Forthcoming in the Journal of Cognitive Neuroscience

Not Every Thing Must Go

Trey Boone, Nina Van Rooy, and Felipe De Brigard*

Duke University

*Corresponding author: felipe.debrigard@duke.edu

Abstract:

In *The Entangled Brain*, Pessoa criticizes standard approaches in cognitive neuroscience in which the brain is seen as a functionally decomposable, modular system with causal operations built up hierarchically. Instead, he advocates for an emergentist perspective whereby dynamic brain networks are associated, not with traditional psychological categories, but with behavioral functions characterized in evolutionary terms. Here, we raise a number of concerns with such a radical approach. We ultimately believe that while much revision to cognitive neuroscience is welcome and needed, Pessoa's more radical proposals may be counterproductive.

Keywords: Functions; Decomposition; Modularity; Brain Networks; Cognitive Ontology

Forthcoming in the Journal of Cognitive Neuroscience

Not Every Thing Must Go

In *The Entangled Brain*, Pessoa advances a bold, everything-must-go framework for cognitive neuroscience. He takes a pessimistic attitude toward standard approaches that seek to understand the brain as a modular system with causal operations built up hierarchically. Ultimately, he argues that we should abandon traditional psychological categories, explanatory strategies that rely on functional decomposition, and even the notions of causality that support those strategies. In their place, Pessoa offers an emergentist perspective that seeks to align dynamic assemblies of neurons and neural regions with psychological categories drawn from an evolutionary perspective on behavior.

While there is much to like in both the negative and positive theses Pessoa defends, we will focus on some points where we feel additional elaboration is needed. We will also argue that Pessoa's more radical proposals may be counterproductive; the evidence he draws on does not demonstrate failures of decompositional approaches to functions but instead points to issues with how we characterize those functions. Specifically, it highlights shortcomings of our mesoscale theoretical understanding of the brain and may also illuminate inadequacies with the higher-level kinds we use to characterize complex behaviors. But functional decomposition remains our best bet for building up this theoretical scaffolding at intermediate scales, and there are costs to jettisoning traditional psychological categories that may be too steep to pay.

Briefly, Pessoa's central claim is that the brain is an interactionally complex system that fails to be near-decomposable. In other words, the functions of neural systems are radically context-sensitive and the networks that support behavior cannot be productively broken down into functional subcomponents. To support these claims, Pessoa cites the massive anatomical interconnectivity of different brain regions and the highly distributed nature of functional interactions between them.

His example of fear extinction helps bring his argument into clearer focus. The amygdala has long been regarded as a central structure involved in fear learning (Hitchcock and Davis, 1986). Early work on fear extinction was thus guided by the search for a structure that inhibits amygdala activity. Research in the 1990s demonstrated that medial prefrontal cortex (mPFC) may play such a role, regulating the amygdala during extinction of a conditioned fear response (Morgan et al., 1993). With additional contributions from thalamus, relaying sensory information, and hippocampus, providing information about environmental context, one can generate a simplistic model of fear extinction according to which sensory (thalamus) and contextual (hippocampus) information trigger an emotion response (amygdala), which cognitive control (mPFC) then suppresses (see Pessoa's Fig. 3a).

Pessoa observes that we now know such a model to be overly simplistic (Fig. 3b). Not only are many additional neural regions implicated in fear extinction, but also many of those areas are reciprocally connected. In particular, he points to feedback connections between basolateral amygdala and mPFC to cast doubt on the idea that either region can be straightforwardly regarded as upstream or downstream of the other. This lack of a clear direction of causal influence introduces the possibility that it is not individual regions, but instead coordinated activity across multiple regions, that performs the functions relevant to fear extinction. This case thus ostensibly illustrates Pessoa's notion of a Type II network, which resides at the heart of his concept of the entangled brain, where functions are instantiated across multiple nodes within a network, and multi-region functions cannot be well characterized in terms of the functions of subregions alone.

