This is a pre-print, please cite the published version
Forthcoming in Synthese

Daniel Kostić and Kareem Khalifa

Does Functional Connectivity Explain?

Abstract:

Many successful explanations show how causally individuated parts are responsible for the occurrence of the phenomena that scientists seek to explain. On this view, parts that are chosen only by convention, and related only through correlations, cannot possibly figure in successful explanations. This is because without some form of causal grounding, it seems unintelligible why any explanatory relation between these parts and the phenomenon of interest would hold. This problem is particularly pronounced in functional connectivity models (FC) in neuroscience. These models typically represent time series of recurrent neural activity in conventionally determined spatial regions (as a network's nodes) and synchronization likelihoods among these time series (as its edges). Many neuroscientists and philosophers maintain that because of this, FC models cannot provide explanations. We formulate this problem more precisely and then show that it rests on an impoverished interpretation of scientific models in general and FC models in

particular. We then provide a positive account of how FC models provide a variety of neuroscientific explanations.

1. Introduction

In the philosophical literature on topological or "network" explanations, a wide variety of debates have emerged.[1] Many of these debates have focused on whether topological explanations ever enjoy a degree of "autonomy" from causal or mechanistic explanations. While these debates have done much to expand our understanding of graph theory's uses and abuses in scientific explanations, they have perhaps obscured other interesting questions one might ask about these kinds of explanations.

In particular, graph theory can be used to model a wide variety of systems, but are some of these systems inherently unsuited for scientific explanations? This question is especially salient in neuroscientists' use of functional connectivity (FC). Some philosophers and neuroscientists contend that graph-theoretic representations of functional connectivity are at best descriptive, but not explanatory (Cole et al. 2016; Craver 2016; Friston 2011; Ramsey et al. 2010; Woodward 2025), while others are more optimistic (Kostić 2023; Kostić and Khalifa 2021; Mill, Ito, and Cole 2017; Reid et al. 2019; Smith 2012).

In this paper, we provide novel arguments for the optimists in this debate: FC models are sometimes explanatory. As we show, this isn't merely an abstract

[1] For reviews of the literature see Kostić (2023) and Rathkopf (2024).

philosophical debate; it has wider ramifications for neuroscientific practice (Section 2). Hence, it's paramount that we first formulate the perceived problem with FC explanations more precisely, which we dub the "Wrong Stuff Objection" (Section 3). This crisper formulation allows us to diagnose where the Wrong Stuff Objection missteps: it assumes an impoverished view of how scientific models are interpreted (Section 4). We then provide neuroscientific examples that more concretely establish our diagnosis (Sections 5 and 6).

2. Background

Before presenting the Wrong Stuff Objection, let's get a clearer picture of what's at stake with FC models in neuroscience. While we'll chiefly be concerned with our titular question, "Does functional connectivity explain?", our answer has implications for two related debates that bear on central issues in neuroscientific methodology.

2.1. Structure and Function

Structure-function relationships are fundamental for understanding not just neural/cognitive systems, but biological systems more broadly (Garson 2016; Novick 2023; Wouters 2003). Think, for example, of how the folding of a protein into a three-dimensional structure enables it to have a specific biological function (e.g., replicating DNA or controlling cell division). In this case, it is plausible to claim that structure explains function, because certain changes to the structure of the protein fold fundamentally change the protein's function. Structure in this context refers to the spatial-geometric features of the fold (Anfinsen and Haber 1961; Gutte and Merrifield 1971; Morange 2006; Sadowski and Jones 2009).

Ostensibly, structural and functional connectivity are supposed to correspond to something analogous in neural/cognitive systems. In structural connectivity networks, the connections between nodes are based on physical anatomical connections (such as white matter tracts) between neuronal populations. For this reason, structural connectivity networks are also called "anatomical connectivity networks."

While there is widespread consensus that structural connectivity tracks the brain's anatomical structure, it is less clear that functional connectivity tracks cognitive functions. In FC models of the brain, nodes represent time series of neural activity, which typically are obtained by measuring blood-oxygen-level-dependent (BOLD) signals in functional magnetic resonance imaging (fMRI) or electrical activity in electroencephalogram (EEG) channels. BOLD signals are attributed to voxels, which are three-dimensional counterparts to pixels. Similarly, FC models' edges represent *synchronization likelihoods*, i.e., correlations between the time series of BOLD signals or EEG channels.

As should be clear, there is no obvious reason why patterns of correlations between events in conventionally defined spatial regions in the brain would adequately represent cognitive functions. Indeed, both philosophers and neuroscientists have expressed skepticism on this front (e.g., Buckner, Krienen, and

Yeo 2013; Craver 2016). In Section 5, we show that these skeptical positions overstate their case. More precisely, we argue that skeptics fail to appreciate that FC models must be properly *interpreted* for their contributions to functional explanation[2] to become apparent.

2.2. What (if Anything) Does Functional Connectivity Represent?

Determining whether FC models explain sheds light on a closely related discussion concerning neuroscientific methodology—namely how neuroscientists are supposed to interpret FC models.

Neuroscientists and philosophers disagree about whether functional connectivity adequately represents brain/cognitive function (Bogen 2002; Friston 2011; Hutchison et al. 2013; Logothetis 2010; Siddiqi et al. 2022; Sporns 2014). A primary concern is its reliance on BOLD signals, which are indirect, temporally "sluggish" proxies of neural activity; too coarse-grained to capture fast electrophysiological processes which take place at the millisecond-scale oscillations (Logothetis, 2008, 2010). Traditional FC modeling was often based on static descriptions of neural activity, ignoring the temporal features of brain dynamics. Even in more recently developed measures of functional connectivity, some skepticism still remains particularly about their interpretability and reliability, because physiological noise such as respiration, heart rate, or head motions can

[2] For reasons we discuss below, we restrict our discussion of functional explanation to contextual-mechanistic explanations.

generate spurious correlations, and arbitrariness in choosing the time-window length can lead to autocorrelations of time series (Hutchison et al. 2013).

Functional connectivity also inherently measures statistical covariation rather than causal interactions, limiting its capacity to disentangle direct neural communication from mediated connections (Friston 2011). Additionally, FC models risk oversimplifying cognition by reducing multifaceted, context-dependent processes—governed by nonlinear, multiscale interactions—to linear correlations between regions, potentially omitting critical mechanisms underlying cognitive states (Sporns 2014).

As our examples below illustrate, these challenges can be overcome, and often, it is precisely because FC models simplify that they are explanatorily useful. Alongside this optimism, we emphasize that there is no uniform template for interpreting FC models. In some cases, they are best interpreted as part of a functional explanation; in other cases, they provide temporal information that is explanatory.

3. The Wrong Stuff Objection

On the face of it, neuroscientists do not seem especially averse to treating functional connectivity as explanatory.[3] Indeed, neuroscientists often present "FC

[3] A simple Google Scholar search suggests that the phrase "functional connectivity explains" (121 results) shows up in at least as many publications as the phrases "structural connectivity explains" (72 results) and "anatomical connectivity explains" (8 results) Similar results occur for other words typically used to describe explanatory relations, e.g., "shapes," "underlies," and "influences."

explanations” as top-line results. For example, Vasileiadi et al. (2023) title a recent article, “Functional connectivity explains how neuronavigated TMS of posterior temporal subregions differentially affect language processing.” Similarly, in their “Preserved network functional connectivity underlies cognitive reserve in multiple sclerosis,” Fuchs et al. (2019) write in their abstract that “preserved functional connectivity explains cognitive reserve in [patients with multiple sclerosis], helping to maintain cognitive capacity despite structural damage.”

However, some express skepticism about functional connectivity’s capacity to explain. For instance, Craver (2016, 705) argues:

> FC matrices are network models. They provide evidence about community structure in the brain. Community structure is relevant to brain function. But the matrices do not explain brain function. They don’t model the right kinds of stuff: the nodes aren’t working parts, and the edges are only correlations.

Woodward (2025) offers a slightly more nuanced but similar concern.[4] While never explicitly denying that FC models are explanatory, he offers an account of “one very common form” of network explanation in which dynamics figure prominently, and then denies that FC models can capture such dynamics:

> (Functional connectivity … describes temporal correlations among patterns of brain activity ...) In such cases, … the edges do not represent constraints on possible interactions. Thus, in such a case it will not make sense to specify an

[4] We note that in an earlier draft of the same paper, Woodward (2023) was more strongly committed to the Wrong Stuff Objection, claiming that functional connectivity “cannot figure in explanations.”

> independent dynamics describing a process occurring along the network. In this case there is nothing more to what the network represents than a pattern of correlation and this does not amount to an activity or process along the network for the dynamics to describe (Woodward 2025, 10).

