HHS Public Access

Author manuscript

J Cogn Neurosci. Author manuscript; available in PMC 2024 April 16.

Published in final edited form as:

J Cogn Neurosci. 2023 March 01; 35(3): 349–360. doi:10.1162/jocn_a_01908.

The Entangled Brain

Luiz Pessoa

University of Maryland

Abstract

The Entangled Brain (Pessoa, L., 2002. MIT Press) promotes the idea that we need to understand the brain as a complex, entangled system. Why does the complex systems perspective, one that entails emergent properties, matter for brain science? In fact, many neuroscientists consider these ideas a distraction. We discuss three principles of brain organization that inform the question of the interactional complexity of the brain: (1) massive combinatorial anatomical connectivity; (2) highly distributed functional coordination; and (3) networks/circuits as functional units. To motivate the challenges of mapping structure and function, we discuss neural circuits illustrating the high anatomical and functional interactional complexity typical in the brain. We discuss potential avenues for testing for network-level properties, including those relying on distributed computations across multiple regions. We discuss implications for brain science, including the need to characterize decentralized and heterarchical anatomical–functional organization. The view advocated has important implications for causation, too, because traditional accounts of causality provide poor candidates for explanation in interactionally complex systems like the brain given the distributed, mutual, and reciprocal nature of the interactions. Ultimately, to make progress understanding how the brain supports complex mental functions, we need to dissolve boundaries within the brain—those suggested to be associated with perception, cognition, action, emotion, motivation—as well as outside the brain, as we bring down the walls between biology, psychology, mathematics, computer science, philosophy, and so on.

INTRODUCTION

Neuroscience tends to study parts of the brain separately. *The Entangled Brain* (Pessoa, 2022b) promotes the idea that, instead, we need to understand the brain as a complex, entangled system. Accordingly, the business of a brain region needs to be situated in the context of multiregion circuits: What does a brain region do “in combination” with other areas? In a sense, when one discusses regions R_1, …, R_4 as part of some function, the decision to “not” discuss other areas is fairly arbitrary. We could have discussed the roles of regions R_5, R_6, and so on. One of the main reasons we don’t is due to the limitations of the

Reprint requests should be sent to Luiz Pessoa, University of Maryland, College Park, MD 20742, or via pessoa@umd.edu.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article’s gender citation balance.

tools available to neuroscientists, which are ill-suited to investigating large-scale, distributed systems (although techniques are advancing fast). As a result, we still do not know much about collective computations involving larger numbers of gray matter components.

The word “entangled” conjures multiple interrelated ideas but is not intended to suggest something like threads that are mixed together but can be separated given enough time. The meaning is closer to “integrated,” but single words do not do justice to the general theme permeating the book—for example, cars are highly integrated systems but are designed with parts with well-defined functions. Instead, the sense of “entangled” is one in which brain parts dynamically assemble into coalitions that support complex cognitive–emotional behaviors, coalitions composed of parts that jointly do their job. Thus, an entangled system is a deeply context-dependent one in which the function of parts (such as a brain region, or a population of cells within a region) must be understood in terms of other parts: an interactionally complex system (as described below).

In this piece, I summarize a few of the key themes of the argument built in *The Entangled Brain*, including that brain functions need to be understood as “emergent properties.” Of course, this is not a new idea. However, neuroscientists still study and explain brain functions in a way that does not heed this assertion. What is more, new generations of students learn about the nervous system in a piecemeal fashion as if processes were fairly localizable—if not in areas, at least in relatively simple networks. Therefore, revisiting these issues is valuable for students of the brain at all levels of expertise. (N.B.: Citations in the text that follows are only illustrative, and in no way seek to be representative. It is my hope that specific work can be given proper credit in the ensuing discussions initiated by this piece.)

WHAT EMERGES?

The prevailing modus operandi of science can be summarized as explaining phenomena by reducing them to an interplay of elementary units that can be investigated independently of one another (Von Bertalanffy, 1950). Such a “reductionistic approach” reached its zenith, perhaps, with the success of chemistry and particle physics in the 20th century. In the present century, its power is clearly evidenced by dramatic progress in molecular biology and genetics. At its root, this attitude to science “resolves all natural phenomena into a play of elementary units, the characteristics of which remain unaltered whether they are investigated in isolation or in a complex” (Von Bertalanffy, 1950, p. 135).

Of course, the reductionistic framework is not the only game in town, as everyone knows that “the whole is greater than the sum of its parts.” Scientists study objects that have many components that interact in manifold ways, so figuring out the parts is not enough—or so the saying implies. Again, in the words of one of early proponents of complex systems, Von Bertalanffy, it is necessary to investigate “not only parts but also relations of organization resulting from a dynamic interaction” leading to “the difference in behavior of parts in isolation and in the whole organism” (Von Bertalanffy, 1950, p. 135). But what does it mean to say that parts behave differently in isolation relative to when they are part of a system?

Enter “emergence,” a term originally coined in the 1870s to describe instances in chemistry and physiology where new and unpredictable properties appear that are not clearly ascribable to the elements from which they arise. For example, when amino acids organize themselves into a protein, the protein can carry out enzymatic functions that the amino acids on their own cannot. More importantly, they behave differently as part of the protein than they would on their own. But it is actually more than that. The dynamics of the system (the protein) closes off some of the behaviors that would be open to the components (amino acids) were they not captured by the overall system (Juarrero, 1999). Once folded up into a protein, the amino acids find their activity regulated—one sense in which they behave differently.

A possible definition of emergence is as follows: a novel, collective property that is observed when multiple elements interact that is not readily reducible to the function of the elements alone. Both scientifically and philosophically speaking, the friction caused by the idea of emergence arises because it is actually unclear what precisely emerges. For example, what is it about amino acids as part of proteins that differs from free floating ones? The question revolves around the exact status of emergent properties. Philosophers refer to this question as the “ontological” status of emergence, that is, one concerning the “proper existence” of the higher-level properties. Do emergent properties point to the existence of new laws that are not present at the lower level? Is something fundamentally irreducible at stake? As the philosopher Alicia Juarrero (1999) says, it is particularly intriguing when “systems exhibit organized and apparently novel properties, seemingly emergent characteristics that should be predictable in principle, but are not in fact” (p. 6). These types of question remain by-and-large unsolved and subject to vigorous intellectual battles (an excellent treatment is provided by Humphreys, 2016; see also Juarrero, 1999).

Fortunately, we do not need to crack the problem and can instead use “lower” and “higher” levels pragmatically when they are epistemically useful—when the theoretical stance advances knowledge. To provide an oversimplified example, we do not need to worry about the status between quarks and aerodynamics. Massive airplanes are of course made of matter, which are agglomerations of elementary particles such as quarks. But when engineers design a new airplane, they consider the laws of aerodynamics, the study of the motion of air, and particularly the behavior of a solid object, such as an airplane wing, in air—they need no training at all in particle physics! So, there is no need to agonize about the “true” relationship between aerodynamics and particle physics (e.g., can the former be reduced to the latter?). The practical thing to do is simply to study the former.

One could object to this example because the inherent levels of particle physics and aerodynamics are very far removed, one level too micro and the other too macro. More interesting cases present themselves when the constituent parts and the higher-level objects are closer to each other, for example, the behavior of an individual ant and the collective behavior of the ant colony, the flight of a pelican and the V-shape pattern of the flock, or amino acids and proteins. And of course, such is the case of the brain.

THE BRAIN AS A COMPLEX SYSTEM?

