Cognitive Ontologies, Task Ontologies, and Explanation in Cognitive Neuroscience.

1. Introduction.

The development of new scientific tools provides opportunities for progress, but also gives scientists reason to reinvestigate, reconsider, and maybe revise their assumptions about the domain under investigation. In cognitive neuroscience, this has manifested in the debate over "cognitive ontology" – that is, the set of mental functions or faculties investigated by the neurosciences. Psychology comes equipped with a series of intuitive mental categories – perception, cognition, memory, imagination, verbal reasoning, emotion, etc. Cognitive neuroscience has traditionally proceeded under the assumption that these, or some suitably explicated set of these, will be realized in the processes that neuroscientists investigate. For better or worse, however, this assumption sits poorly with the current evidence, which indicates the massive multifunctionality of individual parts of the brain, the wide distribution of activity corresponding to intuitive mental categories, and the importance of global network and ecological context in determining what an individual part of the brain does. These data stress, and perhaps break, the "new phrenological" (Uttal, 2001) approach to cognitive and systems neuroscience, invalidating cherished means of analysis such as subtractive methodology and reverse inference.

One powerful thought in the field is that part of the problem is our intuitive conception of psychology. Indeed, Poldrack once said that "*the* fundamental problem is our stone age psychological ontology" (Bunzl, Hanson, & Poldrack, 2010, p. 54). Perhaps the standard mentalistic categories used in the psychological sciences are just too simple, too general, and too crude to capture how the brain implements behavior. Perhaps those categories need to be revised, refined, or even abandoned to understand brain function. A host of questions immediately arises, however, surrounding how committed we should remain to our standard list of cognitive kinds. Do they successfully describe brain function, only at a network level? Could we discover discrete implementations of kinds if they are suitably amended, for instance by subdividing them into more specific kinds? Or should they just be gotten rid of, resulting in a view of the brain on which it is "unanalyzable" (Uttal, 2001) into distinct functions, where its function is "protean" and lacking generalizability (Hutto, Peeters, & Segundo-Ortin, 2017)?

Theorists interested in cognitive ontology are thus facing a conundrum that is both methodological and ontological. What cognitive categories are realized in the brain cannot be determined independently of our methods of investigation. But in cognitive neuroscience, those methods traditionally employ those categories as basic assumptions – i.e., they are what *is being investigated*, and thereby constrain interpretation of otherwise inscrutable brain data. Theorists have begun to use formal tools from databasing, machine learning, and meta-analysis as a way of addressing this problem. The hope is that the use

of these tools can turn the issue of cognitive ontology into a problem for data science, rather than metaphysics. By analyzing large amounts of studies using an agreed upon, publicly shareable taxonomy of cognitive function, neuroscientists hope to be able to *discover* the ways in which cognitive categories relate to brain activation, and thereby provide a groundwork for the substantiation, revision, or abandonment of those categories.

I refer to these projects collectively as "databasing and brain mapping" projects, and in this paper, I assess their status and prospects. Ultimately, I will argue that the problem is not so much with our intuitive mental ontology *per se*, but with the standard explanatory framework assumed by the cognitive neurosciences. The standard framework assumes that categories of mental function are *explanatory* kinds, and that cognitive neuroscience proceeds by showing how these explanatory categories are *instantiated* in brain activity. As such, the standard framework is committed to there being an ultimate taxonomy of distinct and discretely realized cognitive kinds, whose instantiation in the brain causally explains behavior. I will argue that databasing and brain mapping projects, rather than substantiating this standard framework, should inspire us to abandon it. Instead, I advocate an alternative view of neuroscientific explanation on which what explains are ways in which brain systems organize to implement the informational demands of a particular task or context (Author's paper). On this alternative, the best reading of the role of psychological constructs is as *heuristics* for investigation, rather than as explanatory kinds (cf. Feest, 2010).

My aims are both descriptive and normative. I both believe that this is what successful neuroscientific explanation *does* look like, and that it is how we *should* think about it. This comes along with a variety of methodological prescriptions, including a plea for increased focus on *task,* rather than *cognitive* ontologies. I hope to clarify the potential advantages and pitfalls of using formal analytical tools in the cognitive ontology debate along the way.