But what does it mean for multi-region functions to fail to be characterized in terms of the functions of subregions? Pessoa's proposal seems to be that such failures will occur when the behavior of the system as a whole cannot be *predicted* from the behaviors of subcomponents. But prediction and explanation are distinct epistemic aims (Craver, 2014; Salmon, 1978; cf. Douglas, 2009); and predictively useful models need not have mechanistic interpretations, even to guide interventions (Gessell et al., 2020). Traditionally, the epistemic aim of functional decomposition resides squarely with explanation, not prediction (Cummins, 1975). Moreover, even in completely modular, engineered systems it is often difficult to predict how changes to some subcomponent will ultimately affect the behavior of the system. Hence detailed diagnostic tests may be required to determine why a car is malfunctioning, even when that malfunction can be fully explained through functional decomposition. Thus, the absence of predictive relationships between subcomponent functions and whole system functions is lacking as a criterion distinguishing Type I and Type II networks.

A second, related point is that the sources of complex brain connectivity Pessoa cites need not be incompatible with productive notions of modularity and functional segregation, including recent algorithmic strategies to segregate large-scale networks into partially non-overlapping subsets of highly interconnected nodes (Stanley & De Brigard, 2016; Stanley, Gessell, & De Brigard, 2019). Moreover, it is possible for reciprocally connected systems to be productively analyzed through functional decomposition (Boone, forthcoming). Consider the basic thermoregulatory mechanisms involved in shivering. The preoptic area (POA) of the hypothalamus integrates input from core and peripheral thermoreceptors. As body temperature decreases, cold thermoreceptive signals ascend to the POA, disinhibiting an efferent pathway to the ventral horn (Nakamura & Morrison, 2011). The consequent excitation of ventral horn neurons causes rapid, involuntary contractions of skeletal muscles. Those contractions in turn expend chemical energy, generating heat. The resultant increase in body temperature is then fed back to hypothalamus, and when normal body temperature is restored, inhibition of the descending pathway to the ventral horn is likewise restored and the involuntary muscle contractions cease.

A fuller picture of thermoregulation of course becomes much more complex. For instance, concurrent warming mechanisms, also modulated by the POA, effect heat conservation through vasoconstriction and thermogenesis in brown adipose tissue, both of which are caused by norepinephrine release from postganglionic neurons in the sympathetic nervous system (Cannon & Nedergaard, 2004; Mohammed et al., 2018; Morrison et al., 2008). And this is to say nothing of the cooling pathways regulated by other subnuclei in the POA that generate sweating and vasodilation (Shibasaki & Crandall, 2010; Tan et al., 2016).

Critically, all these pathways are integrated and ultimately contained within a closed loop of processes that dynamically engage and disengage to maintain stable body temperature. Changes in afferent thermoreceptive signals trigger regulatory mechanisms in the hypothalamus, which cause changes in afferent thermoreceptive signals, and around it goes. So, it would seem this is exactly the sort of system that should exemplify the principles Pessoa advocates. And yet, our understanding of thermoregulatory mechanisms and their interactions has evolved precisely through systematic iterations of functional decomposition. The reciprocal causal relationships within the system do not undermine decomposition, and subcomponents within the system seem to perform well-defined, modular causal-role functions.

One obvious way Pessoa could respond is to point out that thermoregulation is autonomic and so will be fundamentally different from the processes involved in cognition. But it is unclear why we should expect that the nervous system would adopt entirely different organizing principles to support cognitive behaviors rather than scaffolding off existing functional motifs. Relatedly, such a response requires bright lines separating autonomic from cognitive processes, and it is unclear that such lines really exist, especially when analyzing the nervous system at finer grains. And finally, much of the account Pessoa is advancing pushes us toward understanding the brain as a complex, evolved biological system. As such, drawing stark lines between cognitive systems and autonomic systems, which do have well-defined evolutionary functions, feels somewhat out of step with the general thrust of his project.

Now, it is noteworthy that the functions we attribute to the hypothalamic circuits involved in thermoregulation are characterized more locally and internally, and not with reference to the global homeostatic functions they support. For instance, at finer grains, we encounter such functional attributions as GABAergic neurons in the medial preoptic area disinhibiting projections from dorsomedial hypothalamus to the rostral raphe pallidus nucleus of the medulla oblongata (Nakamura & Morrison, 2011). Similar attributions can be found in the context of fear extinction – for instance with dopaminergic neurons in the ventral tegmental area suppressing feedforward inhibition in basolateral amygdala (Salinas-Hernandez & Duvarci, 2021).

