Convolutionary, Evolutionary, Revolutionary: What's next for Bodies, Brains and AI?

Abstract: In recent years we have made significant progress identifying computational principles that underlie neural function. While not yet complete, we have sufficient evidence that a synthesis of these ideas could result in an understanding of how neural computation emerges from a combination of innate dynamics and plasticity, and which could potentially be used to construct new AI technologies with unique capabilities. I discuss the relevant principles, the advantages they have for computation, and how they can benefit AI. Limitations of current AI are generally recognized, but fewer people are aware that we understand enough about the brain to immediately offer novel AI formulations.

Principles of neural function and plasticity.

The recent unexpected difficulties in realizing truly autonomous vehicles, making reliable medical diagnoses, detecting offensive online content, creating useful chat-bots and even just recognizing faces, show that brains remain functionally more capable than we can currently emulate. Significant differences exist between spiking neural networks (SNNs) and the superficially-similar artificial neural networks (ANNs) used for Deep Learning (DL), and in these differences can be found potential clues to the next AI revolution:

- Real neurons communicate with spikes where spike timing forms an integral part of the neural code[1]. The computational benefits of sparse spike coding are substantial[2]. SNNs are also rigorously more powerful than their real-valued counterparts[3].
- Networks self-organize to represent feedforward input structure[4]. The mechanisms the brain uses to accomplish this have been established (spike timing dependent plasticity (STDP)[5-7], several homeostatic mechanisms[5,6] and local decorrelating inhibition[5-7] together implement sparse non-negative matrix factorization[6,8]).
- Abundant feedback connections build predictive models[9-11] by learning to invert the self-organized feedforward representations (also using STDP[11]).
- Oscillations[12,13] and quasi-chaotic transitions[14,15] continuously reconfigure neural circuits based on the needs of the task at hand. Activity levels, oscillation frequencies and proximity to phase boundaries are controlled through recurrent thalamic projections[16].
- Multi-scale inhibitory mechanisms[7,13,16,17] ensure only best-matching circuits are engaged for any given task, implementing powerful *k-winner-take-all* computations[18].
- Spike conduction delays[13,19], oscillations[13] and short-term plasticity (STP)[19,20] innately represent time in the brain[19,20]. In recurrent neural circuits, STP can maintain memory states (working memory) for indefinite lengths of time[21].
- Dopamine modulates the gain of STDP for model-free reinforcement learning (RL)[22].

The above principles are well-established, even if all are not yet universally accepted. Other principles are more speculative, or the specific underlying mechanisms are not fully established:

- Brains combine predictive models of the world with oscillations and dynamic circuit reconfiguration to create internalized simulations of 'what if' scenarios and future plans[23,24]. These models also flag unexpected and out-of-distribution events[25].
- Closely related to the above is the idea of stochastic sampling[11,26]. This same process probably applies to outputs (actions), including sampling from possible *sequences* of actions (equivalent to 'what if' simulations).
- Dopamine works with oscillations, dynamic circuit reconfiguration and the internal

world models to implement model-based RL. How this occurs is unclear, but prediction of reward and temporal-difference (TD)-style learning will be essential components[27].

- Explicit actions are preceded by sequences of internal neural (cognitive) operations that, from the perspective of neural activity patterns and TD learning, are indistinguishable from those patterns that directly cause movement of the body in the world. High level cognitive functions are therefore simply 'internal actions'[24].
- Subcortical circuits, particularly the basal ganglia, cerebellum, brainstem and spinal cord, fully control well-trained movements and serialization of all other motor outputs. These regions may use partially different mechanisms but are still tightly integrated[28].
- Computational principles apply similarly across all of cortex. Differentiation of function occurs predominantly through structural connectivity.