I proceed as follows. In section 2, I introduce the standard explanatory framework of cognitive neuroscience, articulate its commitments, and discuss methodological and empirical problems for the framework present in the literature. Then, in section 3, I outline some of the formal tools that have been applied to the problem, and in section 4 show that no clear consensus has emerged on how results employing these methods are supposed to relate to the standard framework. In section 5 I outline my preferred approach to understanding the role of psychological constructs in neuroscientific explanation, and in section 6 show how this approach offers distinct normative prescriptions that the standard framework. Section 7 concludes.

2. The Standard Explanatory Framework of Cognitive Neuroscience

As I understand it, the standard explanatory framework in cognitive neuroscience is as follows. First, psychological kinds are explanatory. Behavior is explained by citing mental functions such as memory, attention, language processing, action planning, etc. Second, explanation is *causal* explanation. It is the realization of psychological kinds in neural processes that explains behavior. To take one philosophical gloss on the issue, consider Piccinini and Craver's (2011) account of the role of psychological theories in cognitive neuroscience. Psychology, on their view, provides *mechanism sketches* of cognitive phenomena. They outline a causal sequence of mental functions that can explain the phenomena of interest. A full *mechanism schema*, on the other hand, will show how this abstract functional organization is realized in lower-level causal interactions in the brain, eventually bottoming out in the electrical and chemical processes of individual cells. This can be read as a way of combining Cummins' (1983) functional decomposition approach to psychological explanation, with the explanatory goals of cognitive neuroscience.

This kind of view is very influential (Boone & Piccinini, 2016), describes the traditional explanatory practices of cognitive neuroscientists well, and comes with a number of advantages. First, it gives a metaphysically appealing picture of mental causation, wherein psychological states cause behavior via their realization relation to the physical processes of the brain. Second, it explains the importance of *operationalization* and *localization* in cognitive neuroscience. 'Memory' is not something we can study directly; we can only study behavior. On the standard story, behavioral tasks are designed to dissociate the different components of putative cognitive processes. The brain is then studied to show how those functional differences are causally realized in distinct part of the brain. These changes can be measured either through differences in activation – the traditional "subtraction" methodology in fMRI research – or through intervention, by studying artificially induced or naturally occurring brain injuries. By finding different localizations of distinct functions, one explains the causal differences between, for example, a memory process and an attentional process. A diagram of the standard framework is provided below.

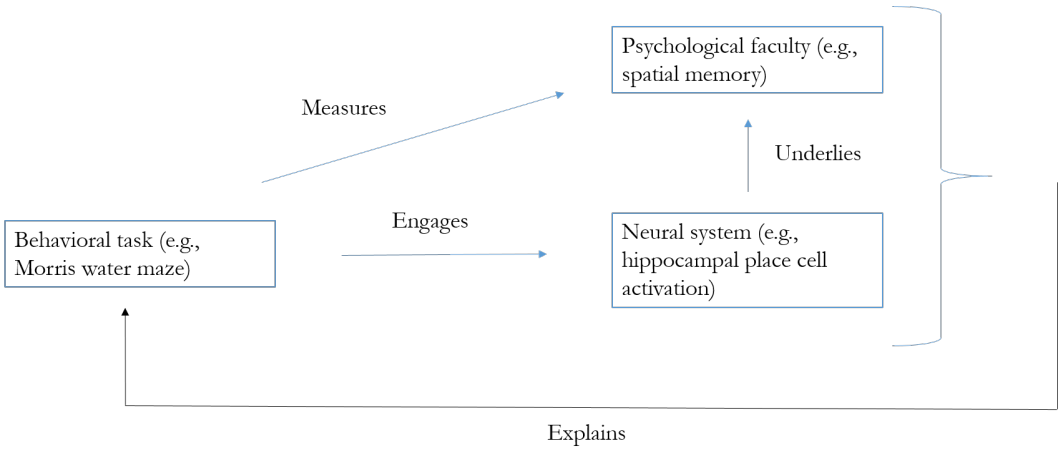

Figure 1. The standard explanatory framework.