We take Woodward's remarks to raise an important challenge. If FC models are explanatory, then we must show that either they provide topological explanations that are distinct from the "independent dynamics explanations" that are Woodward's focus or, contrary to Woodward, functional connectivity does represent dynamics.[5]

These subtleties notwithstanding, Woodward's concern overlaps substantially with Craver's. As he states elsewhere, "in a network representing functional connectivity in the brain, an edge just corresponds to the presence of a temporal correlation in activity between different brain areas, without there necessarily being any causal influence or interaction between these areas," suggesting that he takes the viability of FC explanation to require further justification (Woodward 2025, 6).

Reconstructing these remarks, we get the Wrong Stuff Objection:

WS1. All FC models only represent correlations (synchronization likelihoods) between time series of conventionally defined spatial regions (voxels, EEG channels, etc.).

[5] We meet this challenge in Section 6.3.

WS2. Models that only represent correlations between time series of conventionally defined spatial regions are not explanatory.

WS3. ∴ FC models are not explanatory.

Let's briefly motivate the premises in this argument. FC models are graph-theoretic models. As noted above, in FC models, nodes represent time series of neural activity, which typically are obtained by measuring BOLD signals in fMRI or by measuring electrical activity in EEG. BOLD signals are attributed to voxels, which are three-dimensional counterparts to pixels. Each voxel is a cubical spatial unit on a three-dimensional grid. EEG channels are locations on the skull in which EEG electrodes are attached. Electrodes are typically placed on the skull using the "10-20" system. Roughly stated, electrodes are spaced apart by 10% to 20% of the length of the skull. As such, the nodes in FC models are conventionally defined spatial regions: cubes in a grid that is "drawn" on the brain or EEG channels that have been placed in their location in accordance with the 10-20 convention. Similarly, FC models' edges represent correlations between the time series of BOLD signals or EEG channels. More precisely, these correlations are synchronization likelihoods, i.e., probabilities that pattern recurrence in one time series coincides with pattern recurrence in another time series. Hence, it's understandable why Wrong Stuffers (such as Craver and Woodward) accept the first premise of the Wrong Stuff Objection. However, as we'll argue below, this premise should ultimately be rejected.

By contrast, we will grant the Wrong Stuff Objection's other premise, that correlations between time series of conventionally defined spatial regions do not explain phenomena (WS2), for the purposes of this paper. To that end, let's briefly clarify how we're using the key term "explanation." Here is a plausible argument for this premise: (a) correlations between time series of conventionally defined spatial regions are neither *causal* nor *mechanistic* relations. Coupled with the assumption that (b) models can only be explanatory insofar as they represent causal or mechanistic relations, this would then generate an argument for the claim that models that only represent correlations between time series of conventionally defined spatial regions do not explain phenomena (WS2).

While we have argued elsewhere that graph-theoretic models can provide explanations that are neither causal (Kostić 2020, 2023) nor mechanistic (Kostić and Khalifa 2022), we will assume (a) and (b) for the sake of this paper. Doing so has two advantages. First, it puts the focus squarely on what a topological model must represent in order to be explanatory rather than on larger debates about noncausal or nonmechanistic explanations.[6] Second, if it can be shown that FC models provide causal/mechanistic explanations, then FC models will also provide explanations on more inclusive accounts that permit non-causal and non-

[6] Recall from our Introduction that we see the Wrong Stuff Objection as distinct from the debate over the "autonomy" of topological explanations from causal/mechanistic explanations. This assumption helps to highlight the distinctness of these issues.

mechanistic topological explanations. To that end, Section 5 provides an FC model that is a mechanistic explanation; Section 6, one that is a causal explanation.[7]

4. Topological Explanation and Scientific Representation

In our estimate, the Wrong Stuff Objection's first premise is false. In other words, we will argue that:

> ¬WS1. **Some** FC models **do not** *only* represent correlations between time series of conventionally defined spatial regions.

The key is paying closer attention to what it means for a model to represent its target (Frigg and Nguyen 2017, 2020). In particular, we take it to be a platitude that scientific models must be *interpreted*.[8] To a first approximation, an interpretation maps elements of a model onto elements of the target system it is supposed to represent.

However, only some kinds of interpretations beget the Wrong Stuff Objection. Call these *austere* interpretations. By contrast, *rich* interpretations undermine the Wrong Stuff Objection. They do so by providing examples of FC models that do more than represent correlations and conventionally defined spatial regions (i.e.,

[7] Importantly, the explanations in Sections 5 and 6 align substantially with Craver and Woodward's accounts of mechanistic and causal explanation, respectively.

[8] Several authors explicitly refer to interpretation as part of their analysis of scientific representation (Bueno and French 2018; Contessa 2007; Díez 2020; Frigg and Nguyen 2020; Hughes 1997). Virtually every other account has a (set of) concept(s) that fulfills the same role as interpretation in their analysis. For some of the more explicit connections drawn on this front, see e.g., Khalifa, Millson, and Risjord (2022) and Suárez (2015, 2024). Because the point about model interpretability is platitudinous, we need not commit to any particular account of scientific representation.

counterexamples to WS1). We discuss each kind of interpretation in turn, first illustrating their key points by using an example of a topological explanation involving *structural* connectivity. We do this to show that if these models were only austerely interpreted, they would lose much (if not all) of their explanatory power. This suggests that there is nothing peculiar about FC models; like structural connectivity models, their explanatory power only becomes apparent on a rich interpretation. Put more sharply: Wrong Stuffers typically take structural connectivity models to be immune to their objection, but if they held those models to the same interpretive standards to which they hold FC models, that immunity would be illusory. This suggests that the problem is with austere interpretations—not with functional connectivity.

4.1. Austere Interpretations

As we see it, Wrong Stuffers' austere interpretations of FC models have two characteristic features. First, they only indicate what nodes and edges in a graph denote, leaving additional network structure uninterpreted. For example, Woodward (2025, 10) states of functional connectivity, "there is nothing more to what the network represents than a pattern of correlation." Similarly, in his discussion of functional connectivity, Craver (2016, 704-706) discusses FC models' nodes and edges repeatedly in the ways discussed above: nodes are time series of voxels and edges are correlations. However, defenders of FC explanations very rarely appeal to nodes and edges as their primary explanatory variables. Rather, they typically appeal to more sophisticated network concepts, such as mean functional connectivity and functional hubs (see Sections 5 and 6 for further discussion.)[9]

Second, austere interpretations assume that a graph's edges represent explanatory relations. This best explains Wrong Stuffers' heavy emphasis on FC models' edges being correlations.[10] To see why, note that an explanation has exactly three elements—an *explanandum*, an *explanans*, and an explanatory relation—and that correlations can unproblematically function as *explananda* or *explanantia.* Correlations can be causal *explananda.* For example, a correlation between *X* and *Y* might be explained by a common cause *Z*. Similarly, correlations—even between time series—can also figure unproblematically in causal *explanantia.* For example, the temporal correlation of high winds and dry conditions (causally) explains the onset of wildfires. Problems only arise when correlations are alleged to play the role of *explanatory relations*—paradigmatically causal-explanatory relations—between an explanans and explanandum. As the old adage goes, "Correlation does not imply causation."

So, insofar as the Wrong Stuff Objection's foregrounding of correlations in FC models is not a red herring, its proponents appear committed to the following: (a) a

[9] In fairness to Craver (2016, 706), he briefly mentions functional hubs:

> … 'target hubs' in the cortex… are closely connected to many functional systems and have high participation coefficients. Damage to these areas produces wide-ranging deficits out of proportion to the size of the lesion.

However, this seems to concede that the Wrong Stuff Objection is unsound, for damage to a functional hub now appears to *explain* ("produce") cognitive deficits. So we cannot see how Craver (2016, 705) can consistently endorse both an austere interpretation and his earlier claim that FC models "do not explain brain function."

[10] See the quotes from Craver and Woodward above.

graph-theoretic model is explanatory only if its edges denote explanatory relations. When coupled with the claims that (b) FC models' edges are correlations and (c) correlations are not explanatory relations, the Wrong Stuff Objection's conclusion—that (d) FC models are not explanatory—follows.[11] The first claim, (a), is characteristic of austere interpretations. As we'll see, it is wildly at odds with neuroscientific practice.

Thus, when nodes and edges are the only elements of a network model that are interpreted and when edges are taken to denote explanatory relations, we have an austere interpretation. As we will now argue, if structural connectivity models were held to the same interpretive standards, they would fail to explain, too. Consider a well-known example from Watts and Strogatz (1998). They model the nervous system of *C. elegans* graph-theoretically in order to explain its relative efficiency in processing information. While they ultimately offer a rich interpretation, consider an austerely interpreted version of this model. Watts and Strogatz represent the individual neurons in *C. elegans* as nodes; edges denote synapses or gap junctions between different neurons.

