Journal of Cognitive Neuroscience
Commentary: The Entangled Brain, Luiz Pessoa (2022)

Unraveling the entangled brain: How do we go about it?

Shaul Druckmann, Department of Neurobiology, Stanford University School of Medicine & Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA USA

Nicole C. Rust, Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

An impactful understanding of the brain will require entirely new approaches and unprecedented collaborative efforts. The next steps will require brain researchers to develop theoretical frameworks that allow them to tease apart dependencies and causality in complex dynamical systems, as well as the ability to maintain awe while not getting lost in the effort. The outstanding question is: How do we go about it?

The central premise of The Entangled Brain (Pessoa, 2022) is that the brain is a highly complex system with emergent properties that cannot be understood by studying its parts in isolation. The complexity of the brain's dynamics follow from its highly interconnected and recurrent (feedback) circuits. We applaud Pessoa for drawing attention to the need to tackle this complexity to arrive at an impactful understanding of the brain - for example, to create the type of foundational understanding that can be leveraged to diagnose and treat brain dysfunction.

We support Pessoa's call for richer conceptual and theoretical frameworks that allow us to make sense of complex, recurrent distributed systems. We suspect that a rate-limiting step in brain research in the next decade, if not sooner, will be the absence of those frameworks. Most notably, we lack the frameworks to tease apart 'What causes what?' with regard to the brain and its relationship to the mind and behavior. In contrast, recent progress in biotechnology such as calcium/voltage imaging and multielectrode arrays has been rapid and has already generated rich and complex ways to record and perturb neural activity. Accordingly, much of the current bottleneck of progress in understanding causality in the brain is centered around the conceptual design of experiments and the interpretations of the resulting data (Barack et al., 2022).

Developing these new conceptual frameworks will require the community to foster, embrace and develop a diversity of ideas and new approaches. Some will be more foreign than others and we cannot let that lack of familiarity bias us away from the innovation that we so desperately need. At the same time, the space of possible explorations is infinite and we must figure out how to explore it efficiently. Complicating this exploration is the sheer awe of the human experience, which lends itself to a desire for the explanations of how it arises to be peppered with a bit of mysticism. Some (albeit not all) versions of 'emergence' tilt in that direction, and we agree with Pessoa's observation that many neuroscientists find this off-putting. We join Pessoa's call to understand and appreciate why "More is different" (Anderson, 1972), but we also emphasize the need to navigate this space rationally. We also must recognize that there may not be many easy wins. For example, we suspect that that the complexity theory frameworks that will be needed have not yet been developed.

In sum, we regard the issues that Pessoa describes in the Entangled Brain among the most outstanding and important challenges in contemporary brain research. In our minds, this begs the question: What is the best path forward? Here we describe our own vision as a complement to Pessoa's.

We suggest that in the pursuit to understand the complexity of the brain, we begin by acknowledging that the pursuit amounts to figuring out how to conceptualize, measure and describe something that we don't yet understand. One implication of this is that while we have to start somewhere, we are unlikely to define the problems in the right way from the outset. Pessoa provides an excellent example with his illustration of the problems associated with attaching psychological concepts to the brain areas involved in fear extinction (his Figure 3, left). Crucially, we can all agree that we need to move beyond examples like this, but conceptually, how do we go about it? In this, we and other brain researchers (Fried et al., 2022; Seth, 2021) are drawing inspiration from historical accounts of the development of thermometry and measuring temperature in the 17th century via "epistemic iteration" (Chang, 2007). When applied to the brain, descriptions of brain function and dysfunction and measurements of it are tied to theories about how they work, and both measurements and theories are continuously refined as new information is acquired. This process naturally allows for the replacement of less sophisticated ways of understanding with more sophisticated ones as the field evolves. The key element is being open to change.

What's next? Pessoa argues that we cannot take a reductionist approach toward understanding the brain because its functions emerge from interactions (e.g. between different parts of the brain), and that instead, we need to study it using complex systems approaches. Ultimately, we agree. At the same time, we endorse arriving at crisp definitions of the phenomenon to be explained, whenever possible, before launching into questions around how a system gives rise to it - including when that phenomenon emerges from complex, distributed processing. In many cases, this can be achieved through behavioral observations, followed by experiments that target brain areas in relative isolation. Classic examples include evidence that the brain reflects short term memory representations through activity that persists across a population (Funahashi et al., 1989; Fuster & Alexander, 1971) and that the brain arrives at decisions through the temporal integration of incoming sensory information, often reflected in the ramping of neural activity (Mazurek et al., 2003; Roitman & Shadlen, 2002). We agree that a highly reductionist approach of making inferences about how the brain does what it does by studying its parts in isolation may often fail. At the same time, we feel that rejecting the ability to recognize interesting phenomena from recordings performed in isolation when a system is known to be distributed may be a step too far. We emphasize that there is utility in approaches that focus on single elements, and do not draw on complexity theory.