Despite its appeal, the standard framework faces a large class of problems, which we can generally refer to as *individuation* problems. Individuation problems are the result of the realism about psychological kinds inherent in the standard framework, along with the complexity of causal processes in the brain. Basically, different *explanantia* need to be distinct. The standard framework is committed, in each instance, to establishing exactly what psychological functions are explaining a behavior, how they interact, and how the behavior arises as a result of that interaction. Individuation problems lead us to question whether this commitment is in fact met in cognitive neuroscience – they suggest that, very frequently, we have not, and even *cannot*, establish exactly which psychological faculties are at work and the pattern of their interaction.

There are two kinds of individuation problems that have seen extensive discussion in the literature. The first is an individuation problem with *operationalization* and *measurement*. The question here is whether tasks individuate particular mental faculties. Sullivan (2010), for instance, questions whether the Morris water maze, a famous task in neuroscience, can be taken as a specific measurement of 'spatial memory', the cognitive kind with which it is generally associated, rather than measuring learning, the change of representational capacities, or just ability to find a platform under the surface of some water. Sullivan (2014) also applies the argument to tasks such as Stroop tasks. While standardly thought of as measuring attention, Sullivan persuasively argues that elements of attention, memory, language processing, and perceptual processing are all indexed in the standard versions of Stroop tasks.

The second version of individuation problem applies to the kinds themselves. The worry is that psychological kinds are simply not distinct from one another. So, perhaps memory is not distinct, metaphysically speaking, from action planning or imagination (De Brigard, 2014; Schacter, Benoit, De Brigard, and Szpunar, 2015; but cf. Robins, 2016). Perhaps "basic" emotions (Griffiths, 2002), or "concepts" (Machery, 2009), or psychiatric categories such as "schizophrenia" (Tekin, 2016) do not correspond to natural kinds. Recently, these worries have been extended to core cognitive capacities like working memory (Gomez Lavin, 2020).

I am moved by these considerations, but will not focus on them here. What I will assess in detail is the last variety of individuation problem, which occurs in the purported *realization* of cognitive kinds. The traditional approach to cognitive neuroscience hopes to specify *the* function of each part of the brain. This "atomistic" (Burnston, 2019) approach is motivated by the goal of causal decomposition in the brain. If we can specify the function of each part, then we can understand any given behavior as the result of causal interaction between

these functions. A one-to-one mapping between the mental ontology and the neuroscientific ontology would occasion a particularly powerful form of mechanistic explanation.

Current data, however, sits uneasily with individuating atomic mappings between functions and brain activation. Widespread data from many distinct parts of the brain suggest the *multifunctionality* of individual brain parts, *overlap* between instantiations of distinct mental categories, and *distribution* of brain activation corresponding to distinct kinds. Multifunctionality of particular parts of the brain undermines the ability to say, given activation in a particular part, what function that part is performing, and hence how it is causally interacting with other parts. This is part of the problem with the traditional subtractive methodology and reverse inference. *Overlap* and *distribution* of function undermines the ability to distinguish between the causal contribution of distinct kinds at the neural level.

It is worth considering further why this is. Suppose we are observing a behavior, and activation in a number of parts of the brain. If one can decompose the behavior into distinct psychological processes, and localize each of those processes, then one can theorize about the causal interactions between them. But significant distribution and overlap of instantiation muddy the division between localization and interaction. This is because they allow for too much inferential freedom in how one interprets the instantiation relation. Suppose two putatively distinct faculties overlap in their instantiation. Is this due to the fact that they share some common functional core, and other elements that differ? Is it because we have not explicated them sufficiently relative to each other? Or is it because they are not, in fact, distinct after all? Similarly, given wide distribution corresponding to a given function, is that distribution indication that the construct ranges over multiple distinct sub-functions that interact? Or that our experiments in fact index multiple distinct functions? Or, again, that the construct does not describe the mechanistic functioning of the brain?

In most cases of mental faculties, this is the situation that actually obtains – the data suggests multifunctionality, distribution, and overlap. But the standard framework requires that distinct explanantia be distinct, including in their instantiations. So, the current data conflicts with the standard framework. Given this conflict, one can either attempt further work to substantiate the standard framework, or one can abandon it. To substantiate the framework, one would have to either try to further differentiate the instantiation relations between distinct kinds, or revise the ontology so that more specific mappings emerge. One could, of course, pursue some combination. If abandoning the framework, one would have to specify what kind of explanation results from that abandonment.