However, austerely interpreted models may have complex network structures and topological properties, including small-worldness, mean connectivity, scale-freeness, connectance, and eccentricity. These models will remain austerely interpreted so long as no further interpretation is given to these network structures.

[11] Tying this back to our more general formulation of the Wrong Stuff Objection in Section 2, we can see that (a) and (c) provide an argument for WS2, WS1 entails (b), and (d) is equivalent to WS3. By challenging (a), we are technically arguing against WS2. However, as the examples in Sections 5 and 6 illustrate, this is parasitic on denying WS1—just as we stated at the beginning of this section.

Return to our example. Watts and Strogatz (1998) model the *C. elegans* neural system as a *small-world network*. These networks are characterised by two graph-theoretic variables. The first, the clustering coefficient, represents densely interconnected triplets of nodes in the same neighbourhoods of nodes. The other, average path length, concerns the number of edges that need to be traversed in order to reach any node in the network. Small-world networks have high clustering coefficient and low average path length. However, even if the nodes and edges of the *C. elegans* network are interpreted as above, this still leaves the small-worldness of the network somewhat mysterious. Merely linking neurons, synapses, and gap junctions together through some mathematical formalism does not yet seem like the kind of thing that can explain the neural system's efficiency.

Additionally, we noted above that austere interpretations also assume that one or more edges denote the explanatory relation. In this example, this would mean that one or more synapses or gap junctions between different neurons is the explanatory relation. However, Watts and Strogatz's *explanans* is the small-worldness of the *C. elegans* structural connectivity network and their *explanandum* is efficient information-processing. Neither small-worldness nor information-processing are neurons, and they clearly are not connected by synapses or gap junctions. This aspect of an austere interpretation systematically leads to category mistakes of this kind.

Something is clearly remiss with austerely interpreting structural connectivity networks in this way. To understand how small-worldness is more than a mathematical description, a further question must be asked: “How does a small-world network in which nodes are neurons and edges are synapses and gap junctions explain the efficiency of the corresponding neural system’s information processing?” Scientists transition from austere to rich interpretations by answering questions such as this.[12]

4.2. Rich Interpretations

In what follows, we discuss the rich interpretation of the *C. elegans* network to highlight three characteristic features of rich interpretations (see Table 1). While this list of features is by no means exhaustive, we think that typical rich interpretations exhibit these features, albeit with varying points of emphasis.[13]

	Austere	**Rich**
Graph-theoretic structures other than nodes and edges interpreted?	No	Yes (via, e.g., contrastive *explanantia*, qualitative reasoning, and sharpening)
Edges always interpreted as explanatory relations?	Yes	No

Table 1. Differences between austere and rich interpretations.

[12] More generally, let T be a topological property such as small-worldness, scale-freeness, etc. and assume that the set of nodes and edges in a graph are austerely interpreted in terms of objects O and relations R respectively. Furthermore, assume that the explanandum is Y. Then a rich interpretation answers the question, “How does a network with property T interpreted in terms of O and R explain Y?” Undoubtedly, there are other questions that can elicit rich interpretations as answers.

[13] As should be clear, because both austere and rich interpretations are multidimensional concepts, they lie on a continuum. We are choosing exemplars at both ends of this continuum to make our points. We hope to explore subtler forms of austerity and richness in explanatory interpretations in future work.

4.2.1. Contrastive Explanantia

It's widely thought that causal and mechanistic explanations must support counterfactuals.[14] Consequently, in the present discussion, they must show how the explanandum would have been different had the topological property in the *explanans* been different. Rich interpretations frequently specify ways that the *explanans*-property could have been different. In Watts and Strogatz's example, they contrast small-world networks with two other network structures: (1) *regular* networks, which have both high global clustering coefficients and high average path length, and (2) *random* graphs, which have low global clustering coefficients and low average path length. Recall that (3) *small-world* networks have high clustering coefficient and low average path length. Let us say that these three network structures comprise the *contrast class* of Watts and Strogatz's explanation of the *C. elegans* nervous system.[15]

Determining the contrast class sets the stage for evaluating counterfactuals such as the following:

[14] This is a necessary condition on causal/mechanistic explanations; not a sufficient one. See our brief discussion of 'toolkits' below for how rich interpretations add further information that suffice for causal explanation.

[15] This differs from van Fraassen's (1980) well-known use of the phrase "contrast class," which refers to contrastive *explananda*. Here, we use the phrase "contrast class" to refer to contrastive *explanantia*, though contrastive explananda are also suggested by our discussion.

> Had the *C. elegans* neural network been regular or random (rather than small-world), then efficiency of information transfer would have been different.

In this way, contrast classes clarify how the *explanans* can change. Furthermore, this counterfactual departs from the austere interpretation by eschewing any assumption that edges (synapses and gap junctions) denote the explanatory *relation*. On this rich interpretation, they figure instead in the *explanans,* as they are essential ingredients of the small-worldness of the *C. elegans* network.[16] As we'll now see, they enrich their interpretation in two further ways to clarify the consequent of this counterfactual.

4.2.2. Qualitative Reasoning

Of course, simply contrasting one uninterpreted mathematical structure with another does not yet yield an explanation. To that end, it is frequently useful to be able to *qualitatively reason* about these different graph-theoretic structures. Here, we think de Regt's (2017, 102) influential account of intelligibility is suggestive:

> A scientific theory *T* (in one or more of its representations) is intelligible for scientists (in context *C*) if they can recognize the qualitatively characteristic consequences of *T* without performing exact calculations.

[16] Woodward (personal communication) accepts that correlations can be *explanantia*.

For de Regt, the relevant contextual factors are both features of the theory (whether it is visualizable, simple, etc.) as well as the skills of the scientists. The idea is that when a theory's features "match" a scientist's skills, the scientist can draw the relevant qualitative consequences. Moreover, and important for our purposes, explanatory concepts—such as mechanisms, interventions, and causes—are toolkits that promote intelligibility and qualitative reasoning in different contexts (De Regt 2017, 115).

Insofar as we might quibble with de Regt's account of intelligibility, we do not see anything quite like a theory in our examples. Hence, we suggest that de Regt's account be broadened so that representations more broadly are the loci of intelligibility:

> A scientific representation *R* is intelligible for scientists (in context *C*) if they can recognize the qualitatively characteristic consequences of *R* without performing exact calculations.

Paradigmatically, these consequences will be the conclusions of so-called "surrogative inferences," i.e., inferences from a model to its target (Contessa 2007; Khalifa, Millson, and Risjord 2022; Suárez 2004, 2024). The rough idea is this: scientists draw different qualitative consequences about the explanandum from the different contrastive explanantia outlined above. Insofar as those qualitative consequences comport with true counterfactuals (such as the one later in this section), the interpretation is suggestive of a correct explanation.[17]

In Watts and Strogatz's explanation of the *C. elegans* neural system, the key qualitative concept is that of "spreading."[18] They do this by identifying distinct spreading characteristics across different network architectures. Regular networks, characterized by high clustering but long average path lengths, impose significant delays in spreading, as signals must traverse numerous intermediate connections. Random networks, despite shorter path lengths, exhibit reduced clustering coefficients that limit efficient spreading. By contrast, small-world networks represent an optimal configuration for speedy spreading. These networks maintain substantial local clustering while incorporating critical long-range connections that function as transmission shortcuts. These strategic connections effectively reduce the network's diameter by linking distant neural clusters, enabling them to communicate with near-neighbor efficiency. Small-world networks thereby enjoy a dual advantage: signals can rapidly spread across global network distances while preserving the processing benefits of dense local connectivity. The resulting propagation enhancement allows small-world networks to achieve efficient signal distribution unattainable by either regular or random networks.[19]

[17] Of course, even if it generates true counterfactual claims, the qualitative reasoning can be wildly inaccurate of the target system. Insofar as this cannot be countenanced as a proper use of idealization, approximation, etc., the interpretation is in need of revision.

[18] Our discussion of this example leans heavily on Kostić and Khalifa (2022).

[19] Some of this qualitative reasoning was developed and further motivated after the original Watts and Strogatz article (e.g., Bullmore and Sporns 2012; Latora and Marchiori 2001).