It would be quite surprising if the reciprocal connections between medial PFC and basolateral amygdala do not ultimately give way to similar analyses (McGarry & Carter, 2016). Indeed, Pessoa seems to anticipate a response along these lines in his discussion of the granularity of functions. But we think he dismisses the possibility and its implications too hastily. Functional decomposition generally proceeds by breaking functions down into much simpler processes, and it is a sort of fallacy of division to require that whole system functions be able to be read straightforwardly off the functions of subcomponents.[1] One could certainly make the case that cognitive neuroscientists often fall prey to such a fallacy. To this end, our suggestions here actually align with some of the central concerns we take Pessoa to be raising. Namely, we agree that we should not assume neural systems will always be decomposable into functional subunits that directly relate to behavior. Moreover, we agree that we should give up any insistence that neural functions must have straightforward conceptual interpretations to productively explain such behaviors. But we see these less as problems with the general strategy of functional decomposition and more as shortcomings with the mesoscale concepts we employ to attribute functions to subsystems within the brain.

It is a *big, open* question whether we can find appropriate mesoscale concepts to allow us to seamlessly bridge the gap between purely local, internal functions and the global, external functions we associate with cognitive behaviors. The success or failure of this project ultimately depends not only on the explanatory strategies we adopt, but also on the concepts we develop to characterize cognitive functions – i.e., our cognitive ontology. Cognitive ontology revision has been a hot topic in recent years among both philosophers and cognitive neuroscientists (Anderson, 2015; McCaffrey and Wright, 2022; Poldrack, 2010), and the account Pessoa has developed will undoubtedly be a productive focal point for additional debate on this issue.

But, again, his proposal is radical: we should jettison our familiar psychological categories – e.g., emotion, cognition, memory, attention – in favor of evolutionarily inspired, complex behaviors (Pessoa et al., 2021). We see two concerns with this proposal. First, completely giving up our familiar psychological categories risks losing track of the phenomena cognitive neuroscience is supposed to explain (De Brigard, 2015). A lesson from philosophical attempts at eliminativism in the 1980s is that replacing folk psychological categories that have been with us for centuries is easier said than done. The idea that a novel ontology will not suffer from the same flaws that plague our good-old-fashioned one is little more than a promissory note.

A second concern is that there is tension between the kinds of networks Pessoa argues are the basic functional units in the brain and the idea that these networks perform evolutionary functions. Unlike other biological structures that have resulted from the combined forces of evolutionary pressure, topological characterizations of neural networks very seldom reveal structures that can straightforwardly be understood as the product of genetic selection. We are owed a story of how precisely these biological structures came to be selected for. Such a story is especially critical given Pessoa's desire to reconceive

[1] The fallacy of division consists in assuming that something true of a whole must be true of some or all of its parts (Hansen, 2020).

cognitive ontology in terms of evolutionary functions, and its absence makes that proposal difficult to evaluate.

Pessoa's book will no doubt be a fulcrum for future work aiming to disentangle the complexities of the brain and its relationship to behavior. We have tried to target a few areas of the account presented in his precis that merit additional expansion and clarification. Time will ultimately tell whether and to what extent revolutionary measures are needed for cognitive neuroscience to progress. But it is unclear that the field has, at this point, stalled to such an extent that these drastic measures are necessary. We agree that revisions are needed, but rather than adopting entirely new explanatory strategies and giving up on our familiar psychological categories, we conclude with our title: not every thing must go.

Acknowledgments

This work was supported by a National Science Foundation award to F.D.B. (FAIN 2218556).

References

Anderson, M. (2015). Mining the brain for a new taxonomy of the mind. *Philosophy Compass.* https://doi.org/10.1111/phc3.12155

Boone, W. (forthcoming). Robustness and Modularity. *British Journal for Philosophy of Science*. https://doi.org/10.1086/715627

Cannon, B., & Nedergaard, J. (2004). Brown adipose tissue: function and physiological significance. *Physiological reviews*, 84(1), 277–359. https://doi.org/10.1152/physrev.00015.2003

Craver, C.F. (2014). The Ontic Account of Scientific Explanation. In: Kaiser, M.I., Scholz, O.R., Plenge, D., Hüttemann, A. (eds) *Explanation in the Special Sciences*. Synthese Library, vol 367. Springer, Dordrecht.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy*, 72(20), 741-765.