The idea of revision of the cognitive ontology is appealing, and it is often suggested in mechanistic contexts that higher-level kinds will have to be split or revised in light of causal explanation at a lower level (Bechtel, 2008; Bickle, 2003). Similarly, it is often suggested that databasing and brain-mapping techniques can help us revise our ontology. I will consider these claims in detail in the next section. But it is worth noting that the ontology revision proposal is less anodyne then is normally supposed. Attempting to fine-grain our taxonomy does not guarantee that distinct functions will be discovered – Feest (2010), for instance, nicely explains how continuous attempts to distinguish implicit memory from other forms of memory are what eventuated in the conclusion that implicit memory may not be distinct from perceptual association. And even a successful distinction may not be mechanistically useful – attempts to distinguish face perception, body perception, and place perception, for instance, do not show clearly distinct realizers but interdigitated "archipelagoes" of voxels with statistical preference for one kind of information over another (Kanwisher, 2010), but not clearly distinct parts with causal interactions between them.

In the next two sections, I introduce databasing and brain mapping techniques in more detail, and argue that, while proponents of these techniques are generally *realist* about psychological kinds, they do not clearly opt for either substantiation, revision, or abandonment of the standard framework, instead vacillating between these options. I also raise doubts that the databasing and brain mapping techniques on offer can perform any of these functions. This motivates my own take on the issue, which I will pursue in section 5.

3. Databasing and Brain Mapping

There are two main aspects to the databasing and brain mapping projects I will discuss. The first is the collection of compendious amounts of neural data from across studies. The shareability of scientific data, as well as the best means to collect it, disseminate it, and use it are problems across the biological sciences (see, e.g., Bechtel, 2017; Darden, Pal, Kundu, & Moult, Forthcoming; Leonelli, 2012), and neuroscience is no different. One reaction to the massive amount of research using, for instance, fMRI methodology, is to try to systematize and understand this expansive dataset as a whole. So, collection of the information is the first step. Several open-access databases have been created in cognitive neuroscience to play this role, including the Brain Map (Fox & Lancaster, 2002), Neurosynth (Yarkoni et al., 2011), The Cognitive Atlas (Poldrack et al., 2011), the Experiment Factory (Sochat et al., 2016), and The Cognitive Paradigm Ontology (Turner & Laird, 2012).

While these projects differ in their precise focus, they all share a number of aspects. First, the idea is to collect activation data in a theoretically unbiased way. What is archived is the raw activation data from a set of fMRI studies. One can then ask questions about this data. Second, one of the questions that everyone wants to ask about this data is how, whether,

and in what sense patterns in the data correspond to psychological concepts. This is done in a number of ways. In the Cognitive Atlas, each study, in addition to the data, is categorized according to the type of tasks manipulated. Each task-type is then defined, and the psychological concepts that it is supposed to measure are listed. So, one can look for the ways in which the same tasks/concepts are realized similarly or differently across different tasks, or one can look at how different tasks/concepts diverge or overlap. In Neurosynth, the *text* of papers is archived along with the raw fMRI data, so one can look for ways in which *usage* of key mental terms by scientists varies along with changes in brain activation, and vice-versa. Finally, the hope across these projects is that the search for patterns can be *automated*. Given the scale of the data set, automated data analysis is used in the attempt to find meaningful patterns.

The analytical techniques applied to this collected data range from traditional meta-analyses to statistical classifiers to generative, probabilistic models, each with their associated benefits and detractions. Meta-analytic techniques take already reported correlations between cognitive concepts and activation patterns, and attempt to identify, generalize, and summarize the relationships discovered in the literature. Statistical decoders train models to predict, given the presence of brain activation, what cognitive concepts are being assessed in the range of studies (or vice versa, see below). While there are a range of generative models, one popular technique is *Latent Dirichlet Allocation*, which is a Bayesian algorithm that models the text in a corpus of words (in this case, the text from fMRI studies) as being generated by a grouping of topics, themselves construed as probabilistic groupings of individual words. One can then attempt to correlate greater influence (or "loadings") of those topics with brain activation.