Why does the complex systems perspective matter for brain science? In fact, many renowned neuroscientists consider the issues above a distraction. For example:

> [A]lthough network properties of a system are a convenient explanation for complex responses, they tell us little about how they actually work, and the concept tends to stifle exploration for more parsimonious explanations…[For example, the] highly interconnected nature of the central autonomic control system has for many years served as an impediment to assigning responsibility for specific autonomic patterns [to particular groups of neurons]. (Saper, 2002, p. 460)

Under this view, treating the brain as a complex system is not only a temporary distraction but also an actual impediment to progress.

One scenario that justifies the quote's stance is if we consider the brain to be a "near-decomposable" system. Herbert Simon (1962) proposed that scientists are frequently interested in systems exhibiting "near-complete decomposability" (see also Bechtel & Richardson, 2010), where intrasystem interactions are much stronger than extrasystem ones. Engineered systems work this way, and much research in neuroscience—lesion work in neuropsychology, systems neuroscience, fMRI research, and so on—proceeds from this vantage point. Such systems are "interactionally simple," with parts interacting weakly with anything considered beyond the system's boundaries, under a given, "reasonable" decomposition. A trivial example is a rock, where the atoms within the rock interact strongly but weakly with atoms elsewhere. In contrast, a system is "interactionally complex" to the extent that its elements (under a certain decomposition) cross its boundaries in important ways. Of course, systems exhibit a spectrum of interactivity, from low to high (for in-depth analysis, see Wimsatt, 2007, Chap. 9).

Biologists of the brain, indeed academics across varied disciplines, have been repeatedly turned off by some of the putative mystical features of emergent properties. With this in mind, it might be advantageous to switch the language used to, hopefully, stimulate debate along more productive directions centered around "system interactivity." The question of interest then shifts toward dissecting the types of communication and interactivity we find in the nervous system. Temporarily, at least, it might be productive to have "emergence" and "complex systems" recede into the background (hopefully in way that does not amount to a pure semantic sleight of hand).

The question we face is thus the following: What kind of interactional system is the brain?

PRINCIPLES OF BRAIN ORGANIZATION

To address this question, consider three principles of brain organization: (1) massive combinatorial anatomical connectivity, (2) highly distributed functional coordination, and (3) networks/circuits as functional units.

Massive Combinatorial Anatomical Connectivity

Anatomical pathways are dominated by short-distance connections. In fact, 70% of all the projections to a given locus on the cortical sheet arise from within 1.5–2.5 mm (Markov et al., 2011). Does this not dictate that processing in the brain is local, or quasilocal?

Computational analyses of anatomical cortical pathways gathered from a large number of studies inform this question. Studies initially suggested that the cortex operates as a "small-world" (Sporns & Zwi, 2004), an organization that supports enhanced signal propagation speed and synchronizability between parts, among other properties (Barabási & Albert, 1999; Watts & Strogatz, 1998). In small-world networks, though most of the connectivity is local, a modest amount of long-range random connections suffices to endow networks with these properties. Arguably, the most important insight of these analyses is not that the brain really follows a small-world organization; after all, biological systems would not be expected to exhibit "random" nonlocal connections (that's a mathematical construct!). Instead, the key idea is that it's possible for a system with mostly local physical connections, but some mid- and long-range connections, to display "unexpected" large-scale system properties.

Indeed, cortical organization is not small-world. First, nonlocal pathways are not random and instead target a "core of regions." Although different arrangements have been proposed, they indicate that cortical signals flow via a relatively small subset of richly interconnected and integrated areas, at times called a "rich club." For example, Markov et al. (2013) identified a small subset of areas in temporal cortex, parietal cortex, frontal cortex, and pFC that are very highly connected structurally. It is thus likely that brain communication relies heavily on signals being communicated via a core (Figure 1A). Second, and surprisingly, experiments indicate that cortical regions are considerably more interconnected than previously believed. Some estimates are that 60% of the possible connections between pairs of areas are indeed observed in some cortical patches (Markov et al., 2013)—clearly not a small-world organization! The precise implications of these findings, if confirmed, must also consider pathway strength (not only the existence vs. absence of a connection), which varies over several orders of magnitude, because computational work demonstrates that the type of network organization (is it small-world?) strongly depends on the pattern of pathway strengths (Gallos, Makse, & Sigman, 2012).

Cortical connectivity, although important, is only one ingredient contributing to the anatomical organization of the CNS. In fact, the focus on cortical connections of most of the computational work neglects major connectional properties that shape the overall neuroarchitecture (for a comprehensive treatment, see Nieuwenhuys, Voogd, & van Huijzen, 2008). (1) The entire cortical sheet projects to the striatum and loops back to the cortex via the thalamus, forming the so-called BG–cortical loops (Figure 1B). (2) The cortex and the thalamus are massively interconnected. Most of the thalamic volume is involved in bidirectional circuits with the cortex via the so-called higher-order regions (Sherman & Guillery, 2002). For example, the pulvinar nucleus is bidirectionally connected to the entire cortical sheet. (3) The hypothalamus is frequently viewed as a "descending" controller of autonomic functions. However, the mammalian cerebral cortex and the hypothalamus share massive "bidirectional" connections. In particular, in rodents, there are direct hypothalamic

projections to all parts of the cortical sheet (as well as multiple indirect connectivity systems with cortex; Risold, Thompson, & Swanson, 1997). In primates, the hypothalamus has widespread projections to all sectors of the pFC, including lateral sectors. (4) The basolateral amygdala (BLA) is bidirectionally connected with the entire cortical sheet; these connections are quite substantial with parts of frontal and temporal cortex, leading to the suggestion that this amygdala sector be called the "frontotemporal amygdala" (Swanson & Petrovich, 1998). (5) The cerebellum not only receives inputs from broad swaths of the cerebral cortex but also projects to many, if not all, of these areas. In particular, a significant portion of the output from the dentate nucleus of the cerebellum projects to nonmotor areas, including regions of pFC and posterior parietal cortex (Bostan, Dum, & Strick, 2013). Other major connectivity systems involve the claustrum, the septum, and the brainstem (Nieuwenhuys et al., 2008).

Clearly, a more complete elucidation of the properties of the connectional neuroarchitecture requires combining both cortical and noncortical pathway systems. The overall picture is one of massive interconnectivity, leading to "combinatorial" pathways between sectors. In other words, one can go from point A to point B in a multitude of ways. We propose that, combined, connectivity systems spanning the entire neuroaxis (cortical forebrain, subcortical forebrain, midbrain, and hindbrain) provide the basis for both broadcasting and integration of diverse signals linked to the external and internal worlds. Such crisscrossing connectional systems support the interaction and integration of signals that are typically associated with standard mental domains, including emotion, motivation, perception, cognition, and action (Pessoa, 2013) but, critically, in a manner that does not abide by putative boundaries between these categories (see below). I propose that this general architecture supports a degree of computational flexibility that enables animals to cope successfully with complex and ever-changing environments. The overall architecture may produce circuits with local specificity while attaining large-scale sensitivity, a type of "global-within-local design," which likely contributes to more sophisticated, plastic, and context-sensitive behaviors (Pessoa, Medina, & Desfilis, 2022).

Highly Distributed Functional Coordination

The complexity of anatomical pathways allows signals to flow across the brain in a staggeringly large set of ways. Anatomy provides a backbone that constrains function, but the structure–function relationship is anything but simple when one considers the abundance of bidirectional connections and loop-like organization (as in the BG), combined with excitation, inhibition, and nonlinearities. In this manner, the anatomy supports a large range of "functional interactions," namely, particular relationships between signals in disparate parts of the brain (e.g., they might fire coherently). For one, the anatomy will support the efficient communication of signals, even when strong direct pathways are absent, such as the functional coordination between signals in the amygdala and lateral pFC, although the two are not strongly connected physically. These ideas, of course, are related to the notion of functional connectivity, which in its most basic form can be indexed via the correlation coefficient between two time series (e.g., of the amygdala and the lateral pFC).