To compute means to control the flow of information, and to store, recall, organize, integrate and transform information in pursuit of a defined outcome or ongoing effect. In the case of the brain, it also means to flexibly adapt to unforeseen conditions in ways that no computers can currently achieve. The brain accomplishes this by flexibly activating neural assemblies (groups of neurons) in combinatorial patterns that best represent the confluence of sensory input and current internal state. But what controls which assemblies should be active? Neural assemblies respond when they recognise (i.e. are keyed by) particular afferent spike patterns, and only when this key-match is good enough to win the multi-scale-inhibition competition. The vital insight is that neurons and assemblies respond *when they are required* and *without centralized control*. **This fortuitous outcome is the result of the brain's multiple plasticity mechanisms that bias the dynamics towards activity that is ultimately rewarded[29], while extracting features that predict reward, and maintaining activity within useful dynamical bounds.**

<u>Neural assemblies, dynamics, cognition and creativity.</u>

When an assembly becomes active in response to a recognized input, it outputs a transformation of its inputs that is shaped by self-organization and reward learning. In so doing it is performing a computation on its inputs that meets either innate (self-organized) or external (reward-bearing) criteria. Neural connections are convergent, divergent, hierarchical and re-entrant, supporting spatiotemporal activity patterns that are exquisitely intricate and inter-dependent. Assemblies couple in novel patterns in response to novel inputs, and exploit the chaotic nature of the transitions between states to form novel patterns any time, in a manner related to binding of representations, fluid intelligence and creativity[30]. However they are far from random; constrained by neural architecture and shaped by plasticity, they are finely honed to be task-specific. These patterns underlie the combinatorial computational power of the brain.

The brain does not follow a program. Brain regions do not encode packets of information which are transmitted to receiving regions for decoding and processing, and brains do not work 'despite the noise'. Due to efficient coding and stochastic sampling, what we are tempted to think of as noise is in fact the entire computation[11]. Engineering-style reductionist simplifications yield no insights into neural function. The brain is the ultimate bootstrapped physical dynamical system. A neuron simply sits and listens[31]. When it hears an incoming pattern of spikes that matches a pattern it knows, it responds with a spike of its own. That's it! Repeat this process recursively tens to trillions of times, and suddenly you have a brain controlling a body in the world or doing something else equally clever. Our challenge is to understand how this occurs. **We require a new class of theories that dispose of the simplistic stimulus-driven encode/**

transmit/decode doctrine. We must embrace the brain's inherent dynamic complexity and emergent properties, and explain how plasticity molds the dynamics to capture useful couplings across brain regions and between the brain, the body and the world. My contention is that the above principles are sufficient to meet this challenge at some level, in order to both better understand the brain and to construct better brain-inspired AI.

Brains to AI.

The differences between brains and ANNs lead to significant concrete differences in capabilities:

- AI is difficult to train and typically requires huge amounts of clean labelled data. Due to its ability to combine dynamic assemblies into novel patterns, the brain implements transfer learning by chunking, often requiring just a handful of training examples.
- AI systems have enormous energy requirements for training and operation. Due to sparse efficient spike-time coding, the brain runs on less than the power of a light bulb.
- AI systems need to be pre-trained and any new information typically requires complete re-training. The brain learns online continuously using transfer learning and dynamical processes (activity patterns) for integrating that knowledge into long-term networks.
- AI is terribly brittle and can be easily fooled by adversarial input that needs to be shifted only slightly outside the training distribution. Brains generalize exceptionally well due to modular self-organization, predictive feedback and transient combinatorial dynamics.
- AI systems can perform only the task for which they are trained. Due to its ability to dynamically reconfigure through oscillations and internal actions, the brain can perform multiple tasks and switch between them as required.

What we currently call AI is fundamentally still big data and correlation analysis, predominantly used to generate classifications and occasionally predictions. There are exceptions to this rule[32]; interestingly these exceptions tend to draw inspiration directly from the brain in order to improve on the capabilities of DL. While improvements are often achieved, the insights are applied in piecemeal fashion and many of the compelling advantages of neural processing remain ignored and unharnessed. SNNs are dynamically and architecturally ideal for representing spatiotemporal patterns and for building causal models of the world. It is reasonable to expect that there will be classes of problem, relevant for our usage of AI, for which ANNs will fail in practice but that can be learned and performed efficiently by SNNs.