There are a range of attitudes taken by brain mappers to their projects, which I will discuss in detail below. In general, however, I think there are two fundamental assumptions they share. First, they are *realists* about mental faculties. Second, and relatedly, they are committed to the *measurement* relation between tasks and those faculties. For instance, Hastings et al. (2014) describe the project as one on which "ontological realism is a foundation" (p. 4). Lenartowicz et al. (2010) suggest that "the elements of the mental ontology are not directly accessible but rather must be accessed through experimental manipulations and measurements (i.e., tasks)" (p. 680).

In these quotes, theorists are committing to the idea that mental functions are real entities, and that tasks are measurements of them. This is reflected in much of the databasing work. In the Cognitive Atlas and Brain Map, for instance, tasks are explicitly categorized according to the mental constructs they are supposed to measure. While Neurosynth collects a range of textual data, the preprocessing of that data indicates a realist commitment. Generally, LDA models using Neurosynth focus on the abstracts of paper, and specifically on the cognitive terms contained therein. In attempting to map these uses

to the brain, then, these projects assume that, at least at an abstract level, the concepts we employ in thinking about the brain correspond to physical categories within the brain.

As I will show below, this set of commitments interacts in complicated ways with the standard framework. For now, I want to discuss a few early results from these frameworks to show that, far from solving individuation problems, brain mapping projects tended to illustrate them. The question will then be what attitudes brain mappers take to these results.

In a meta-analysis of fMRI research, Anderson, Kinnison, and Pessoa (2013) compared different patterns of activation according to the cognitive categories listed in Brain Map. They were interested in a number of properties, including the range of cognitive concepts associated with each area's activation, the distribution of activation corresponding to those concepts, and the degree to which distinct brain areas were likely to be active in studies measuring the same mental concepts. What they showed was that individual parts of the brain exhibit a range of "diversity profiles," but that most areas' activation corresponded with significantly more than one cognitive concept. Moreover, areas within previously identified functional-connectivity networks tended to be highly "assorative," meaning they tended to be active for similar cognitive concepts, suggesting both the distribution of individual functions and the overlap in neural activation between functions. Importantly, taking functional networks such as the "fronto-parietal" network and the "ventral attention" network, as basic units to correlate with cognitive concepts showed a similar pattern of results.

Poldrack, Halchenko, and Hanson (2009) performed a decoding analysis of the results from eight different fMRI studies investigating a range of cognitive constructs. They began with statistical maps of the entire brain – i.e., z-scored activation coordinates from every condition in the eight studies. The question was then whether one could train a decoder to predict which tasks and/or cognitive concepts were named in the studies, such as "risk-taking". They trained a support vector machine to predict, on the basis of a given brain-wide activation, what task and what cognitive concept was being measured in a case, and showed that the classifier could successfully classify both with 80-90% accuracy. They further trained a neural network with six hidden nodes to match the predictive accuracy of the support vector machine. Importantly, however, the nodes operated over a widely-distributed set of voxels. When analyzed as a six-dimensional system (one for each hidden node), each cognitive concept was shown to be related to a combination of each dimension, and each dimension was associated with a range of cognitive concepts.

Poldrack et al. (2012) performed a topic modeling analysis with the following structure. First, they took the results and text from over 5,000 papers in the *Neurosynth* database. They began by exploring the topic structure in the text. They then selected topics that

corresponded to mental concepts in the Cognitive Atlas, and measured how the topic loadings on these topics correlated with brain activity. Here, however, they also show multifunctionality and distribution in the results. For instance, they report: "topic 43 (with terms related to visual attention) was associated with activity in the bilateral lateral occipital cortex, parietal cortex, and frontal cortex. Topic 86 (with terms related to decision making and choice) was associated with regions in the ventral striatum, medial, orbital, and dorsolateral prefrontal cortex. Topic 93 (with terms related to emotion) was associated with bilateral activity in the amygdala, orbitofrontal cortex, and medial prefrontal cortex" (2012, p. e1002707).

These results are perfectly interesting in their own right, in that they quantify the "specificity" with which our intuitive cognitive concepts interact with brain activation. It is just that, on their face, they are in conflict with the standard model because they show multifunctionality and distribution rather than univocal relationships between psychological constructs and activation. The question is what to do in response to these results with regards to the standard framework.