As an illustration of functional interactions, consider an experiment that acquired fMRI in monkeys when they were not performing an explicit task (Grayson et al., 2016). The study observed robust signal correlation between the amygdala and several regions that are not connected to it (as far as it is known). They asked, too, whether functional connectivity was more related to direct (monosynaptic) pathways or multipath (polysynaptic) connectivity by undertaking graph analysis. Are there efficient routes of travel between regions even when they are not directly connected? To address this question formally, they estimated a graph measure called "communicability" (related to the concept of "efficiency") and found that amygdala functional connectivity was more closely related to communicability than would be expected by considering only monosynaptic pathways. Their finding illustrates that the relationship between signals in disparate parts of the brain is not determined by structural pathways in a straightforward manner.

Networks/Circuits as Functional Units

The combination of the prior two principles leads to the present one. In a highly interconnected system, to understand function, we need to shift away from thinking in terms of individual brain regions: The network itself is the functional unit, not the brain area (Figure 2A). Processes that support behavior are not implemented by an individual area but depend on the interaction of multiple areas, which are dynamically recruited into multiregion assemblies. Such functional networks are based on the relationships between signals across disparate parts of the brain.

But how are networks/circuits defined? Let us consider here large-scale networks, such as those studied with fMRI in humans and rodents (Grandjean et al., 2020; Yeo et al., 2011). (Other examples of networks/circuits will be discussed in the context of extinction learning below.) The most popular partitioning schemes parse individual elements (brain regions or parcels) into unique groupings—a node belongs to one and exactly one community. (A community refers to a subdivision of a larger network, namely, a subnetwork. At times we will refer to subnetworks as "networks," as in "default network," given common usage in the literature.) Based on fMRI data in the absence of a task, Yeo et al. (2011) described a seven-community division of the entire cortex, where each local patch of tissue was assigned to a single community. In other words, the overall space was broken into disjoint communities. Their elegant work has been very influential, and their seven-network partition has been adopted as a sort of "canonical" division of the cortex. Whereas discrete clusters simplify the description of a system, do they capture the underlying organization?

Multiple types of networks (social, biological) exhibit nontrivial "overlapping organization" (Palla, Derenyi, Farkas, & Vicsek, 2005). For example, the study of chemical interactions reveals that a substantial fraction of proteins interacts with several protein groups, indicating that actual networks are made of interwoven sets of overlapping communities. Another way to motivate overlapping organization in networks is by considering "hub regions." Both structural and functional analyses of brain data have revealed the existence of particularly well-connected regions, called hubs. For example, as discussed above, Markov et al. (2013) described a set of areas in temporal cortex, parietal cortex, frontal cortex, and pFC that are very highly connected structurally. Regions that work as "connector hubs" (Guimera

& Nunes Amaral, 2005) are distinctly interesting because they have the potential to integrate diverse types of signals (if they receive inputs from disparate sources) and/or to distribute signals widely (if they project to disparate targets). They are a good reminder that communities are not islands; regions within a (disjoint) community have connections both within and outside the community.

Although there are many ways to operationalize overlapping networks, a simple way is to allow each brain region to participate in all communities simultaneously but in a graded fashion. Thus, if region A does not participate in community C_1, its membership value is 0; conversely, a membership value of 1 indicates that it belongs maximally to C_1. It is also useful to conceive of membership as a finite resource, such that it sums to 1. Applying these notions formally, we found that functional brain networks based on fMRI data both when tasks are not required and during task conditions are highly overlapping (Najafi, McMenamin, Simon, & Pessoa, 2016). In other words, a considerable fraction of regions shared their memberships across multiple communities. Indeed, overlapping organization has been detected via multiple networks analysis techniques (Faskowitz, Esfahlani, Jo, Sporns, & Betzel, 2020; Yeo, Krienen, Chee, & Buckner, 2014).

The brain is a dynamic, constantly moving object, and so are its networks. Functional relationships between groups of regions are constantly fluctuating based on cognitive, emotional, and motivational demands. Paying attention to a stimulus that is emotionally significant (say, paired with mild shock in the past), increases functionally connectivity between the visual cortex and the amygdala (Lojowska, Ling, Roelofs, & Hermans, 2018). Performing a challenging task in which an advance cue stimulus indicates that participants may earn extra cash for performing it correctly increases functional connectivity between the parietal/frontal cortex (important for performing the task) and the ventral striatum (important for reward-related processes; Padmala & Pessoa, 2011).

A vast literature has documented such changes in functional connectivity between pairs of regions, but distributed, large-scale changes have been observed, too (Cole et al., 2013). For example, in the reward study just described, the nucleus accumbens and the caudate each increased their functional connectivity with nearly all cortical regions engaged by the cue. Network analysis identified two communities, one cortical and another composed mostly of subcortical regions (including the nucleus accumbens and caudate). It also revealed a decrease of modularity when potential reward cues were encountered, consistent with the notion that particular conditions (in this case, the possibility of reward) reorganize large-scale functional organization (Kinnison, Padmala, Choi, & Pessoa, 2012). In another study, we observed a progression of network-level changes when participants experienced threat, uncovering how network organization unfolds across time during anxious apprehension (McMenamin, Langeslag, Sirbu, Padmala, & Pessoa, 2014), a reminder that network functional organization must be understood dynamically, as further illustrated by studies of time-varying functional connectivity (Lurie et al., 2020).

The ideas of network overlap and dynamic organization are related. If brain areas can belong to multiple networks, what determines the strength of a region's affiliation to a specific network? Here, context plays a pivotal role: region *A* will participate strongly in network

N_1 during a certain context C_1 but will be more strongly linked with network N_2 during context C_2. These ideas resonate with the "flexible hub theory" (Cole et al., 2013), where some regions are suggested to adjust their functional connectivity patterns as a function of task demands. This conceptualization brings us back to the functions of brain regions: The processes carried out by an area will depend on its network affiliations (i.e., the regions it clusters with) at a given time.

Overlap and dynamics promote a view in which networks do not consist of fixed collections of regions but instead are made of coalitions that form and dissolve to meet computational needs. In contrast, in the literature, networks frequently are described in terms of fixed sets of nodes; for example, the "salience network" might refer to nine bilateral regions plus the dorsal ACC (Hermans et al., 2011). But conceptualizing networks in more dynamic fashion is fruitful. For instance, at time t_1, regions R_1, R_2, R_7, and R_9 might form a natural cluster; at a later time t_2, regions R_2, R_7, and R_{17} might coalesce. This shift in perspective challenges the notion of a network as a stable unit, at least for longer periods of time, and raises new questions. At what point does a coalition of regions become something other than network N? Conversely, can we think of the "salience network," for example, as a set of regions that varies temporally (see Figure 2A).

THE INTERACTIONAL COMPLEXITY OF FEAR EXTINCTION

We now discuss the neural circuits of fear extinction as an example of the types of network studied by systems neuroscientists, which also illustrates the challenges of trying to unravel the mapping between structure and function. When a conditioned stimulus no longer predicts the unconditioned stimulus to which it was paired in the past (say, a sound no longer is followed by a shock), the conditioned stimulus gradually stops eliciting the conditioned response. This process is called "fear extinction." Understanding it is of potentially enormous consequence given the prevalence of anxiety and other related disorders.

