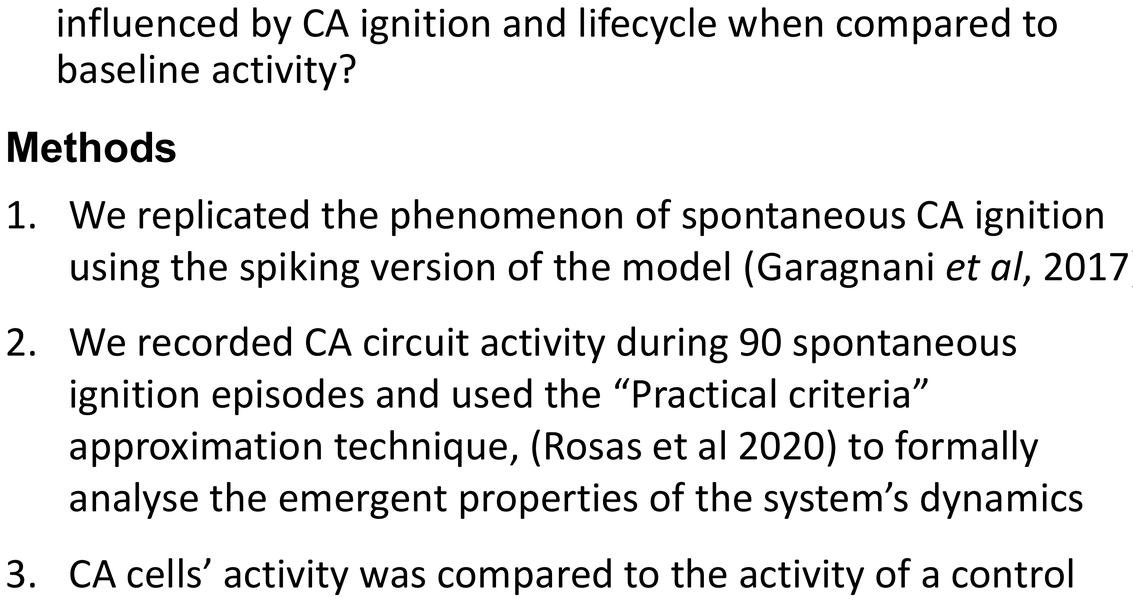

Figure 2: Causal decoupling shows how an emergent feature influencing itself without interaction with the micro elements. Downward causation shows the emergent property influencing micro elements but not itself.

Results

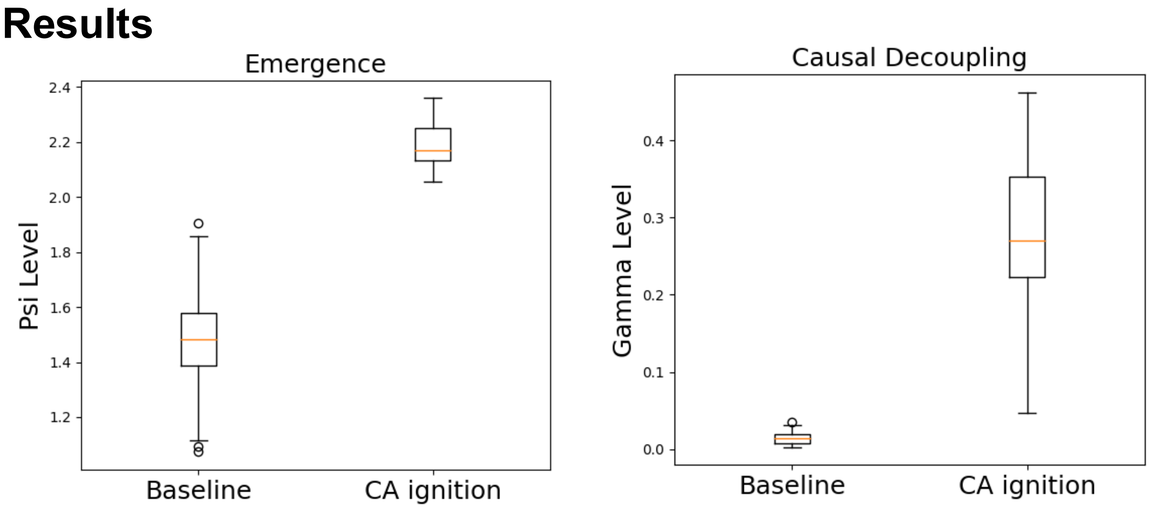

Figure 3: Average level of emergence is consistently higher during CA ignition than baseline activity (Left). Peak levels of causal decoupling during CA ignition are higher and much more varied than baseline (right).

We found:

- a significant increase in the average level of emergence (psi) in the CA Ignition group vs control group $t(178)=36$, $p=<.001$. This demonstrates that CA ignition is associated with increased levels of emergence when compared to baseline firing alone;
- during peak CA ignition a smaller, more varied, but significant level of causal decoupling (gamma) $t(178)=86$, $p=<.001$;
- that downward causation (delta) is quickly driven negative during CA ignition and so is unlikely to play any significant role in this activity.

Summary

We applied a fully brain constrained spiking model which is known to closely reflect, and explain, patterns of real brain activity during "free" action decisions. We investigated the spontaneous ignitions of distributed cell assembly circuits in it:

- We demonstrated that CA ignition presents measurably emergent properties (i.e., it produces synergistic information) that could not be explicitly pre-encoded in the system, but that result from the interaction of the system with itself;
- The observed causal decoupling and lack of downward causation indicates that CA activity is predictive of itself as a whole but not of its parts; namely, the individual neurons it is composed of.

Conclusions & Further work

We submit that *the observed spontaneous CA ignitions can be interpreted as the model correlate of endogenous action decisions, or as a form of "artificial free will".*

This claim rests on two main observations:

1. As seen here, the CA circuit ignitions do constitute a truly spontaneous, non-predetermined phenomenon of the system;

   and

2. the system in which such phenomenon happens is a brain-constrained model, and one whose activity is known to mimic brain pattern observed during spontaneous action (Garagnani & Pulvermüller, 2013).

To further understand the dynamics observed here, future investigation should focus on incorporating multiple simultaneous and sequential CA ignitions and activity patterns in the analysis.

References

Garagnani M. & Pulvermüller, F. (2013). *Brain and Language*, **127**(1):75-85.

Garagnani M, Lucchese G, Tomasello R, Wennekers T and Pulvermüller F. (2017) *Frontiers in Computational Neuroscience* **10**:145.

Rosas FE, Mediano PAM, Jensen HJ, Seth AK, Barrett AB, Carhart-Harris RL, *et al.* (2020) *PLoS Comput Biol* **16**(12): e1008289. https://doi.org/10.1371/journal. pcbi.1008289