This kind of reasoning furnishes Watts and Strogatz with clearer consequents for the relevant counterfactuals. In this case:

> Had the *C. elegans* neural network been regular or random (rather than small-world), then efficiency of information transfer would have been *lower (rather than its actual level).*

While the counterfactual in 4.2.1 only indicated that changes in network structure would lead to *different* levels of efficiency in information transfer, the qualitative reasoning discussed in this section entails that those same changes would have led to *lower* levels of efficiency in information transfer. In this way, the qualitative reasoning that is characteristic of rich interpretations makes the network model more intelligible.

At this point, it seems that we have most of what is needed to explain efficient information propagation in terms of small-worldness. This is largely because small-worldness went from being a mere mathematical description of an austerely interpreted graph to having its own interpretation: as long as something can "spread" from one node to another via an edge, this qualitative interpretation of small-worldness is available as a potential explanation. This will be true regardless of whether nodes and edges are interpreted as neurons and synapses, individuals and contagion relations, beliefs and interpersonal communication, etc. Furthermore, precisely because this rich interpretation is available under a wide range of austere interpretations, it provides a useful explanatory strategy that can be applied to diverse systems.

4.2.3. Sharpening

Of course, qualitative reasoning without exact calculation faces its own liabilities—most notably imprecision. For this reason, qualitative reasoning is often supplemented with another level of mathematical or formal representation. We call this *sharpening* the rich interpretation. All else being equal, sharper rich interpretations will license finer-grained counterfactuals and thus substantially increase the kinds of things that a topological model can explain.

One of the most useful “sharpening strategies” involves identifying rules or equations that accurately describe the processes and interactions in a network that unfold over time—what are frequently called *dynamics*. Spreading in a small-world network is clearly something that can be modeled dynamically.[20] (We take it that “spreading” is already a dynamical concept, albeit a qualitative one. Dynamical equations render that concept quantitative.)

However, other kinds of assumptions can sharpen a rich interpretation. For instance, Bullmore and Sporns (2012) adopt “economic assumptions” to establish the efficiency of small-world network brain structure. Qualitatively, their idea is that small-world networks achieve a high level of adaptive value while keeping wiring costs low. Here, adaptive behavior includes information processing capacity

[20] See Woodward (2025) for further discussion.

and resilience to "adverse" perturbations. They define wiring cost as "[t]he fixed cost of making anatomical connections between neurons, often approximated by the wiring volume of anatomical connections" (Bullmore and Sporns 2012, 338). They sharpen this idea thusly: "The difference between efficiency and connection distance (each expressed as a proportion of its maximum value so as to lie in the same numerical range, 0–1) — so-called cost-efficiency — increases as a function of connection density to a maximum (when the networks are about 20% connected) and declines thereafter" (Bullmore and Sporns 2012, 344).

Taking stock, *austere* interpretations of a topological model only map its nodes onto objects and its edges onto relations between those objects, while leaving additional topological structure uninterpreted. Furthermore, austere interpretations assume that edges must represent explanatory relations. By contrast, *rich* interpretations add flesh to the graph-theoretic bones of a topological *explanans* in at least three important ways: (1) embedding it in a contrast class of other potential *explanantia*, (2) rendering it intelligible through qualitative reasoning, and (3) sharpening that qualitative reasoning through dynamical equations and other formal devices. When compared to their austere counterparts, it's far easier to see how richly interpreted models are explanatory. Indeed, rich interpretations may be necessary to establish that a topological model is not merely descriptive. As we've seen, this is largely because rich interpretations provide for more precise and easily evaluable counterfactuals. Moreover, throughout this section, we have used an example that only involves the "right stuff" according to our critics. This suggests that rich interpretations are needed not only for FC models, but for explanatory models in general.[21]

With these points in place, we can now see where the Wrong Stuff Objection goes wrong. Just as an austere interpretation of structural connectivity of the *C. elegans* neural network must be replaced by a rich interpretation in order to explain (i.e., about how small-world networks spread information more efficiently), so too must an austere interpretation of an FC model. The first premise of the Wrong Stuff Objection (WS1) states that FC models represent nothing more than edges as correlations and nodes as time series of conventionally defined spatial regions. Hence, it assumes that FC models are always austerely interpreted. However, as we'll now argue, this is unfaithful to neuroscientific practice. In what follows, we neutralize the Wrong Stuff Objection by offering counterexamples in which FC models are *richly* interpreted and provide causal and mechanistic explanations.

5. Functional Connectivity and Contextual Mechanisms

As its name suggests, functional connectivity can figure in functional explanation. While different philosophers understand “functional explanation” to refer to different kinds of explanations, we will focus on a kind of functional explanation amenable to the causal/mechanistic framework we adopted in Section 3—what Craver (2001, 2013) calls “contextual” mechanistic explanation.

[21] We suspect that when other models lack the three features associated with rich interpretations, it will be difficult to evaluate their explanatory power. This includes causal models.

We proceed as follows. We first describe a widespread use of functional connectivity in identifying different brain regions' functions (Section 5.1). We then argue that this involves richly interpreted FC models (Section 5.2). These rich interpretations, in turn, allow functional connectivity to play a central role in contextual-mechanistic explanation (Section 5.3). We then show how our analysis rebuts the Wrong Stuff Objection (Section 5.4).

5.1. Functional Identification in Neuroscience

FC models are frequently used in experiments designed to identify the functions of different brain regions. In many of these experiments, experimental subjects are asked to perform a task such that successful performance provides evidence of a specific cognitive function (e.g., working memory, language comprehension, reasoning, cognitive control). Certain areas of the brain will typically exhibit different levels of functional connectivity during the performance of the task than they did during the subjects' resting state. Resting-state functional connectivity is designed to measure brain activity when a subject is not performing any particular cognitive task. When there is sufficient evidence that two or more brain regions' change in functional connectivity is associated with correct performance of the task, those brain regions are assumed to be (partly) responsible for the corresponding cognitive function.

Neuroscientists frequently link brain structure to brain function by using probabilistic brain atlases to connect functional connectivity to structural connectivity. Brain atlases are typically three-dimensional models of the brain that include anatomical regions, functional areas, and connectivity maps. When two or more anatomical regions exhibit high functional connectivity when a subject is performing a specific task, they are interpreted as jointly contributing to the function associated with that task. As we will see below, brain atlases are central when richly interpreting FC models.

For example, consider the role of the dorsal medial system in "mentalizing," in the psychological capacity to attribute mental states to explain and predict others' actions (Andrews-Hanna, Saxe, and Yarkoni 2014). Subjects are asked to perform a false-belief task, a paradigmatic protocol for testing mentalizing capacities, within a magnetic resonance imaging chamber. In these tasks, participants must infer the false mental state of another person (in this case, a character in a short story), e.g.,

> Jenny put her chocolate in the cupboard. Then she went outside. While Jenny was outside, Alan moved her chocolate into the fridge. When Jenny returns, she will look for her chocolate in the: Fridge/Cupboard (Andrews-Hanna, Saxe, and Yarkoni 2014, 326).

The aforementioned resting state serves as the control condition. The brain regions associated with the dorsal medial system are activated during the experimental condition. These brain regions include the posterior dorsal medial prefrontal cortex, cingulate cortex, angular gyrus, temporoparietal junction, lateral temporal cortex, and anterior temporal pole. However, in *both* the experimental and control conditions, they exhibit high levels of functional connectivity. These results are interpreted as showing that the aforementioned brain regions "are organized into [a] stable functional-anatomic brain network" that plays "an important role … in reflecting on the mental states of other people" (Andrews-Hanna, Saxe, and Yarkoni 2014, 328, 331).

5.2. Neuroscientists Richly Interpret Functional Connectivity

To make sense of this widespread neuroscientific practice, rich interpretations of functional connectivity are often needed. The key qualitative concept is "activation," which is used to interpret BOLD signals in an FC model. The basic idea is simple: as brain cells become increasingly active, they require more oxygen, leading to increased blood flow and oxygenated blood in the surrounding area. A brain region that is activated when a cognitive function is being executed is taken to play a role in that function. When two regions exhibit a high degree of functional connectivity, they activate and deactivate under similar conditions. In this way, they are seen to be "functionally joined," i.e., they both contribute to the function with which they are associated.[22, 23]

So construed, functional connectivity plays a role in sharpening these qualitative concepts. A brain region's level of activation can be quantified using the strength of the BOLD signals associated with it. Two brain regions' functional jointness can similarly be quantified using functional connectivity, which is simply the magnitude of their synchronization likelihood. Because BOLD signals and functional connectivity are both quantifiable, contrast classes can include continuous, numerical levels of each. This supports the following kinds of counterfactuals:

> Had brain region *B* been less active (i.e., had a lower BOLD signal), the capacity to execute cognitive function *F* would have been diminished.