The medial pFC plays an important role in regulating the amygdala during fear extinction (Morgan, Romanski, & LeDoux, 1993). Extinction critically depends on context, too. For instance, a sound may no longer signal an aversive event, but not necessarily in a completely novel environment, and the hippocampus is thought to provide such critical contextual information to the amygdala. Another region influencing extinction is the nucleus reuniens of the thalamus, which allows discrimination of dangerous from safe contexts (Ramanathan, Jin, Giustino, Payne, & Maren, 2018). At first glance, fear extinction appears to fit the scheme of separate contributions interacting to generate a new behavior: cognition (tied to the medial pFC) controlling emotion (tied to the amygdala) in a top–down fashion, with additional contributions related to the context of extinction and other factors (Figure 3A). However, such characterization does not do justice to the behavioral and neural richness of the phenomenon.

Let's consider a few additional findings about fear extinction (Figure 3B). Multiple cell groups in the BLA actually project to the medial pFC whose outputs in turn influence amygdala signals. Some studies even have suggested that the BLA is upstream of the

medial pFC, because a population of extinction neurons in the BLA (which project to the medial pFC) increase their activity during extinction learning (Herry et al., 2008). The medial pFC is also the target of the hippocampus, and this input potentiates medial pFC signals during extinction. Furthermore, the medial pFC receives substantial inputs from the thalamus, itself a major subcortical–cortical connectivity hub. Additional contributors to this circuit include the ventral tegmental area, where dopamine neurons are activated by the omission of the aversive unconditioned stimulus during extinction (Salinas-Hernández et al., 2018) and are suggested to influence the BLA (possibly via indirect projections). Although the role of the medial pFC is well established in extinction, it is likely that both the OFC and the ventrolateral pFC are important for behavioral regulation in the presence of aversive stimuli, too (Shiba, Santangelo, & Roberts, 2016). Finally, the locus coeruleus in the brainstem, which is a primary source of forebrain norepinephrine, has important neuromodulatory effects on extinction. In fact, stress-related engagement of the locus coeruleus opposes extinction (Maren, 2022). (To simplify the discussion, we described the medial pFC as a unit, but in rodents, the main contributions during extinction involve the ventral/infralimbic component, which probably corresponds to the ventromedial pFC in humans. In addition, complex microcircuits exist within critical nodes of the circuit, such as the amygdala [Whittle et al., 2021], but are not discussed here.)

Now, let us return to the initial scheme of Figure 3A. This description of extinction instantiates a boxes-and-arrows arrangement that is a mainstay of psychology and neuroscience, where semantic labels can be added to some of the interactions when their interpretation can be distilled to a convenient concept. However, the depiction is fundamentally wanting not only because it lacks some regions and connections but also because of the implicit assumption that well-defined functions are implemented by individual regions, with their outputs being read by downstream regions. For example, the hippocampus determines context, and the medial pFC some kind of appraisal that determines when the amygdala should be downregulated; hence, one can place these functions at the regions. Their outputs are then read by the amygdala to determine what to do given the inputs.

In contrast, an alternative mode of thinking considers how multiple regions jointly and dynamically implement key processes (Figure 3A). For example, as discussed above, extinction neurons in the BLA are reciprocally connected to the medial pFC (Herry et al., 2008). Thus, whereas they actually could be thought to be “upstream” of the medial pFC (thus flipping the typical way of thinking about the two regions, as indicated previously), it is important to evaluate the possibility that the two regions work in a coordinated fashion during extinction learning. Another reason fear extinction should be considered a circuit/network property is that extinction has to do with processes that convert a fear-inducing stimulus back to a status of neutrality—an “off” switch, if you will. However, as animals navigate their environment, there are many stimuli that do exactly the opposite, and they can be considered “on” switches. Accordingly, to understand complex behaviors, one needs to consider how defensive behaviors are dynamically engaged and disengaged. Even more broadly, the defensive circuits involved intersect and interact heavily with those that promote exploratory and appetitive behaviors. Should the animal withdraw, stay, or approach?

In any case, even under a more constrained conceptualization, the preceding discussion should help highlight the interactional complexity of systems-level processes neuroscientists often focus on. For one, the circuit contains both unidirectional and bidirectional connections, as well as both excitatory and inhibitory components. The case advanced here is that the field should in fact embrace this level of interactivity, not attempt to side step it. Otherwise, it will continue to be no surprise how little progress has been made in ameliorating the debilitating impacts of fear- and anxiety-related mental health conditions in the lives of so many people. Broadly speaking, extinction can be studied in the context of Pavlovian or instrumental learning (an example of the latter is avoidance learning where the animal makes a response to avoid foot shock). In both laboratory animals and humans, procedures to extinguish behavior ("instrumental extinction") elicit well-documented "side effects" (for a discussion, see Bouton, Maren, & McNally, 2021). Examples include temporary increase of the very behavior being extinguished, a return of other behaviors previously extinguished, as well as increased frequency of undesirable behaviors such as aggression. The claim made here is that these examples appear to be secondary consequences when regions are considered as well-defined investigative units with putatively specialized functions. Instead, they are exactly the types of effects routinely found in nondecomposable, interactionally complex systems, where cascades of interactions generate "side effects."

WHAT KIND OF NETWORK?

Let us consider two scenarios to further clarify what is meant by "network properties." In a Type I network, brain regions carry out (compute) fairly specific functions. For example, in the context of extinction, the hippocampus determines contextual information, and the ventral tegmental area computes omission prediction errors. In this scenario, a process of interest (say, fear extinction) is still viewed as a network property that depends on the interactions of the brain regions involved. That is to say, it is necessary to investigate the orchestration of multiple regions to understand how the regions, collectively, carry out the processes of interest. Importantly, however, the collective properties of the system are not accessible, or predictable, from the behavior of the individual regions alone: The multiregion function, $F(R_1, R_2, \ldots, R_n)$, is poorly characterized from considering $f(R_1)$, $f(R_2)$, and so on.

Poorly characterized in what sense? In a near-decomposable system, lesion of R_1, for example, will cause a deficit to the network that is directly related to the putative function of R_1. However, this is not the outcome in an interactionally complex system. Consider multispecies ecological systems in which the introduction of a new species or the removal of an existing one causes completely unexpected knock-on effects (Levine, Bascompte, Adler, & Allesina, 2017). The claim being made here is that, in many cases, we need to consider brain networks in much the same way: A complex system that is not well approximated by simple decompositions; $F(R_1, R_2, \ldots, R_n)$ will not be well approximated by considering $f(R_1)$, $f(R_2)$, …, $f(R_n)$ (Figure 2B1).

Now let us turn to Type II networks, where areas do not instantiate specific functions. Instead, two or more regions working together instantiate the basic function of interest, such that its implementation is distributed across regions. It is easy to provide an example

of Type II networks if we consider computational models where undifferentiated units are trained together to perform a function of interest. But, are there examples of this type of situation in the brain? Multiarea functions are exemplified by reciprocal dynamics between the FEFs and the lateral intraparietal area in macaques supporting persistent activity during a delayed oculomotor task (Hart & Huk, 2020). Based, among others, on the tight link between these areas at the trial level, the authors suggested that the two areas be viewed as a single functional unit (see Murray, Jaramillo, & Wang, 2017, and Kang & Drukmann, 2020, for a computational model; see also Mejías & Wang, 2022; Figure 2B2).

In rodents, motor preparation requires reciprocal excitation across multiple brain areas (Guo et al., 2017). Persistent preparatory activity cannot be sustained within cortical circuits alone but in addition requires recurrent excitation through a thalamocortical loop. Inactivation of the parts of the thalamus reciprocally connected to the frontal cortex results in strong inhibition of frontal cortex neurons. Conversely, the frontal cortex contributes major driving excitation to the higher-order thalamus in question. What is more, persistent activity in frontal cortex also requires activity in the cerebellum and vice versa (Gao et al., 2018), revealing that persistent activity during motor planning is maintained by circuits that span multiple regions. The claim, thus, is that persistent motor activity is a circuit property that requires multiple brain regions. In such case, one cannot point to a brain region (or even a sector) and label “working memory” as residing there.