Next, when we are ready to take on the question of *how?,* what is the best approach? It begins by developing theoretical frameworks targeted at understanding the complexity of recurrent, distributed systems with a practical eye to understanding brain function. Crucial in this effort is that these frameworks can be leveraged in the design and interpretation of experiments. As an illustration, we draw on a series of studies from one of our own works (S.D. and collaborators) that demonstrates this type of approach. This work was predated by the understanding that a particular brain area known as Anterior Lateral Motor cortex (ALM) is a key node in short term memory representations of action preparation in mice (Guo et al., 2014). The most central question was: how do the neural dynamics in this circuit support short term memory? We found evidence that persistent activity in ALM supports short term memory, but at the same time these dynamics depend not just on circuits confined to ALM, but are supported instead by a multi-

regional network of interactions that operate as a complex, dynamical system (Li et al., 2016). Foundational to this discovery was our development of a theoretical framework in which the population dynamics of a brain area are not viewed as a single entity (as is typically the case), but instead as a set of overlaid dynamical patterns, or factors, each with their own computational and functional meaning (Druckmann & Chklovskii, 2010, 2012). This framework allowed us to interpret the results of otherwise perplexing optogenetic brain perturbation experiments. There we determined that persistent memory activity remains robust to perturbations along the dimensions, or patterns, that are linked to short term memory, but not along other dimensions (despite the fact that these other dimensions capture substantial variance in the population response). We also determined that these short term memory representations (and behavior) were surprisingly robust to perturbation as a consequence of distributed interactions between ALM across the two hemispheres (Li et al., 2016). In sum, our understanding of how persistent memory activity arises in ALM depended crucially on the tools we had developed to think about population dynamics along multiple dimensions, and our development of those tools predated a detailed understanding of what we would ultimately use them for.

It is now more common to describe the dynamics reflected in a brain area as an interaction between two or more brain areas, described in terms of the subspaces, or patterns, that they operate upon. Others have used similar approaches to investigate, for example, motor preparation (Kaufman et al., 2014; Stavisky et al., 2017) and visual processing (Huang et al., 2019; Semedo et al., 2019; Srinath et al., 2021), as reviewed more extensively by (Gallego et al., 2017; Kohn et al., 2020; Rust & Cohen, 2022; Vyas et al., 2020).

We've only just begun to scratch the surface in thinking about complexity. What might the next steps look like? The most promising paths forward are difficult to anticipate, but we point to what we regard as some of the best ways to begin. One exceedingly good approach involves experiments that characterize population dynamics across multiple brain areas on single-trials and their relationships with behavior. The results of these experiments can be used to constrain models of how brain areas interact to create brain dynamics, and how behavior is produced from brain activity. These models can be tested for their ability to accurately predict the outcomes of experiments that perturb neural activity, as well as their ability to generalize to new behavioral conditions. A key difficulty is that these experiments often require animals to perform tasks that require relatively reduced computations, such as two alternative forced choice tasks, and the full richness of within and between circuit dynamics may not be revealed. As such, particularly interesting results may be those that defy our expectations of how circuits compute, since these mismatches can lead to advances in our conceptual frameworks. The consequences of the observed differences from our expectations can then be theoretically studied, both in the context of the task in which the experiments were performed and in more generalized tasks. In addition, the fuller richness of circuit dynamics are likely to be revealed through behaviors that are themselves dynamic. As such, establishing interpretable dynamic behaviors is a crucial direction for future research.

More broadly, among the many challenges to be faced in this overall pursuit is the extraction of principles of brain function that we can reason with and build upon, amidst all of this complexity. Given that we currently understand so little about these principles, a compelling argument can be made for investigating nearly any behavior across a broad range of species, largely absent translational considerations. Ultimately, the most clinically impactful types of understanding will be those of systems that go awry in human brain dysfunction, and going forward, we will need to carefully consider how to best draw equivalencies between behaviors and species.

Here we've sketched what we regard as a reasonable approach to tackle the vast complexity of the "entangled brain". It appreciates the fact that theories and measurements of the brain must inform one another and must both evolve together (epistemic iteration), and it emphasizes the need to develop new theoretical frameworks that capture the brain's complex, distributed recurrent networks. In our effort to build on Pessoa's contributions and constructively point to the best next paths forward, we do not underestimate the challenge that brain researchers are up against. The human brain has been described as "the most complex system humanity has ever been confronted with … by far by any metric, the most complex piece of highly organized active matter in the universe" (Koch, 2022). We acknowledge that we have only just begun to conceptualize the vast complexity of the brain. Clearly, the path forward will require new approaches and unprecedented collaborative efforts. Conceptually, we anticipate that it will proceed along the path that we have described here. What inspires us to forge ahead along this path is our intense curiosity regarding how our own brains seem to be using a computational style so alien to our understanding, about what makes us "us", as well as the vast unmet needs of people with brain dysfunction who so desperately need solutions.