> Had brain regions *B* and *B** been less functionally conjoined (i.e., had lower FC), the capacity to execute cognitive function *F* would have been diminished.

[22] We have baptized our own term "functional joining" to minimize confusion: functional *connectivity* is operating largely as a theory-neutral term between us and purveyors of the Wrong Stuff Objection. If we put this to the side, then another way to discuss the edges in functional connectivity networks is as follows: Wrong Stuffers must claim that neuroscientists are engaged in loose talk when they describe these edges as functional connections; our interpretation takes neuroscientists' terminological choice at face value.

[23] Importantly, we make no assumptions that functionally conjoined brain regions causally influence *each other*. This is something that FC models typically cannot deliver. Our claim is only that they jointly realize a cognitive function.

Counterfactuals of the latter sort are often corroborated empirically by measuring the functional connectivity of individuals with known deficits in the cognitive function. For instance, individuals with mentalizing deficits tend to have dorsal medial subsystems with lower functional connectivity (Harenski et al. 2018; Schilbach et al. 2016).

Interestingly, sharpening in this example is quite different from the *C. elegans* example. In that example, there were additional assumptions (about spreading) needed to get the qualitative reasoning off the ground. Insofar as those additional assumptions are imprecise, sharpening is required. However, in this case, the precision/sharpening resides in what needs to be interpreted, i.e., the FC model.

Taking stock, among the most fundamental qualitative concepts associated with functional identification in neuroscience are that brain regions "activate" in order to perform a cognitive function and that multiple brain regions activate and co-active in systematic ways that "functionally conjoin" those regions. FC models sharpen these concepts to allow for finer-grained contrast classes and more testable counterfactuals.

5.3. Contextual Mechanisms

Crucially, this rich interpretation highlights the explanatory power of FC models. As noted above, we are assuming (if only for the sake of argument) that FC models are only explanatory if they are causal or mechanistic. Since Craver is one of the chief purveyors of the Wrong Stuff Objection, we use his account of how functional or "contextual" description contributes to mechanistic explanation to establish this point. According to Craver (2013, 135), "functional description can serve as a form of causal/mechanistic explanation; it is a means of situating an item in the causal structure of the world." More precisely:

> A *contextual description* of some *X*'s ϕ-ing characterizes its mechanistic role; it describes *X* (and its ϕ-ing) in terms of its contribution to a higher (+ 1) level mechanism. The description includes reference not just to *X* (and its ϕ-ing) but also to *X*'s place in the organization of *S*'s ψ-ing (Craver 2001, 63).

In the examples that interest us, *X* is a brain region *B*, ϕ-ing is either *B's* activating or functionally joining with another brain region *B**, *S* is a cognitive system/person, ψ-ing is a cognitive function such as *S's* mentalizing. Hence, even for committed mechanists such as Craver, there is little reason to deny that FC models are explanatory.

Importantly, rich interpretations of this sort rely on two conditions being satisfied:

(1) Reliable performance on the task is evidence for correct cognitive functioning. For example, reliable performance on the false-belief task is evidence of a capacity to mentalize.

(2) Functional connectivity can be mapped onto structural connectivity. For example, the clusters of voxels in the FC model are associated with the brain regions constituting the dorsal medial system.

As we've already seen, rich interpretations frequently must make assumptions such as these. Still, Wrong Stuffers might be tempted to argue that this second assumption shows that functional connectivity doesn't really explain *per se*. After all, their initial concern was that functional connectivity is not explanatory *if it is isolated from structural connectivity*.

However, this argument is unsound. Return to the two counterfactuals above. In these cases, the structural connectivity structure remains fixed (*B* and *B**) and it is the richly interpreted FC properties—varying levels of activation and functional joining—on which the cognitive function counterfactually depends. So, functional connectivity is doing the explanatory heavy-lifting. If the sense in which functional connectivity must explain "independently" of structural connectivity means that fixed structural connectivity structures cannot even be presupposed, then Wrong Stuffers' arguments are self-defeating. To see why, note that a wide array of causal/mechanistic models in neuroscience will presuppose all sorts of fixed structures (e.g., evolutionary histories, chemical structures, fundamental physics). Yet, Wrong Stuffers are not inclined to harbor skepticism about causal/mechanistic models in neuroscience simply because of these presuppositions. By parity of reasoning, they should not harbor skepticism about FC models' explanatory status because it presupposes invariant anatomical structure.

5.4. Austere Interpretations are the Wrong Stuff

Given the preceding, it is hard to see how the Wrong Stuff Objection even gets off the ground. The core mistake, as we see it, is in the austere interpretation of the nodes of an FC model. We first rehearse what that interpretation involves, then indicate where it goes wrong, and finally turn to the interpretation of edges in an FC model.

On an austere interpretation of an FC model, nodes represent nothing more than the time series of BOLD signals, which measure the amount of oxygen being used in a particular voxel. If this were all there were to the nodes of an FC model, their role of pairing brain regions with cognitive functions would be utterly mysterious. However, this overlooks the importance of *brain atlases* in this kind of inquiry. Brain atlases are reference templates that provide a standard anatomical framework for brain research. The process of brain atlas *registration* maps clusters of voxels to anatomical regions in the brain specified by such atlases. Because of this, brain atlases ensure that FC models are representing precisely the "working parts" that Wrong Stuffers insist are where proper explanations reside.

However, brain atlas registration is part of richly interpreting an FC model. Brain regions are typically more intelligible than clusters of voxels, for they more readily enable the qualitative reasoning discussed above. Indeed, the brain regions in the aforementioned counterfactuals ("*B*" and "*B**") could only be specified through the kind of mappings that brain atlas registration provides. In highlighting brain atlas registration, however, we emphasize that functional connectivity still plays an indispensable role in identifying anatomical regions with cognitive functions—as we said above, it can do the explanatory heavy-lifting. Without functional connectivity, the brain atlas would simply enumerate anatomical regions and their structural connections, but would not tell us which regions are activated during different cognitive tasks. It's precisely for this reason that functional connectivity is deserving of its "functional" moniker.

With the difference between austere and rich interpretations of FC models' nodes in place, we turn now briefly to their edges. These are synchronization likelihoods. When those are interpreted austerely as mere correlations between voxels' time series and taken to represent the explanatory *relation*, their explanatory significance is obscure. However, once a cluster of voxels can be interpreted as an anatomical region and BOLD signals as the extent to which such anatomical regions are activated while executing a function, synchronization likelihoods can be richly interpreted as indicating when two anatomical regions activate and deactivate in concert with each other while executing a given function—what we called "functional joining" above. Thus, contrary to the Wrong Stuff Objection's leading assumption (WS1), many FC models in neuroscience represent far more than correlations between conventionally defined spatial regions. Moreover, *contra* austere interpretations, the edges in FC models only figure in the *explanans*. Neuroscientists also do not confuse those correlations with causal-explanatory relations.

In sum, we have seen that rich interpretations of FC models underscore their explanatory significance. Moreover, we have used a relatively simple example—the dorsal medial system's role in mentalizing. Our arguments can be expanded to more complex examples. For instance, functional identification is performed at both higher levels of organization, e.g., the default mode network (Uddin et al. 2009), of which the dorsal medial is a subsystem. Functional identification is also performed at lower levels, e.g., the posterior cingulate cortex (Khalsa et al. 2014), which is part of the dorsal medial system. Additionally, we have used a relatively basic measure of functional connectivity, but other topological measures and concepts—most notably "functional hubs" (van den Heuvel and Sporns 2013; Tomasi and Volkow 2011a, 2011b)—also figure prominently in functional identification. The preceding suggests that most (if not all) of these functional identifications will be fodder for contextual-mechanistic explanation.

6. Explaining with Temporal Properties

FC models figure in other kinds of explanations as well. For instance, FC models contain a wealth of temporal information. Their nodes are time series and their edges are synchronization likelihoods. As examples from relativity theory, evolution, geology, and archaeology vividly illustrate, temporal information is sometimes explanatory. Hence, a suggestive idea is that richly interpreted FC models can recruit temporal information for explanatory purposes. In this section, we argue that some neuroscientific explanations work in precisely this manner. We first provide an austere and then a rich interpretation of an FC model of epileptic seizures. In doing the latter, we provide another counterexample to the Wrong Stuff Objection's assumption that FC models only represent correlations and conventionally defined spatial regions (WS1).