It could be argued that, in the brain, the two types of networks discussed here—with and without well-defined node functions—are not really distinct and that what differs is the granularity of the function. After all, if above one could decompose the function “persistent motor activity” into basic primitives, it is conceivable that they could be carried out in separate regions. In such case, we would revert back to the situation of networks with nodes that compute well-defined functions. Put another way, a skeptic could quibble that, in the brain, a putative Type II network is a reflection of our temporary state of ignorance. The conjecture advanced here is that, in the brain, such reductive reasoning will fare poorly in the long run: It is not the case that one can develop a system of primitive properties that, together, span the functions/processes of interest. In many cases, network properties are not reducible to component interactions of well-defined subfunctions—they are inexorably distributed.

SOME IMPLICATIONS FOR BRAIN SCIENCE

In the preceding sections, we discussed one of the major themes of *The Entangled Brain*: Neural processes that accompany behavior are profitably viewed through the lens of complex, networked systems. Here, we summarize some of the implications of the framework to the general goal of elucidating brain functions and how they relate to brain parts.

Interactional Complexity

The brain is a system of interacting parts. At a local level, say within a specific Brodmann’s region or subcortical area, populations of neurons interact. But interactions are not only local. Massive anatomical connectivity provides the substrate for communication

crisscrossing the entire span of the neuroaxis (hindbrain, midbrain, and forebrain). This structural interactional complexity has important implications for brain function: Simpler decompositions that insulate brain regions from one another will capture only a slice of the contributions of the parts in question.

Anatomical interactional complexity implies that network or circuits are the functional unit of interest. This conclusion, when considered in a broad sense, may seem incontrovertible to many (most?) neuroscientists. The question then is what kind of network/circuit are we studying? I contend that simply enlarging the functional unit from an area to a standard, fixed network is only a modest step. Networks should be considered inherently overlapping and dynamic. Parts of the brain (say, populations of neurons within areas) affiliate dynamically with other elements in a highly context dependent manner driven by the current endogenous and exogenous demands and opportunities present to the animal. Critically, network properties are novel (with respect to that of individual regions), and key functions are distributed across regions or neuronal populations.

From this perspective, it is no surprise that neuroscientists are constantly discovering that brain regions participate in novel and unexpected ways in previously studied circuits and/or processes. Examples abound, but in the context of extinction learning, for example, a growing number of critical contributions of the thalamus are being discovered (Silva et al., 2021; Ramanathan et al., 2018). Finally, the inflexible nature of laboratory testing plays no small role in the apparent low interactional complexity of functional brain circuits (Paré & Quirk, 2017). By restricting the conditions under which circuits are interrogated, it appears that neuronal populations, areas, and circuits are considerably more selective for the properties and functions investigated.

Decentralization, Heterarchy, and Causation

In many systems, and the brain is no exception, it is instinctive to think that many of its important functions depend on centralized processes; for example, the pFC may be viewed as a convergence sector for multiple types of information, allowing it to control behavior. The view advanced here favors PDP (Goldman-Rakic, 1988). Instead of information flowing hierarchically to an “apex region” where signals are integrated, information travels in multiple directions without a strict hierarchy. An organization of this sort is termed a “heterarchy” to emphasize the multidirectional flow of information. As discussed previously, this does not imply an absence of organization. The anatomical backbone itself is highly structured, and several cortical regions are important anatomical/functional core regions (Markov et al., 2013). Other noncortical areas are also important hubs, including the thalamus, hypothalamus, BLA, and parts of the midbrain, including the superior colliculus.

The decentralized nature of processing should be understood temporally, too. When a novel stimulus and/or context is encountered by an animal, signals might flow first along the most direct and potent routes. However, behavior evolves temporally, and signal flow will progress in complex, decentralized ways. In fact, the spiraling pathways of the neuroarchitecture support communication and integration of signals across different spatial extents. The processing of new stimuli always take place against ongoing activity reflecting the immediate, recent past (and of course the more remote past), further decoupling

functional states from what would be anticipated by considering the most immediate anatomical pathways.

Investigating systems that are conceptualized as relatively decentralized instigates different classes of research questions about brain and behavior. If it is the coordination between the multiple parts that leads to the properties of interest, the object of scientific studies shifts to unraveling how such interactions work. For example, instead of investigating how property P is encoded in area A, the question becomes one of elucidating how the property arises from decentralized coordination. The coordination framework also moves the goalpost away from deciphering what information is passed from region to region—or relatedly how a region decodes the signals of other regions—to how the coordinated activity of multiregion assemblies generates signals with specific properties.

The view has important implications for causation, too, as the concept needs to be substantially reformulated. Neuroscientists often operate with an implicit billiard ball model of causation, a Newtonian scheme in which signals in one region affect the response in another, much like billiard balls affect each other. However, Newtonian causality provides an extremely poor candidate for explanation in interactionally complex systems like the brain because of the distributed, mutual, and reciprocal nature of the causal contributions. This is not to promote a Lashleyan view of causal equipotentiality; the brain is clearly very highly structured. So, how should one proceed?

A possible strategy to advance the understanding of causation is to investigate circuit controllability (Tang & Bassett, 2018; Liu & Barabási, 2016). By using tools from network science and mathematical control theory, one seeks to determine the extent to which certain network nodes can steer the system into different states. Thus, controllability of future system states provides a promising tool to understand the emergence of multipart properties. An interesting concept from this field is the notion of “pinning control,” where multiple inputs are applied (“pinned”) and propagate through the system, with the goal of defining the (future) trajectory of the system (Wang & Chen, 2002; Figure 4). Transplanting such reasoning to neuroscience, one could see how it could inform perturbation experiments. The ability to determine specific future states will depend on simultaneously stimulating and/or silencing sets of regions, not a single one, in particular ways.

This perspective offers avenues for testing network-level properties, too. To observe a certain function, *F*, instantiated by a certain future circuit state requires pinning multiple regions (or neuronal subpopulations); pinning a single one is insufficient to attain the collective state in question (see also Fakhar & Hilgetag, 2021, for arguments that multiregion lesion experiments are necessary). This type of approach also helps evaluate Type II networks, where function is not well defined at the area level. In such cases, manipulating two or more regions is necessary for the function in question to be instantiated (such as R_1 and R_3 in Figure 2B2). More generally, neuroscience will benefit from the development of mathematical techniques to investigate causation in complex systems, as in other areas such as weather prediction (Runge et al., 2019) and ecology (Sugihara et al., 2012).

Mental Categories and the Entangled Brain

Categories such as perception, cognition, action, emotion, and motivation organize how we understand and study brain function. But are such mental domains consistent with the framework described here? In a nutshell, no. The standard decomposition adopted by neuroscientists requires an organization that is fairly modular, which is inconsistent with the principles of the anatomical and functional neuroarchitecture discussed. In general, mental processes of interest cut across domains and do not respect putative boundaries between traditional systems (e.g., emotion, cognition). In fact, crisscrossing anatomical/functional connectional systems dissolve potential lines of demarcation.

More broadly, brains have evolved to provide adaptive responses to problems faced by living beings, promoting survival and reproduction. In this context, even the mental vocabulary of neuroscience (attention, cognitive control, etc.), with origins disconnected from the study of animal behavior, provides problematic theoretical pillars. Instead, approaches inspired by evolutionary considerations provide potentially better scaffolds to sort out the relationships between brain structure and function (Cisek, 2022; Pessoa et al., 2022).