The true computational power of the brain lies in the synergistic integration of all the principles of neural computation. To the author's knowledge, such an integration has never been attempted at scale. Oscillations and spike-time coding have rarely, if ever, been combined with RL to flexibly route information through a neural network, for example. Further combining such a network with self-organizing plasticity could then create a network that can generalize and respond flexibly to new inputs; feedback could allow for attention to unexpected inputs; and so on. There is a feeling in the ANN community that the dynamics of spiking networks are difficult to conceptualize and control. On the contrary, it is quite possible that the synergistic combination of principles will lead to intuitive dynamics and a deeper understanding of the underlying mechanisms. **All the principles that have been discussed here are mechanistic, evidence-based, and *realistically implementable in neural circuit models and AI prototypes.* It is probable that revolutionary computational systems can be created in this way with only moderate expenditure of resources and effort.**

References

1 Brette, R. Philosophy of the spike: Rate-based vs. spike-based theories of the brain. *Frontiers in Systems Neuroscience* **9** (2015).

2 Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Current Opinion in Neurobiology* **14**, 481-487 (2004).

3 Maass, W. Noisy spiking neurons with temporal coding have more computational power than sigmoidal neurons. *Advances in Neural Information Processing Systems* **9**, 211–217 (1997).

4 Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607 (1996).

5 Zylberberg, J., Murphy, J. T. & DeWeese, M. R. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Computational Biology* **7**, e1002250 (2011).

6 Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N. & Krichmar, J. L. Neural correlates of sparse coding and dimensionality reduction. *PLoS Computational Biology* **15**, e1006908 (2019).

7 Vogels, T., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569-1573 (2011).

8 Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* **5**, 1457-1469 (2004).

9 Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76**, 695-711 (2012).

10 Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**, 79 (1999).

11 Hartmann, C., Lazar, A., Nessler, B. & Triesch, J. Where's the noise? Key features of spontaneous activity and neural variability arise through learning in a deterministic network. *PLoS Computational Biology* **11**, e1004640 (2015).

12 Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences* **9**, 474-480 (2005).

13 Izhikevich, E. M. Polychronization: Computation with spikes. *Neural Computation* **18**, 245-282 (2005).

14 Tognoli, E. & Kelso, J. The metastable brain. *Neuron* **81**, 35-48 (2014).

15 Rabinovich, M., Huerta, R. & Laurent, G. Transient dynamics for neural processing. *Science* **321**, 48-50 (2008).

16 Stratton, P. & Wiles, J. Global segregation of cortical activity and metastable dynamics. *Frontiers in Systems Neuroscience* **9**, doi:10.3389/fnsys.2015.00119 (2015).

17 Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience* **13**, 51 (2012).

18 Maass, W. On the computational power of winner-take-all. *Neural Computation* **12**, 2519-2535 (2000).

19 Szatmáry, B. & Izhikevich, E. M. Spike-timing theory of working memory. *PLoS Computational Biology* **6**, e1000879 (2010).

20 Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience* **10**, 113 (2009).

21 Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543-1546 (2008).

22 Izhikevich, E. M. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex* **17**, 2443-2452 (2007).

23 Suddendorf, T. & Corballis, M. C. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* **30**, 299-313 (2007).

24 Buzsáki, G., Peyrache, A. & Kubie, J. in *Cold Spring Harbor Symposia on Quantitative Biology.* 024679 (Cold Spring Harbor Laboratory Press, 2014).

25 Wacongne, C., Changeux, J.-P. & Dehaene, S. A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience* **32**, 3665-3678 (2012).

26 Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology* **7**, e1002211 (2011).

27 Frémaux, N., Sprekeler, H. & Gerstner, W. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Computational Biology* **9**, e1003024 (2013).

28 Schmahmann, J. D. & Caplan, D. Cognition, emotion and the cerebellum. *Brain* **129**, 290-292 (2006).

29 Savin, C. & Triesch, J. Emergence of task-dependent representations in working memory circuits. *Frontiers in Computational Neuroscience* **8** (2014).

30 Sandkühler, S. & Bhattacharya, J. Deconstructing insight: EEG correlates of insightful problem solving. *PLoS ONE* **3**, e1459 (2008).

31 Buzsáki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68**, 362-385 (2010).

32 Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245-258 (2017).