Acknowledgements:

This work was supported by the Simons Collaboration on the Global Brain (award 543033 to NCR and 542969 to SD), the National Science Foundation (award 2043255 to NCR), and the McKnight Foundation (to SD).

References:

Anderson, P. W. (1972). More is different. *Science (New York, N.Y.)*, *177*(4047), 393–396. https://doi.org/10.1126/science.177.4047.393

Barack, D. L., Miller, E. K., Moore, C. I., Packer, A. M., Pessoa, L., Ross, L. N., & Rust, N. C. (2022). A call for more clarity around causality in neuroscience. *Trends in Neurosciences*, *45*(9), 654–655. https://doi.org/10.1016/j.tins.2022.06.003

Chang, H. (2007). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Druckmann, S., & Chklovskii, D. (2010). Over-complete representations on recurrent neural networks can support persistent percepts. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 23). Curran Associates, Inc.

https://proceedings.neurips.cc/paper/2010/file/39059724f73a9969845dfe4146c5660e-Paper.pdf

Druckmann, S., & Chklovskii, D. B. (2012). Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology: CB*, *22*(22), 2095–2103. https://doi.org/10.1016/j.cub.2012.08.058

Fried, E., Flake, J., & Robinaugh, D. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*, 1–11. https://doi.org/10.1038/s44159-022-00050-2

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science (New York, N.Y.)*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978–984. https://doi.org/10.1016/j.neuron.2017.05.025

Guo, Z. V., Li, N., Huber, D., Ophir, E., Gutnisky, D., Ting, J. T., Feng, G., & Svoboda, K. (2014). Flow of cortical activity underlying a tactile decision in mice. *Neuron*, *81*(1), 179–194. https://doi.org/10.1016/j.neuron.2013.10.020

Huang, C., Ruff, D. A., Pyle, R., Rosenbaum, R., Cohen, M. R., & Doiron, B. (2019). Circuit Models of Low-Dimensional Shared Variability in Cortical Networks. *Neuron*, *101*(2), 337-348.e4. https://doi.org/10.1016/j.neuron.2018.11.034

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: Permitting preparation without movement. *Nature Neuroscience*, *17*(3), 440–448. https://doi.org/10.1038/nn.3643

Koch, C. (2022). *Lab Notes | Why don't we understand the brain?* [Interview]. https://alleninstitute.org/news-press/articles/lab-notes-why-dont-we-understand-brain

Kohn, A., Jasper, A. I., Semedo, J. D., Gokcen, E., Machens, C. K., & Yu, B. M. (2020). Principles of Corticocortical Communication: Proposed Schemes and Design Considerations. *Trends in Neurosciences*, *43*(9), 725–737. https://doi.org/10.1016/j.tins.2020.07.001

Li, N., Daie, K., Svoboda, K., & Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, *532*(7600), 459–464. https://doi.org/10.1038/nature17643

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex (New York, N.Y.: 1991)*, *13*(11), 1257–1269. https://doi.org/10.1093/cercor/bhg097

Pessoa, L. (2022). The Entangled Brain. *MIT Press*. https://mitpress.mit.edu/9780262544603/the-entangled-brain/

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *22*(21), 9475–9489.

Rust, N. C., & Cohen, M. R. (2022). Priority coding in the visual system. *Nature Reviews. Neuroscience*, *23*(6), 376–388. https://doi.org/10.1038/s41583-022-00582-9

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. *Neuron*, *102*(1), 249-259.e4. https://doi.org/10.1016/j.neuron.2019.01.026

Seth, A. (2021). *Being You*.

Srinath, R., Ruff, D. A., & Cohen, M. R. (2021). *Attention improves information flow between neuronal populations without changing the communication subspace* (p. 2021.03.31.437940). https://doi.org/10.1101/2021.03.31.437940

Stavisky, S. D., Kao, J. C., Ryu, S. I., & Shenoy, K. V. (2017). Motor Cortical Visuomotor Feedback Activity Is Initially Isolated from Downstream Targets in Output-Null Neural

State Space Dimensions. *Neuron*, *95*(1), 195-208.e9. https://doi.org/10.1016/j.neuron.2017.05.023

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, *43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115