FINAL THOUGHTS

Neuroscience strives to elucidate the neural underpinnings of behaviors and has done so in a preponderantly reductionistic fashion for over a century and a half. The time is ripe for transitioning into a period when a truly dynamic and networked view of the brain takes hold. Future research will need to strive to make progress along several fronts: dynamics, decentralized computation, and, yes, emergence.

Why do we need the perspective advocated here? The claim goes back to the interactional complexity of the brain. If some of the ideas above are correct, neuroscience needs to stop treating the brain as a near-decomposable system. Doing so distorts our view of the very system we're attempting to decipher. For example, we will think we can make progress in understanding fear and anxiety by focusing on a few regions at a time, or even isolated circuits. The contention made here is that this strategy is deficient (see Pessoa, 2022a).

How can the shift advocated be implemented? At least in part, current limitations stem from the neurotechniques available. Novel neurotechniques will play a major role. In particular, developments that allow recording over a larger number of regions simultaneously, as well as accomplishing multiregion perturbations. At the same time, a science of the mind–brain must stand on a solid foundation of understanding behavior (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017), while employing computational and mathematical tools in an integral manner. The field needs to take stock and invest on the development of conceptual and theoretical pillars. Bigger and shinier tools and techniques alone will not yield the necessary progress; we run the risk of being able to measure every cell (or subcellular component even) in the brain in a theoretical vacuum. The current obsession in the field with causation is equally problematic. Without conceptual clarity—how should we even think of causation in highly entangled systems?—causal explanations in fact might miss the point.

Ultimately, to explain the cognitive–emotional brain, we need to dissolve boundaries within the brain—perception, cognition, action, emotion, and motivation—as well as outside the brain, as we bring down the walls between biology, psychology, ecology, mathematics, computer science, philosophy, and so on.

Acknowledgments

The author is grateful for support from the National Institute of Mental Health (MH071589 and MH112517) and Brad Postle for constructive feedback on earlier versions of the article.

Funding Information

Luiz Pessoa, National Institute of Mental Health (https://dx.doi.org/10.13039/100000025), grant numbers: MH071589, MH112517.

REFERENCES

Barabási AL, & Albert R (1999). Emergence of scaling in random networks. Science, 286, 509–512. 10.1126/science.286.5439.509 [PubMed: 10521342]

Bechtel W, & Richardson RC (2010). Discovering complexity: Decomposition and localization as strategies in scientific research. Cambridge, MA: MIT Press. 10.7551/mitpress/8328.001.0001

Bostan AC, Dum RP, & Strick PL (2013). Cerebellar networks with the cerebral cortex and basal ganglia. Trends in Cognitive Sciences, 17, 241–254. 10.1016/j.tics.2013.03.003, [PubMed: 23579055]

Bouton ME, Maren S, & McNally GP (2021). Behavioral and neurobiological mechanisms of Pavlovian and instrumental extinction learning. Physiological Reviews, 101, 611–681. 10.1152/physrev.00016.2020, [PubMed: 32970967]

Cisek P (2022). Evolution of behavioural control from chordates to primates. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 377, 20200522. 10.1098/rstb.2020.0522 [PubMed: 34957850]

Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, & Braver TS (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. Nature Neuroscience, 16, 1348–1355. 10.1038/nn.3470 [PubMed: 23892552]

Fakhar K, & Hilgetag CC (2021). Systematic perturbation of an artificial neural network: A step towards quantifying causal contributions in the brain. bioRxiv. 10.1101/2021.11.04.467251

Faskowitz J, Esfahlani FZ, Jo Y, Sporns O, & Betzel RF (2020). Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. Nature Neuroscience, 23, 1644–1654. 10.1038/s41593-020-00719-y [PubMed: 33077948]

Gallos LK, Makse HA, & Sigman M (2012). A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. Proceedings of the National Academy of Sciences, U.S.A, 109, 2825–2830.10.1073/pnas.1106612109

Gao Z, Davis C, Thomas AM, Economo MN, Abrego AM, Svoboda K, et al. (2018). A cortico-cerebellar loop for motor planning. Nature, 563, 113–116. 10.1038/s41586-018-0633-x [PubMed: 30333626]

Goldman-Rakic PS (1988). Topography of cognition: Parallel distributed networks in primate association cortex. Annual Review of Neuroscience, 11, 137–156. 10.1146/annurev.ne.11.030188.001033

Grandjean J, Canella C, Anckaerts C, Ayranci G, Bougacha S, Bienert T, et al. (2020). Common functional networks in the mouse brain revealed by multi-Centre resting-state fMRI analysis. Neuroimage, 205, 116278. 10.1016/j.neuroimage.2019.116278 [PubMed: 31614221]

Grayson DS, Bliss-Moreau E, Machado CJ, Bennett J, Shen K, Grant KA, et al. (2016). The rhesus monkey connectome predicts disrupted functional networks resulting from Pharmacogenetic inactivation of the amygdala. Neuron, 91, 453–466. 10.1016/j.neuron.2016.06.005 [PubMed: 27477019]

Guimera R, & Nunes Amaral LA (2005). Functional cartography of complex metabolic networks. Nature, 433, 895–900. 10.1038/nature03288 [PubMed: 15729348]

Guo ZV, Inagaki HK, Daie K, Druckmann S, Gerfen CR, & Svoboda K (2017). Maintenance of persistent activity in a frontal thalamocortical loop. Nature, 545, 181–186. 10.1038/nature22324 [PubMed: 28467817]

Hart E, & Huk AC (2020). Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietalnetwork. eLife, 9, e52460. 10.7554/eLife.52460 [PubMed: 32379044]

Hermans EJ, Van Marle HJ, Ossewaarde L, Henckens MJ, Qin S, Van Kesteren MT, et al. (2011). Stress-related noradrenergic activity prompts large-scale neural network reconfiguration. Science, 334, 1151–1153. 10.1126/science.1209603 [PubMed: 22116887]

Herry C, Ciocchi S, Senn V, Demmou L, Müller C, & Lüthi A (2008). Switching on and off fear by distinct neuronal circuits. Nature, 454, 600–606. 10.1038/nature07166 [PubMed: 18615015]

Humphreys P (2016). Emergence: A philosophical account.Oxford University Press. 10.1093/acprof:oso/9780190620325.001.0001

Juarrero A (1999). Dynamics in action: Intentional behavior as a complex system. Cambridge, MA: MIT Press. 10.7551/mitpress/2528.001.0001

Kang B, & Druckmann S (2020). Approaches to inferring multi-regional interactions from simultaneous population recordings. Current Opinion in Neurobiology, 65, 108–119. 10.1016/j.conb.2020.10.004 [PubMed: 33227602]

Kinnison J, Padmala S, Choi JM, & Pessoa L (2012). Network analysis reveals increased integration during emotional and motivational processing. Journal of Neuroscience, 32, 8361–8372. 10.1523/JNEUROSCI.0821-12.2012 [PubMed: 22699916]

Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, & Poeppel D (2017). Neuroscience needs behavior: Correcting a reductionist bias. Neuron, 93, 480–490. 10.1016/j.neuron.2016.12.041 [PubMed: 28182904]

Levine JM, Bascompte J, Adler PB, & Allesina S (2017). Beyond pairwise mechanisms of species coexistence in complex communities. Nature, 546, 56–64. 10.1038/nature22898 [PubMed: 28569813]

Liu YY, & Barabási AL (2016). Control principles of complex systems. Reviews of Modern Physics, 88, 035006. 10.1103/RevModPhys.88.035006