6.1. Austere Interpretation

Helling, Petkov, and Kalitzin (2019) offer a topological model in which mean functional connectivity explains the onset of epileptic seizure (also called *ictogenicity*). In Helling et al.'s FC model, nodes are time series of readings from EEG channels. For each EEG channel, a time series was constructed by sampling its readings several times per second. The edges in Helling et al.'s model are synchronization likelihoods between the time series data generated by two or more EEG channels. Thus, mean functional connectivity is the average strength of the synchronization likelihoods that exist between any two nodes in a functional connectivity network. Helling et al. conducted prospective studies involving subjects with focal seizures either starting treatment with an anti-epileptic drug or undergoing drug tapering over several days. They conclude that changes in mean functional connectivity explain changes in ictogenicity. As evidence, Helling et al. found that mean functional connectivity decreased for those who responded positively to their drug treatment and increased for those who responded negatively.

6.2. Rich Interpretation

If this is all that Helling et al. provided, then the interpretation is austere and the Wrong Stuff Objection appears quite plausible. However, Helling et al. offer a richer interpretation of this FC model. As noted above, rich interpretations frequently provide qualitative insight that show how interventions are possible. Qualitatively speaking, ictogenicity's dependence on mean functional connectivity suggests that "oversynchronization" of the brain explains seizures (Kalitzin et al. 2019, 7). At this point, all of the worrisome aspects of FC models—that their nodes are conventionally defined spatial regions and that their edges are correlations—become virtues. First, high mean functional connectivity can be richly interpreted as oversynchronization precisely because the edges of an FC model are *synchronization likelihoods*, which once again are the likelihoods that a pattern in one time series will coincide with a pattern in another. So, far from being mere correlations that beget the Wrong Stuff Objection, synchronization likelihoods can convey important temporal information about the brain.

Second, and more interestingly, oversynchronization is a *global* property of the brain, so the spatial positions of the individual nodes (EEG channels) do not matter. Helling et al. establish this point using a computational or "*in silico*" model to validate their hypothesis, modeling structural connectivity as a random 128-node graph with weighted edges to demonstrate the invariance of the link between mean functional connectivity and ictogenicity with respect to various anatomical factors.

This model does not represent *specific* brain regions, but simply has generic "brain units" as its nodes. Hence, the position of the EEG channels goes away in the rich interpretation. Indeed, their computational model shows that the link between mean functional connectivity and ictogenicity holds *regardless* of whether fluctuations in mean functional connectivity were driven by changes in either of the two most plausible structural factors: the connectivity strength of the anatomical network or a local tissue parameter (Helling, Petkov, and Kalitzin 2019, 4). So, it appears that very few "working parts"—to use Craver's phrase from above—matter for oversynchronization to be explanatory.

So far, we have mostly been focusing on qualitative reasoning in this rich interpretation. Additionally, the contrast class in this example is interestingly different than our previous two examples, for it consists not only of different levels of mean functional connectivity or synchronization. Helling et al. also consider alternative structures in networks that have *different* interpretations than the FC network. Prior to the simulation, it was a live option that various parameters of the structural connectivity network—especially global anatomical connectivity and the local tissue parameter—could have explained away the link between mean functional connectivity and ictogenicity. However, the simulations reveal this not to be the case. This highlights how finding a suitable contrast class involves not only *including* important foils to the explanatory factor of interest, but also *ruling out* other suggestive candidate foils.

In summary, on an austere interpretation, this FC model traffics only in EEG channels and correlations. However, this is not how the model should be interpreted for explanatory purposes. Our scientists' rich interpretation involved interpreting high mean functional connectivity as oversynchronization, as well as establishing its invariant link to ictogenicity to show that the conventional placement of EEG channels is irrelevant. Because this FC model represents a global temporal property of the brain (oversynchronization), some FC models represent more than correlations between conventionally defined spatial regions. Thus, the Wrong Stuff Objection rests on a false premise (WS1). Craver's wrong stuff is actually the right stuff for this explanation.

6.3. Causal Explanation

One might still worry that this model, even when richly interpreted, is still not explanatory. Here, we can follow Woodward (2003) in treating intervention as a mark of causation to argue that mean functional connectivity (interpreted as oversynchronization) is a *cause* of ictogenicity.[24] Helling et al. used anti-epileptic drugs to intervene on the level of brain synchronization to bring about changes in ictogenicity. Medical trials involving pills in a randomized controlled trial are often

[24] Recall that we are assuming that FC explanations must be causal or mechanistic to address the Wrong Stuff Objection.

cited as textbook examples of interventions.[25] Hence, even on the narrow view that requires all explanations to be causal, the Wrong Stuff Objection misses its mark.

We conclude our discussion of this explanation by tying it to Woodward's (2025) recent discussion of functional connectivity's explanatory power. Commenting on our earlier discussion of this explanation (2021), Woodward (2025, 18) writes:

> …functional connectivity (*E*) is an effect of (and explained by) the combination of C_1 neural dynamics and C_2 structural connectivity. Claiming that *E* can be used to explain C_1 amounts to attempting to explain a cause by appeal to one of its effects. Thus, [Kostić and Khalifa's account] misclassifies this case – we need a condition that is more discriminating. By contrast, intervening on or manipulating *E* (e.g. by altering C_2) will not change C_1, assuming as, we have been, an independent dynamics.

Thus, on "Woodward's framing", as we'll call it, mean functional connectivity is an effect (his "*E*") of neural dynamics (his "C_1") and structural connectivity (his "C_2"); see Figure 1A. However, this faces several challenges. First, even if we grant this framing, it is possible for neural dynamics and structural connectivity to explain mean functional connectivity, while mean functional connectivity explains ictogenicity, i.e., they have different explananda (Figure 1B). Second, Woodward

[25] Woodward (2003, 53, 198, 247) discusses cases with pills, as well as medical trials (Woodward 2003, 111, 127). All told, his discussion of these examples comports well with the idea that Helling et al. were intervening on MFC/synchronization.

provides no description of the "neural dynamics" operant in this example, and Helling et al.'s computer simulations indicate that oversynchronization's influence on ictogenicity is largely independent of neural dynamics and structural connectivity.

Third, if there are "neural dynamics" in this example, then they do not accord with Woodward's framing. Observe that on the rich interpretation, the causal character of this explanation is straightforward: doctors can intervene on the global synchronization of the brain to make predictable changes in patients' chances of having seizures. So, either there are neural dynamics in this rich interpretation or there are not. If, there are neural dynamics in this explanation, the level of synchronization in the brain is the most obvious candidate. However, that is just mean functional connectivity richly interpreted. So, $E = C_1$, and Woodward's suggestion that E is an effect of C_1 no longer appears plausible (Figure 1C). If, on the other hand, there are no neural dynamics in this example, then neural dynamics are not necessary for this causal explanation (Figure 1D).

We tend to favor this last interpretation, as it isn't obvious that Woodward's broader philosophical analysis of "independent dynamics explanations"[26] fits with this explanation, for at least three reasons:

[26] Woodward (2025, 17) seems especially concerned with our (2021, 2022) passing remarks that functional connectivity can explain dynamics. However, we are simply following Helling et al.'s (2019) language. Perhaps Woodward and these neuroscientists define the term "dynamics" differently. Our discussion here attempts to circumvent these verbal disputes.

(a) Woodward (2025, §2) takes networks to represent constraints on dynamics. In this example, it's not clear what, if anything, oversynchronization is "constraining."

(b) On Woodward's view, dynamics are possible only if components causally interact with each other via the relationships represented by a graph's edges. However, as Woodward (2025, 10), acknowledges, synchronization likelihoods are not interactions between components.

(c) It's possible (though not actually the case) that a system with high mean functional connectivity is one in which every node in an FC network is highly synchronized with every other. Hence, even if synchronization was a kind of interaction, it would contravene Woodward's (2025, 5) claim that networks with dynamics are "systems in which there are constraints on which components can affect others [that] contrast with systems in which each component can interact with any other."

Now, Woodward may claim that all of this strongly suggests that mean functional connectivity does *not* explain ictogenicity. However, all of Woodward's misgivings with FC models were predicated on austere interpretations. The richer alternative we have offered is one that, we repeat, offers a causal explanation replete with Woodwardian interventions. As we see it, to deny this includes undercutting claims that are either platitudinous (e.g., that models need to be interpreted in the right way in order to explain) or that are supported by the science (e.g., that high mean functional connectivity can be interpreted as oversynchronization of the brain and that there are medical interventions on oversynchronization with respect to ictogenicity). Hence, it appears that the Wrong Stuff Objection loses much of its force once rich interpretations are countenanced.

