Population-Level Representations and Computations Afford Causal Explanations

The neural population doctrine has emerged as a central framework for explaining cognitive capacities with the advent of large-scale neural recording techniques (Barack & Krakauer, 2021; Ebitz & Hayden, 2021; Yuste, 2015). A prominent approach models population activity in a state space, where low-dimensional patterns of activity called manifolds represent the task-relevant variables and the computations are performed by the dynamics of the population state over the manifold (Vyas et al., 2020). Despite the prominent role of these models in explaining cognitive capacities, their explanatory status has received little philosophical attention.

Because of the high-level and mathematical nature of these models, they may appear as mere phenomenological models that only compactly describe the phenomenon. In this talk, I argue to the contrary that they afford proper explanations by capturing difference-makers at the population-level. Thus, I show how representational vehicles and computations specified at the level of the population may provide causal explanations of cognitive capacities.

First, **I argue for an interventionist approach to support population-level representational vehicles and computations as difference-makers**. A common way to demonstrate that a model is explanatory is to provide a mapping onto the parts of the system that play causal roles. To this end, a popular approach relies on decomposition and localisation (Bechtel & Richardson, 1993) and has been used to show how dynamical models can be explanatory (Bechtel & Abrahamsen, 2010; Kaplan & Craver, 2011). I argue that this approach fails to accommodate the representational vehicles which are specified at the population-level due to the mixed selectivity of many cortical neurons (Fusi et al., 2016; Hardcastle et al., 2017; Shea, 2007).

Population-level vehicles have been identified in previous work on connectionist and neural systems (Burnston, 2021; O'Brien & Opie, 2006; Shagrir, 2012), and a key challenge is to account for how population-level vehicles may play causal roles and figure in genuine representational explanations (Ramsey, 2007). This is precluded if computations are taken to occur at the level of individual neurons. To resolve this, I apply the interventionist account of causation (Woodward, 2003), which provides level-neutral criteria for assessing whether population-level variables are difference-makers for cognitive capacities.

Next, **I argue that dynamical models of population activity specify population-level difference-makers for explanations of cognitive capacities**. Analysing two examples from the empirical literature (Mante et al., 2013; Sohn et al., 2019), I argue that the dynamical

models satisfy the interventionist criteria for causal explanation. Specifically, they provide invariant counterfactual dependencies between the computation performed by the network and interventions on both the input and the population-level variables. Moreover, I argue that these population-level variables are genuine representational vehicles because we can intervene on their particular content such that a counterfactual relation additionally exist between the content of the vehicles and the output of the system (Ramsey, 2007; Schulte, 2023). I conclude, therefore, that the dynamical models identify population-level difference-makers that support answers to what-if-things-had-been-different questions that figure in genuine representational explanations.

A worry that may occur is that because the variables—the explanans—are population-level properties then they can only serve to redescribe but not explain the computation, i.e., the explanandum (Chirimuuta, 2018). Against this interpretation, I argue that the explanans enable surgical and independent interventions that can be used to change specific properties of the network computations, and, moreover, that we can intervene on the computation of the network to change the explanans of the models (Sadtler et al., 2014). This is akin to the mutual manipulability criteria developed within the mechanistic framework (Craver, 2007). Consequently, I conclude that the dynamical models provide explanations of how the computations are produced on an interventionist account of causal explanation.

Finally, **I respond to objections that the models fail to satisfy the interventionist requirements for causal explanation**. One might object that the models fail to provide counterfactual dependencies that can be interpreted as outcomes of interventions. If one accepts that the models answer w-questions, these models would then provide non-causal explanations (Chirimuuta, 2018). The real causal work—the computations—might instead be placed at the level of individual neurons. In response to these objections, I first argue that the population-level variables constitute the appropriate level of explanation before demonstrating that they satisfy the criteria for intervention.

An important part in developing causal explanations is to find stable difference-makers that are proportional to the explanandum. I argue those considerations favour the low-dimensional manifolds and dynamics provided by dynamical models (Woodward, 2021). First, these models are stable across changes in background conditions, including behavioural states and animals (e.g., Chaudhuri et al., 2019; Nieh et al., 2021). Second, the population-level variables are invariant across changes in the neural population and capture the appropriate

contrastive focus with the explanandum (Gallego et al., 2020; Woodward, 2008). There are therefore good reasons to take seriously the population-level variables as difference-makers.

In evaluating whether they are targets for interventions, I first consider what variables need to be held fixed when assessing causal claims (Shapiro & Sober, 2007; Woodward, 2015). On this basis, I argue that the causal powers of the population-level variables are not excluded by their underlying realisation base. Moreover, I argue that since the variables can be intervened on independently to change the explanandum while keeping other variables fixed, the criteria for unconfounded interventions are satisfied (cf. Woodward, 2025). Second, I argue that the information required to answer w-questions is available exclusively at the population-level and can be exploited with precise methods in experimental practice (Vinograd et al., 2024). I therefore conclude that dynamical models of population activity satisfy the interventionist criteria for causal explanations and thus explain cognitive capacities through population-level difference-makers.

References

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), Article 6.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A, Computation and Cognitive Science*, *41*(3), 321–333.

Bechtel, W., & Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* MIT Press.

Burnston, D. C. (2021). Contents, vehicles, and complex data analysis in neuroscience. *Synthese*, *199*(1), 1617–1639.

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., & Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, *22*(9), 1512–1520.

Chirimuuta, M. (2018). Explanation in Computational Neuroscience: Causal and Non-causal. *The British Journal for the Philosophy of Science*, *69*(3), 849–880.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology, Neurobiology of Cognitive Behavior*, *37*, 66–74.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., & Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, *23*(2), 260–270.

Hardcastle, K., Maheswaranathan, N., Ganguli, S., & Giocomo, L. M. (2017). A Multiplexed, Heterogeneous, and Adaptive Code for Navigation in Medial Entorhinal Cortex. *Neuron*, *94*(2), 375-387.e7.

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*. *Philosophy of Science*, *78*(4), 601–627.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature*, *595*(7865), Article 7865.

O'Brien, G., & Opie, J. (2006). How do connectionist networks compute? *Cognitive Processing*, *7*(1), 30–41.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., & Batista, A. P. (2014). Neural constraints on learning. *Nature*, *512*(7515), 423–426.

Schulte, P. (2023). *Mental Content*. Cambridge University Press.

Shagrir, O. (2012). Structural Representations and the Brain. *The British Journal for the Philosophy of Science*, *63*(3), 519–545.

Shapiro, L., & Sober, E. (2007). *Epiphenomenalism – the Do's and the Don'ts*.

Shea, N. (2007). Content and Its Vehicles in Connectionist Systems. *Mind & Language*, *22*(3), 246–269.

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. *Neuron*, *103*(5), 934-947.e5.

Vinograd, A., Nair, A., Kim, J. H., Linderman, S. W., & Anderson, D. J. (2024). Causal evidence of a line attractor encoding an affective state. *Nature*, 1–9.

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, *43*(1), 249–275.

Woodward, J. (2008). Mental Causation and Neural Mechanisms. In J. Hohwy & J. Kallestrup (Eds), *Being Reduced: New Essays on Reduction, Explanation, and Causation* (p. 0). Oxford University Press.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347.

Woodward, J. (2021). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese*, *198*(1), 237–265.

Woodward, J. (2025). Networks, dynamics and explanation. *Synthese*, *205*(5), 204.

Woodward, J. F. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), Article 8.