Let me stress that I am reconstructing positions here – I think each of the options I am about to articulate is present, to some degree, across papers and theorists within the field. As far as I can tell, there are three options with regards to the standard framework. First, one could attempt to substantiate the framework by pursuing more fine-grained analyses in an attempt to discover more and more specific activation patterns for particular cognitive concepts, perhaps further leading to decomposition and causal explanation. Second, one could attempt to use the analyses to *revise* our cognitive ontology. On this view, it might be the case that the standard framework can be maintained, but only after the appropriate revisions to the ontology. Third, one might use these results as motivation to *abandon* the standard framework altogether, and opt for some other kind of project. In the next section, I outline each of these perspectives, along with examples from the literature which might suggest them, and give reasons to question them. This will motivate my own proposal about psychological constructs in section 5.

4. Three options with regards to the standard framework.

4.1 Substantiate?

One view one could take towards the standard framework is that it is basically right, *and* our ontology is basically right, but that the results of multifunctionality and distribution are due to insufficiently fine-grained measurement. The solution, if this is one's perspective, would be to refine analyses so that the "true" and univocal association between psychological constructs and brain activation can be uncovered.

I think that this is the position that is least strongly considered in the literature, but there are a few trends that suggest it. Indeed, one direction in which the literature has gone over the last few years is in the direction of more fine-grained analysis and the search for increased specificity. Poldrack and Yarkoni (2016) thus describe the project as one of "quantifying the true specificity of hypothesized structure-function associations" (p. 589). This, one assumes, means that they indeed take there *to be* relations there to be discovered, further indicating realism about psychological kinds. Moreover, recent projects take the goal of brain mapping projects to be enabling both *forward* and *reverse* inference – that is, the predictive ability of brain mapping models between constructs and activation patterns should be bidirectional. This, to me at least, further suggests a belief in the importance of the instantiation relation. Finally, there is how these projects are qualitatively described. For instance, Varoquaux et al (2018) suggest that one of the goals of mapping projects is "precisely describing the function of any given brain region" (2018, p. 1).

Varoquaux et al. performed a decoding analysis using a hierarchical general linear model (GLM) framework. In particular, their reverse-inference required on multiple layers of linear regressions on activity in the brain. The first layer was tuned to individual oppositions between task conditions. Then, a second layer used another regression that compared each cognitive term to all others, predicting which term was overall most relevant. They compared the results of this decoder to other approaches, showing that it resulted in sharper divisions between distinct functions.

There are a few things to be said here, however. First, this study measured terms in the Cognitive Paradigm Ontology rather than the Brain Map or the Cognitive Atlas, and these terms more directly describe task conditions (e.g. "response with left hand") than psychological constructs (e.g., "motor control"). Second, they focused primarily on perceptual and motor areas for which there are already more-or-less well-understood general function ascriptions. Finally, even *these* results showed distributed and interdigitated functional populations, with, for instance, "face" and "place" areas being more or less separated, but each involving multiple subpopulations distinct from each other.

Another recent approach to bidirectional decoding is from Rubin et al. (2017), which employs LDA on over 11,000 articles from Neurosynth. They start by noting that previous studies show mainly wide patterns of activation for particular constructs, and thus are no help in finding "relatively simple, well-defined functional-anatomical atoms." To overcome this, they performed an LDA analysis constrained *both* by the semantics of the terms and by groupings in spatial coordinates. They report that, not only were they able to uncover topics with relatively clear functional upshot, (e.g., topics related to 'emotion'), but that each topic "is associated with a single brain region". At first, this sounds a lot like the

explanatory aims of the standard framework – i.e., to find a constrained localization corresponding to each psychological function.

A closer reading questions this analysis, however. As the researchers note, the probabilistic nature of the model suggests that the decoding analysis uncovers the construct *most likely* associated with a given area, but not the only one. This is further illustrated by the fact that individual topics were allowed to spatially overlap in the model, and many multifunctional areas did indeed show significant overlap between related topics. Further specifying to individual topics in many cases required conditioning further on more spatial coordinates, hence suggesting, again, distribution of function. So, while the results in this model are *predictive* at a very specific construct-spatial level, it is not clear that this reflects the reality of the system. And this is noted explicitly by the researchers. It is worth quoting them in full:

"While the topics produced by the model generally have parsimonious interpretations that accord well with previous findings, they should be treated as a useful, human-comprehensible approximation of the true nomological network of neurocognition, and not as a direct window into reality. For the sake of analytical tractability, our model assumes a one-to-one mapping between semantic representations and brain regions, whereas the underlying reality almost certainly involves enormously complex many-to-many mappings. Similarly, rerunning the GC-LDA model on different input data, with different spatial priors, a different number of topics, or with different analysis parameters would necessarily produce somewhat different results" (Rubin et al., 2017, p. 14).