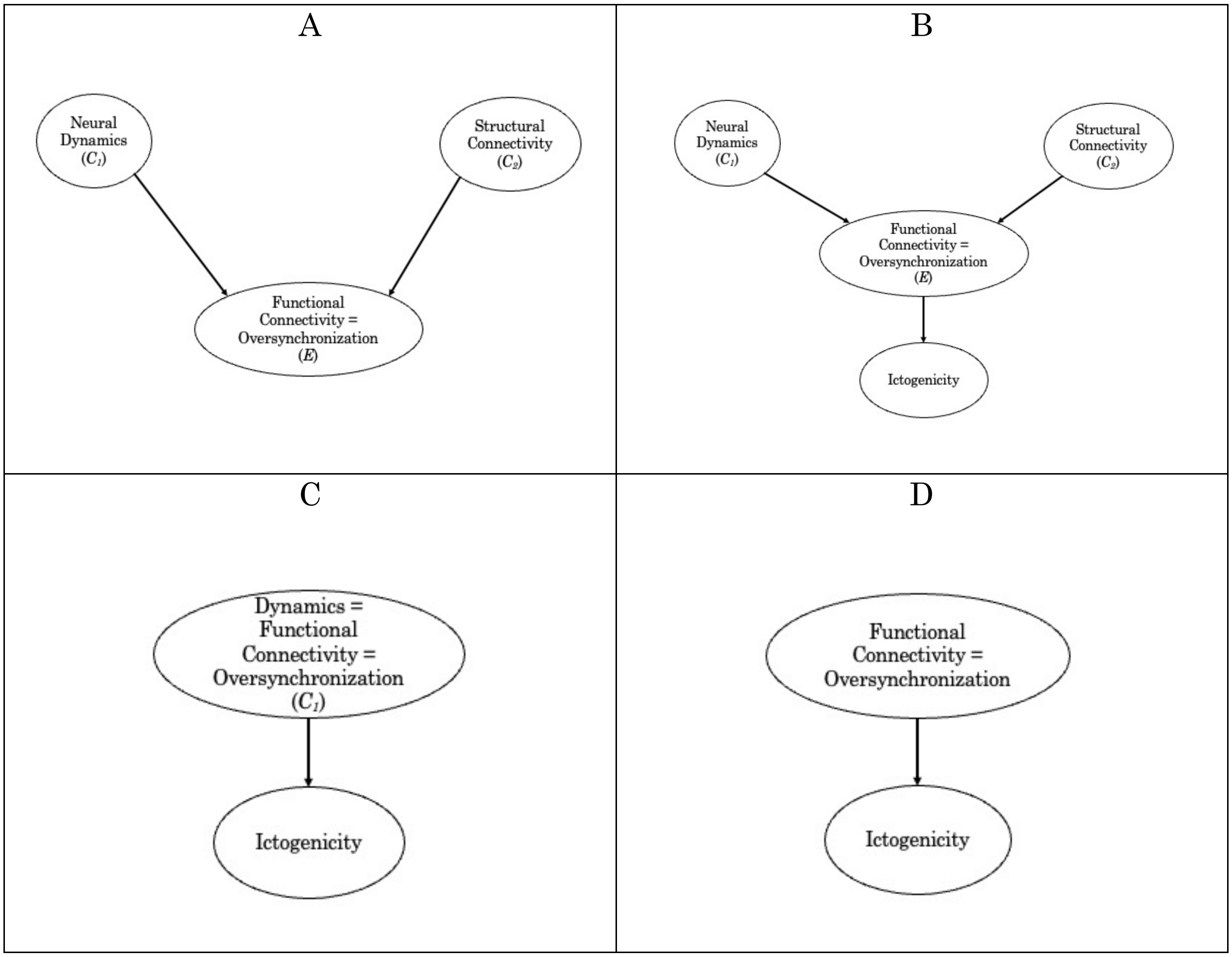
A
Neural Dynamics (C_1)
Structural Connectivity (C_2)
Functional Connectivity = Oversynchronization (E)
B
Neural Dynamics (C_1)
Structural Connectivity (C_2)
Functional Connectivity = Oversynchronization (E)
Ictogenicity
C
Dynamics = Functional Connectivity = Oversynchronization (C_1)
Ictogenicity
D
Functional Connectivity = Oversynchronization
Ictogenicity

Figure 1. Different Interpretations of Functional Connectivity's Causal Role in Helling, Petkov, and Kalitzin (2019)

7. Conclusion

The analysis of functional connectivity (FC) models and their explanatory role in neuroscience challenges a common skepticism that views them as merely correlational and thus non-explanatory. This is what we dubbed the "Wrong Stuff

Objection." In refuting it, we have shown that FC models can, under a rich interpretation, provide genuine scientific explanations. This argument is reinforced by case studies demonstrating how FC models contribute to our understanding of both cognitive functions, such as mentalizing, and pathological conditions, such as epilepsy.

By engaging with key philosophical debates on explanation, the paper situates FC models within a broader landscape of scientific reasoning. Functional explanations, though sometimes dismissed as lacking causal depth, offer indispensable insights into how brain regions work together to generate cognition and behavior. Neuroscientists, in practice, do not rely solely on anatomical connectivity to explain cognitive processes; they incorporate FC models as part of a larger ensemble of explanatory tools that include structural connectivity, dynamical models, and causal interventions.

This has implications not only for neuroscience but also for other disciplines, such as systems biology and complex systems research, where explanations often rely on abstract and high-level patterns of interaction rather than strictly mechanistic decompositions. Beyond its theoretical significance, this discussion has direct consequences for how neuroscientists approach experimental design, data interpretation, and clinical interventions. Similarly, in cognitive neuroscience, FC models are frequently used to delineate large-scale networks underlying mental functions like attention, memory, and decision-making. Accepting that these models have explanatory value legitimizes their use in framing hypotheses about brain function and refining theoretical accounts of cognition.

Additionally, this discussion sheds light on the broader methodological and epistemological challenges in neuroscience. The field has long struggled with how to bridge different levels of explanation—from molecular and cellular processes to systems-level organization and behavior. FC models exemplify this challenge by providing an intermediate level of description that captures large-scale coordination without necessarily specifying the mechanistic details at the neuronal level. As the field moves toward increasingly sophisticated, data-driven approaches, it must grapple with the nature of explanation.

Acknowledgments: We would like to thank Jim Woodward for his helpful discussions and comments on earlier versions of this paper, as well as two anonymous reviewers for their helpful comments.

Andrews-Hanna, Jessica R., Rebecca Saxe, and Tal Yarkoni. 2014. "Contributions of Episodic Retrieval and Mentalizing to Autobiographical Thought: Evidence from Functional Neuroimaging, Resting-State Connectivity, and Fmri Meta-Analyses." *NeuroImage* 91: 324-335.

Anfinsen, Christian B., and Edgar Haber. 1961. "Studies on the Reduction and Re-Formation of Protein Disulfide Bonds." *Journal of Biological Chemistry* 236 (5): 1361-1363.

Bogen, Jim. 2002. "Epistemological Custard Pies from Functional Brain Imaging." *Philosophy of Science* 69 (S3): S59-S71.

Buckner, Randy L., Fenna M. Krienen, and B. T. Thomas Yeo. 2013. "Opportunities and Limitations of Intrinsic Functional Connectivity Mri." *Nature Neuroscience* 16 (7): 832-837.

Bueno, Otávio, and Steven French. 2018. *Applying Mathematics: Immersion, Inference, Interpretation*. Oxford University Press.

Bullmore, Ed, and Olaf Sporns. 2012. "The Economy of Brain Network Organization." *Nature Reviews Neuroscience* 13 (5): 336-349.

Cole, Michael W., Genevieve J. Yang, John D. Murray, Grega Repovš, and Alan Anticevic. 2016. "Functional Connectivity Change as Shared Signal Dynamics." *Journal of Neuroscience Methods* 259: 22-39.

Contessa, Gabriele. 2007. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74 (1): 48-68.

Craver, Carl F. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68 (1): 53-74.

---. 2013. "Functions and Mechanisms: A Perspectivalist View." In *Functions: Selection and Mechanisms*, edited by Philippe Huneman, 133-158. Dordrecht: Springer Netherlands.

---. 2016. "The Explanatory Power of Network Models." *Philosophy of Science* 83 (5): 698-709.

De Regt, Henk W. 2017. *Understanding Scientific Understanding*. New York: Oxford University Press.

Díez, José A. 2020. "An Ensemble-Plus-Standing-for Account of Scientific Representation: No Need for (Unnecessary) Abstract Objects." In *Abstract Objects: For and Against*, edited by José L. Falguera and Concha Martínez-Vidal, 133-149. Cham: Springer International Publishing.

Frigg, Roman, and James Nguyen. 2017. "Models and Representation." In *Springer Handbook of Model-Based Science*, edited by Lorenzo Magnani and Tommaso Bertolotti, 49-102. Cham: Springer International Publishing.