Lojowska M, Ling S, Roelofs K, & Hermans EJ (2018). Visuocortical changes during a freezing-like state in humans. Neuroimage, 179, 313–325. 10.1016/j.neuroimage.2018.06.013 [PubMed: 29883732]

Lurie DJ, Kessler D, Bassett DS, Betzel RF, Breakspear M, Kheilholz S, et al. (2020). Questions and controversies in the study of time-varying functional connectivity in resting fMRI. Network Neuroscience, 4, 30–69. 10.1162/netn_a_00116 [PubMed: 32043043]

Maren S (2022). Unrelenting fear under stress: Neural circuits and mechanisms for the immediate extinction deficit. Frontiers in Systems Neuroscience, 39, 888461. 10.3389/fnsys.2022.888461

Markov NT, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Toroczkai Z, & Kennedy H (2013). Cortical high-density Counterstream architectures. Science, 342, 1238406. 10.1126/science.1238406 [PubMed: 24179228]

Markov NT, Misery P, Falchier A, Lamy C, Vezoli J, Quilodran R, et al. (2011). Weight consistency specifies regularities of macaque cortical networks. Cerebral Cortex, 21, 1254–1272. 10.1093/cercor/bhq201 [PubMed: 21045004]

McMenamin BW, Langeslag SJ, Sirbu M, Padmala S, & Pessoa L (2014). Network organization unfolds over time during periods of anxious anticipation. Journal of Neuroscience, 34, 11261–11273. 10.1523/JNEUROSCI.1579-14.2014 [PubMed: 25143607]

Mejías JF, & Wang XJ (2022). Mechanisms of distributed working memory in a large-scale network of macaqueneocortex. eLife, 11, e72136. 10.7554/eLife.72136 [PubMed: 35200137]

Morgan MA, Romanski LM, & LeDoux JE (1993). Extinction of emotional learning: Contribution of medial prefrontal cortex. Neuroscience Letters, 163, 109–113. 10.1016/0304-3940(93)90241-C [PubMed: 8295722]

Murray JD, Jaramillo J, & Wang XJ (2017). Working memory and decision-making in a frontoparietal circuit model. Journal of Neuroscience, 37, 12167–12186. 10.1523/JNEUROSCI.0343-17.2017 [PubMed: 29114071]

Najafi M, McMenamin BW, Simon JZ, & Pessoa L (2016). Overlapping communities reveal rich structure in large-scale brain networks during rest and task conditions. Neuroimage, 135, 92–106. 10.1016/j.neuroimage.2016.04.054 [PubMed: 27129758]

Nieuwenhuys R, Voogd J, & van Huijzen C (2008). The human central nervous system: A synopsis and atlas(4th ed.). Springer Science and Business Media. 10.1007/978-3-540-34686-9

Padmala S, & Pessoa L (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. Journal of Cognitive Neuroscience, 23, 3419–3432. 10.1162/jocn_a_00011 [PubMed: 21452938]

Palla G, Derenyi I, Farkas I, & Vicsek T (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814–818. 10.1038/nature03607 [PubMed: 15944704]

Paré D, & Quirk GJ (2017). When scientific paradigms lead to tunnel vision: Lessons from the study of fear. NPJ Science of Learning, 2, 1–8. 10.1038/s41539-017-0007-4 [PubMed: 30294452]

Pessoa L (2013). The cognitive–emotional brain: From interactions to integration. Cambridge, MA: MIT Press. 10.7551/mitpress/9780262019569.001.0001

Pessoa L (2022a). How many brain regions are needed to elucidate the neural bases of fear and anxiety? OSF Preprints.

Pessoa L (2022b). The entangled brain: How perception, cognition, and emotion are woven together. Cambridge, MA: MIT Press. 10.7551/mitpress/14636.001.0001

Pessoa L, Medina L, & Desfilis E (2022). Refocusing neuroscience: Moving away from mental categories and towards complex behaviours. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 377, 20200534. 10.1098/rstb.2020.0534 [PubMed: 34957851]

Ramanathan KR, Jin J, Giustino TF, Payne MR, & Maren S (2018). Prefrontal projections to the thalamic nucleus reuniens mediate fear extinction. Nature Communications, 9, 1–12. 10.1038/s41467-018-06970-z

Risold PY, Thompson RH, & Swanson LW (1997). The structural organization of connections between hypothalamus and cerebral cortex. Brain Research Reviews, 24, 197–254. 10.1016/S0165-0173(97)00007-6 [PubMed: 9385455]

Runge J, Bathiany S, Bollt E, Camps-Valls G, Coumou D, Deyle E, et al. (2019). Inferring causation from time series in earth system sciences. Nature Communications, 10, 1–13. 10.1038/s41467-019-10105-3

Salinas-Hernández XI, Vogel P, Betz S, Kalisch R, Sigurdsson T, & Duvarci S (2018). Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. eLife, 7, e38818. 10.7554/eLife.38818 [PubMed: 30421719]

Saper CB (2002). The central autonomic nervous system: Conscious visceral perception and autonomic pattern generation. Annual Review of Neuroscience, 25, 433–469. 10.1146/annurev.neuro.25.032502.111311

Sherman SM, & Guillery RW (2002). The role of the thalamus in the flow of information to the cortex. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 357, 1695–1708. 10.1098/rstb.2002.1161 [PubMed: 12626004]

Shiba Y, Santangelo AM, & Roberts AC (2016). Beyond the medial regions of prefrontal cortex in the regulation of fear and anxiety. Frontiers in Systems Neuroscience, 10, 12. 10.3389/fnsys.2016.00012 [PubMed: 26941618]

Silva BA, Astori S, Burns AM, Heiser H, van den Heuvel L, Santoni G, et al. (2021). A thalamo-amygdalar circuit underlying the extinction of remote fear memories. Nature Neuroscience, 24, 964–974. 10.1038/s41593-021-00856-y [PubMed: 34017129]

Simon HA (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106, 467–482.

Sporns O, & Zwi JD (2004). The small world of the cerebral cortex. Neuroinformatics, 2, 145–162. 10.1385/NI:2:2:145 [PubMed: 15319512]

Sugihara G, May R, Ye H, Hsieh CH, Deyle E, Fogarty M, et al. (2012). Detecting causality in complex ecosystems. Science, 338, 496–500. 10.1126/science.1227079 [PubMed: 22997134]

Swanson LW, & Petrovich GD (1998). What is the amygdala? Trends in Neurosciences, 21, 323–331. 10.1016/S0166-2236(98)01265-X [PubMed: 9720596]

Tang E, & Bassett DS (2018). Colloquium: Control of dynamics in brain networks. Reviews of Modern Physics, 90, 031003. 10.1103/RevModPhys.90.031003

Von Bertalanffy L (1950). An outline of general system theory. British Journal for the Philosophy of Science, 1, 134–165. 10.1093/bjps/I.2.134

Wang XF, & Chen G (2002). Pinning control of scale-free dynamical networks. Physica A: Statistical Mechanics and its Applications, 310, 521–531. 10.1016/S0378-4371(02)00772-0

Watts DJ, & Strogatz SH (1998). Collective dynamics of'small-world' networks. Nature, 393, 440–442. 10.1038/30918 [PubMed: 9623998]

Whittle N, Fadok J, MacPherson KP, Nguyen R, Botta P, Wolff SB, et al. (2021). Central amygdala micro-circuits mediate fear extinction. Nature Communications, 12, 1–11. 10.1038/s41467-021-24068-x

Wimsatt WC (2007). Re-engineering philosophy for limited beings: Piecewise approximations to reality. Harvard University Press. 10.2307/j.ctv1pncnrh

Yeo BTT, Krienen FM, Chee MW, & Buckner RL (2014). Estimates of segregation and overlap of functional connectivity networks in the human cerebral cortex. Neuroimage, 88, 212–227. 10.1016/j.neuroimage.2013.10.046 [PubMed: 24185018]