De Brigard, F. (2015). What Was I Thinking? Dennett's Content and Consciousness and the Reality of Propositional Attitudes. In C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*, *Studies in Brain and Mind* 7. DOI 10.1007/978-3-319-17374-0_3

Douglas, H. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, 76(4), 444-463. https://doi:10.1086/648111

Gessell, B., Stanley, M., Geib, B. & De Brigard, F. (2020). Prediction and Topological Models in Neuroscience. In Fabrizio Calzavarini & Marco Viola (eds.), *Neural Mechanisms: New challenges in the philosophy of neuroscience*. Springer.

Hansen, Hans, "Fallacies", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/fallacies/>.

Hitchcock J, Davis M. Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behav. Neurosci*. 1986; 100: 11–22.

McCaffrey, J. and Wright, J. (2022). Neuroscience and Cognitive Ontology: A Case for Pluralism. In F. De Brigard and W. Sinnott-Armstrong (Eds.) *Neuroscience and Philosophy*. (pp. 427-266). MIT Press.

McGarry, L. M., & Carter, A. G. (2016). Inhibitory Gating of Basolateral Amygdala Inputs to the Prefrontal Cortex. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 36(36), 9391–9406. https://doi.org/10.1523/JNEUROSCI.0874-16.2016

Mohammed, M., Madden, C., Burchiel, K., & Morrison, S. (2018). Preoptic area cooling increases the sympathetic outflow to brown adipose tissue and brown adipose tissue thermogenesis. *The American Journal of Physiology - Regulatory Integrative and Comparative Physiology*. 315(4): R609-R618.. 10.1152/ajpregu.00113.2018.

Morgan M., Romanski L., LeDoux J. Extinction of emotional learning: contribution of medial prefrontal cortex. *Neurosci. Lett*. 1993; 163:109–13.

Morrison, S. F., Nakamura, K., & Madden, C. J. (2008). Central control of thermogenesis in mammals. *Experimental physiology*, 93(7), 773–797. https://doi.org/10.1113/expphysiol.2007.041848

Morrison, S. F., & Nakamura, K. (2011). Central neural pathways for thermoregulation. *Frontiers in Bioscience* (Landmark edition), 16(1), 74–104. https://doi.org/10.2741/3677

Pessoa, L., Medina, L., & Desfilis, E. (2021). Refocusing neuroscience: moving away from mental categories and towards complex behaviours. *Philosophical Transactions of the Royal Society, B*, 377. https://doi.org/10.1098/rstb.2020.0534

Poldrack, R. (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? *Perspectives on Psychological Science*, 5(6) 753–761. https://doi.org/10.1177/1745691610388777

Salinas-Hernández, X. I., & Duvarci, S. (2021). Dopamine in Fear Extinction. *Frontiers in Synaptic Neuroscience*, 13, 635879. https://doi.org/10.3389/fnsyn.2021.635879

Salmon, W. C. (1978). Why Ask, "Why?"? An Inquiry concerning Scientific Explanation. *Proceedings and Addresses of the American Philosophical Association*, 51(6), 683–705. https://doi.org/10.2307/3129654

Shibasaki, M., & Crandall, C. G. (2010). Mechanisms and controllers of eccrine sweating in humans. Frontiers in bioscience (Scholar edition), 2(2), 685–696. https://doi.org/10.2741/s94

Stanley, M., & Brigard, F. (2016). Modularity in network neuroscience and neural reuse. *Behavioral and Brain Sciences*, 39, E133. doi:10.1017/S0140525X15001673

Stanley, M.L., Gessell, B.S., & De Brigard, F. (2019). Network modularity as a foundation for neural reuse. *Philosophy of Science*. 86(1): 23-46

Tan, C. L., Cooke, E. K., Leib, D. E., Lin, Y. C., Daly, G. E., Zimmerman, C. A., & Knight, Z. A. (2016). Warm-Sensitive Neurons that Control Body Temperature. *Cell*, 167(1), 47–59.e15. https://doi.org/10.1016/j.cell.2016.08.028