---. 2020. *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Cham: Springer.

Friston, Karl J. 2011. "Functional and Effective Connectivity: A Review." *Brain Connectivity* 1 (1): 13-36.

Fuchs, Tom A., Ralph H. B. Benedict, Alexander Bartnik, Sanjeevani Choudhery, Xian Li, Matthew Mallory, Devon Oship, Faizan Yasin, Kira Ashton, Dejan Jakimovski, Niels Bergsland, Deepa P. Ramasamy, Bianca Weinstock-Guttman, Robert Zivadinov, and Michael G. Dwyer. 2019. "Preserved Network Functional Connectivity Underlies Cognitive Reserve in Multiple Sclerosis." *Human Brain Mapping* 40 (18): 5231-5241.

Garson, Justin. 2016. *A Critical Overview of Biological Functions*. Springer.

Gutte, Bernd, and R. B. Merrifield. 1971. "The Synthesis of Ribonuclease A." *Journal of Biological Chemistry* 246 (6): 1922-1941.

Harenski, Carla L., Vince D. Calhoun, Juan R. Bustillo, Brian W. Haas, Jean Decety, Keith A. Harenski, Michael F. Caldwell, Gregory J. Van Rybroek, Michael Koenigs, David M. Thornton, and Kent A. Kiehl. 2018. "Functional

Connectivity During Affective Mentalizing in Criminal Offenders with Psychotic Disorders: Associations with Clinical Symptoms." *Psychiatry Research: Neuroimaging* 271: 91-99.

Helling, Robert M., George H. Petkov, and Stiliyan N. Kalitzin. 2019. "Expert System for Pharmacological Epilepsy Treatment Prognosis and Optimal Medication Dose Prescription: Computational Model and Clinical Application." Proceedings of the 2nd International Conference on Applications of Intelligent Systems.

Hughes, R. I. G. 1997. "Models and Representation." *Philosophy of Science* 64: S325-S336.

Hutchison, R. Matthew, Thilo Womelsdorf, Elena A. Allen, Peter A. Bandettini, Vince D. Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H. Duyn, Gary H. Glover, Javier Gonzalez-Castillo, Daniel A. Handwerker, Shella Keilholz, Vesa Kiviniemi, David A. Leopold, Francesco de Pasquale, Olaf Sporns, Martin Walter, and Catie Chang. 2013. "Dynamic Functional Connectivity: Promise, Issues, and Interpretations." *NeuroImage* 80: 360-378.

Kalitzin, Stiliyan, George Petkov, Piotr Suffczynski, Vasily Grigorovsky, Berj L. Bardakjian, Fernando Lopes da Silva, and Peter L. Carlen. 2019. "Epilepsy as a Manifestation of a Multistate Network of Oscillatory Systems." *Neurobiology of Disease* 130: 104488.

Khalifa, Kareem, Jared Millson, and Mark Risjord. 2022. "Scientific Representation: An Inferentialist-Expressivist Manifesto." *Philosophical Topics* 50 (1): 263-292.

Khalsa, Sakh, Stephen D Mayhew, Magdalena Chechlacz, Manny Bagary, and Andrew P Bagshaw. 2014. "The Structural and Functional Connectivity of the Posterior Cingulate Cortex: Comparison between Deterministic and Probabilistic Tractography for the Investigation of Structure–Function Relationships." *Neuroimage* 102: 118-127.

Kostić, Daniel. 2020. "General Theory of Topological Explanations and Explanatory Asymmetry." *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1796): 20190321.

---. 2023. "Topological Explanations, an Opinionated Appraisal." In *Scientific Understanding and Representation: Mathematical Modeling in the Life and Physical Sciences*, edited by Kareem Khalifa, Insa Lawler and Elay Shech. London: Routledge.

Kostić, Daniel, and Kareem Khalifa. 2021. "The Directionality of Topological Explanations." *Synthese* 199 (5): 14143-14165.

---. 2022. "Decoupling Topological Explanation from Mechanisms." *Philosophy of Science* 90 (2): 245-268.

Latora, Vito, and Massimo Marchiori. 2001. "Efficient Behavior of Small-World Networks." *Physical Review Letters* 87 (19): 198701.

Logothetis, Nikos K. 2010. "Neurovascular Uncoupling: Much Ado About Nothing." *Frontiers in Neuroenergetics* 2.

Mill, Ravi D., Takuya Ito, and Michael W. Cole. 2017. "From Connectome to Cognition: The Search for Mechanism in Human Functional Brain Networks." *NeuroImage* 160: 124-139.

Morange, Michel. 2006. "Post-Genomics, between Reduction and Emergence." *Synthese* 151 (3): 355-360.

Novick, Rose. 2023. *Structure and Function*. Cambridge: Cambridge University Press.

Ramsey, J. D., S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. 2010. "Six Problems for Causal Inference from Fmri." *NeuroImage* 49 (2): 1545-1558.

Rathkopf, Charles. 2024. "How Network Models Contribute to Science." In *The Routledge Handbook of Philosophy of Scientific Modeling*, edited by Tarja Knuuttila, Natalia Carrillo and Rami Koskinen, 535-548. Routledge.

Reid, Andrew T., Drew B. Headley, Ravi D. Mill, Ruben Sanchez-Romero, Lucina Q. Uddin, Daniele Marinazzo, Daniel J. Lurie, Pedro A. Valdés-Sosa, Stephen José Hanson, Bharat B. Biswal, Vince Calhoun, Russell A. Poldrack, and Michael W. Cole. 2019. "Advancing Functional Connectivity Research from Association to Causation." *Nature Neuroscience* 22 (11): 1751-1760.

Sadowski, M. I., and D. T. Jones. 2009. "The Sequence–Structure Relationship and Protein Function Prediction." *Current Opinion in Structural Biology* 19 (3): 357-362.

Schilbach, Leonhard, Birgit Derntl, Andre Aleman, Svenja Caspers, Mareike Clos, Kelly M. J. Diederen, Oliver Gruber, Lydia Kogler, Edith J. Liemburg, Iris E. Sommer, Veronika I. Müller, Edna C. Cieslik, and Simon B. Eickhoff. 2016. "Differential Patterns of Dysconnectivity in Mirror Neuron and Mentalizing Networks in Schizophrenia." *Schizophrenia Bulletin* 42 (5): 1135-1148.

Siddiqi, Shan H., Konrad P. Kording, Josef Parvizi, and Michael D. Fox. 2022. "Causal Mapping of Human Brain Function." *Nature Reviews Neuroscience* 23 (6): 361-375.

Smith, Stephen M. 2012. "The Future of Fmri Connectivity." *NeuroImage* 62 (2): 1257-1266.

Sporns, Olaf. 2014. "Contributions and Challenges for Network Models in Cognitive Neuroscience." *Nature Neuroscience* 17 (5): 652-660.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation." *Philosophy of Science* 71 (5): 767-779.

---. 2015. "Deflationary Representation, Inference, and Practice." *Studies in History and Philosophy of Science Part A* 49 (0): 36-47.

---. 2024. *Inference and Representation: A Study in Modeling Science*. University of Chicago Press.

Tomasi, Dardo, and Nora D. Volkow. 2011a. "Association between Functional Connectivity Hubs and Brain Networks." *Cerebral Cortex* 21 (9): 2003-2013.

---. 2011b. "Functional Connectivity Hubs in the Human Brain." *NeuroImage* 57 (3): 908-917.

Uddin, Lucina Q., A.M. Clare Kelly, Bharat B. Biswal, F. Xavier Castellanos, and Michael P. Milham. 2009. "Functional Connectivity of Default Mode Network Components: Correlation, Anticorrelation, and Causality." *Human Brain Mapping* 30 (2): 625-637.

van den Heuvel, Martijn P., and Olaf Sporns. 2013. "Network Hubs in the Human Brain." *Trends in Cognitive Sciences* 17 (12): 683-696.

van Fraassen, Bas C. 1980. *The Scientific Image*. New York: Clarendon Press.

Vasileiadi, Maria, Anna-Lisa Schuler, Michael Woletz, David Linhardt, Christian Windischberger, and Martin Tik. 2023. "Functional Connectivity Explains How Neuronavigated Tms of Posterior Temporal Subregions Differentially Affect Language Processing." *Brain Stimulation* 16 (4): 1062-1071.

Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393 (6684): 440-442.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

---. 2023. Networks, Dynamics and Explanation. *https://philsci-archive.pitt.edu/22694/*.

---. 2025. "Networks, Dynamics and Explanation." *Synthese* 205 (5): 204.

Wouters, Arno G. 2003. "Four Notions of Biological Function." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34 (4): 633-668.