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, et al. (2011). The Organization of the Human Cerebral Cortex Estimated by intrinsic functional connectivity. Journal of Neurophysiology, 106, 1125–1165. 10.1152/jn.00338.2011 [PubMed: 21653723]

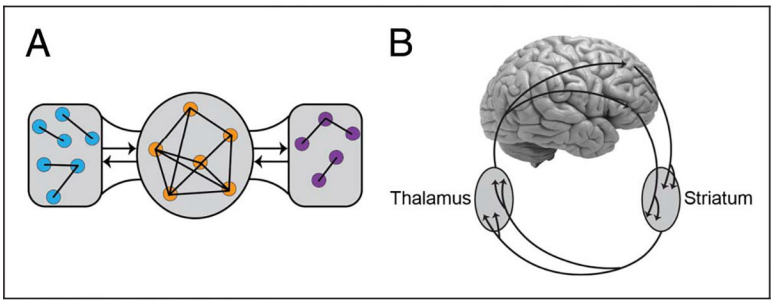

Figure 1.

Combinatorial anatomical connectivity. (A) Computational analysis of cortical pathways suggests that a subgroup of regions works as a “rich club” (orange circles in the middle): a set of highly interconnected nodes that play a major role in determining the flow of signals across the brain. (B) The neuroarchitecture also includes multiple large-scale connectional systems, such as via the BG, as illustrated here. Additional systems include those involving the thalamus, the hypothalamus, the BLA, and the cerebellum, among others.

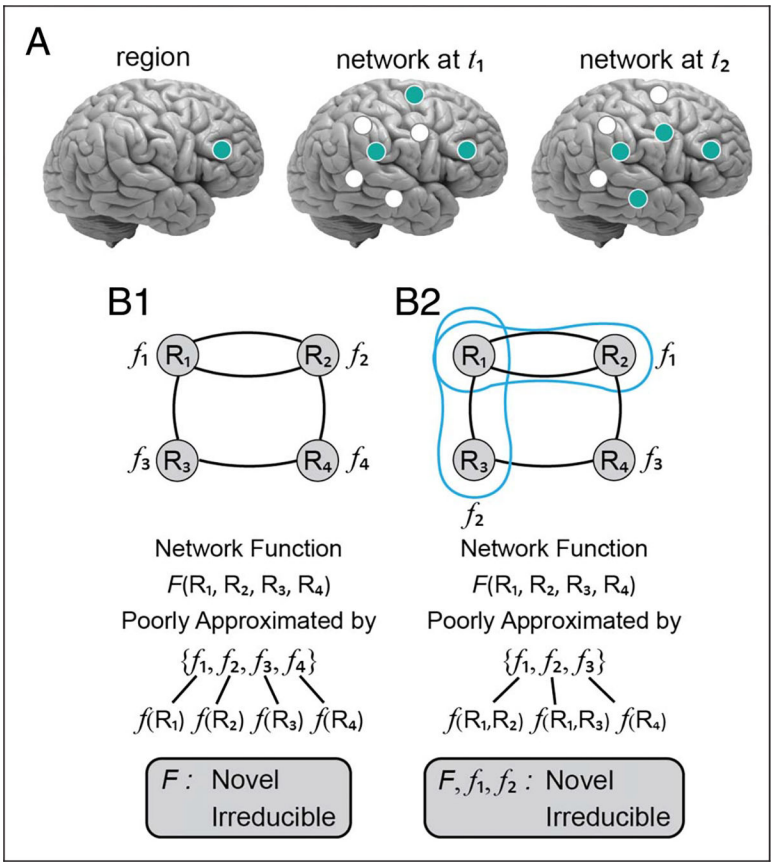

Figure 2.

From regions to networks. (A) The meaningful functional unit is not the brain region (left) but networks of brain regions that aggregate and disassemble as a function of time (middle and right). (B) Illustration of network properties (functions) in a scenario in which regions carry out well-defined "primitives" (B1) and when they do not (B2); in the latter, the function in question needs to be understood in terms a set of regions (blue contours). Note that in both cases, understanding the circuit behavior (function *F*) is not well approximated by considering the individual functions, *f*. Instead, it is necessary to consider the coordinated (emergent) circuit function. In B1, individual functions can be specified based on single regions (e.g., $f(R_1)$), but in B2 depend on more than one region in some cases (e.g., $f(R_1, R_2)$). Boxes at the bottom indicate criteria to determine network-level properties, as well as Type II networks (in B2).

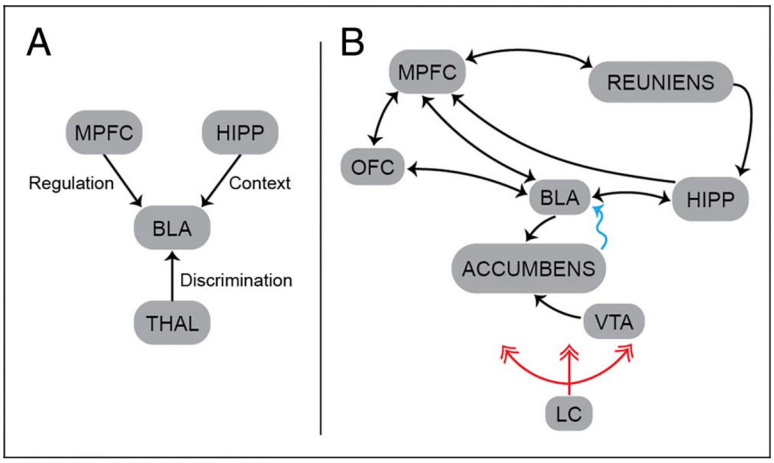

Figure 3.

Fear extinction circuits. (A) Basic extinction circuit centered on the BLA. The contributions of a few key regions are labeled with their putative functional contributions. (B) Extended circuit, with a larger set of brain regions believedtobeinvolved(notintended to becomprehensive). The blue arrows indicate indirect anatomical connectivity. The red arrows indicate the extensive norepinephric projections of the locus coeruleus. HIPP, hippocampus; LC, locus coeruleus; MPFC, medial pFC; THAL, thalamus; VTA, ventral tegmental area.

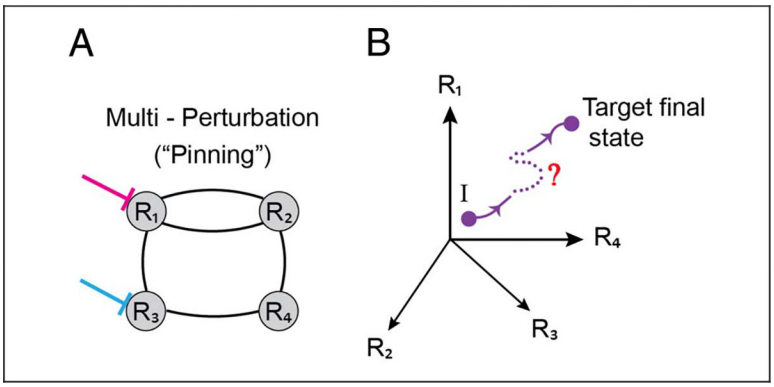

Figure 4.

Network controllability. (A) Multiperturbation methods can be used to activate and/or silence multiple brain regions simultaneously (here R_1 and R_3). (B) Such perturbations can be used to attempt to steer the trajectory of the system from some initial state, I, toward a final target state. Here, the state of the system can be represented as a point in four dimensions, each corresponding to the activity level at an individual region. The temporal evolution of the system along the state space corresponds to a trajectory.