So, while one trend in the literature is to look for increasingly specific relationships between extant psychological constructs and patterns of activation, it is not clear that even successful results in this endeavor substantiate, or should be read as attempting to substantiate, the standard framework.

4.2. Revise?

The idea that the databasing and brain mapping project can help us revise our cognitive ontology is extremely common. For example, Poldrack and Yarkoni (2016) suggest that "formal cognitive ontologies [are useful] in helping to clarify, refine, and test theories of brain and cognitive function" (p. 587), and that "biological discoveries can and should inform the continual revision of psychological theories" (p. 599).

These quotes suggest that, ultimately, the role of the databasing and brain mapping project will be in helping us to explicate our mental ontology. Sometimes, this is pitched in terms of a discovery science – we should let the brain tell us what its functional categories are, and revise our ontology accordingly (Poldrack et al., 2012). In this section, I suggest two

related problems for this view. The first is the *interpretability* problem, and the second is the *seeding* problem. In general, however, the issue is this: without a *rubric* for how and when to revise our mental categories in light of brain mapping data, we lack the ability to use results from brain mapping to revise the ontology in any specific way. This suggests that metaphysical commitments about the nature of mental states *precede*, rather than being compelled by, brain mapping data.

The interpretability problem is akin to a problem discussed by Carlson et al. (2018; cf. Ritchie, Kaplan, and Klein, 2016) for uncovering neural *representations* via machine learning techniques. They argue that, given a particular ability to decode some stimulus from neural activity, it is unclear how to *interpret* that result in terms of representational content. The worry, I take it, is that the ability to decode a stimulus from an activation does not mean that the activity represents the stimulus under anything like the way we would describe it. The analogue problem here is that, simply showing that a pattern of activity in the brain is specific to, say, decision-making (or to a high topic loading on a topic that happens to comprise words we associate with decision-making), doesn't give us any indication of whether the pattern of activity is in fact performing something we would call "decision-making." The more distribution and overlap uncovered in the analysis, the more exacerbated this problem becomes, because of the inferential freedom discussed in section 2.

So, given the association of a pattern of activity with a mental construct, should we take that construct as substantiated, as in need of explication, or what? What degree of correlation/predictability, or what degree of specificity, is required to count the kind as substantiated, and at what point should we consider it in need of revision? The brain mapping results themselves provide no rubric for how to make these decisions.

The seeding problem is related to the interpretation problem, and is based on the fact that even *constructing the analyses* requires adhering, to an unspecified degree, to our extant cognitive constructs. In an analysis based on Brain Map or the Cognitive Atlas, one only *considers* concepts that are a current part of our mental ontology. This presumes that the basic structure of the brain corresponds closely enough to those categories in order for them to be useful in understanding the brain. But what justifies this assumption? In principle, a specific-enough correlation between mental constructs and brain activity might justify the assumption, but it is precisely a *lack* of specificity of this type that prompts the idea of ontology revision.

In topic-modeling analyses, the topics that are often focused upon are the ones that one can intuitively or statistically pair with an already-known mental construct. Varoquaux et al., for example, advertise that 100 of the 200 topics in the model correspond to well-understood mental constructs. What about the other ones, however? Even given a

substantiation of some of our mental categories, what the analysis would suggest is that our ontology is at least impoverished, and it does not come with any prescriptions for what to say in these other cases.

Again, the point of this is not to discount the analysis. The point of if it is just to deny that, the analysis on its own offers us any principled way of revising our cognitive ontology. Put differently, the principles for ontology revision cannot be uncovered bottom-up from these analyses. Metaphysical commitments must be undertaken in constructing and interpreting the analyses themselves. Again, theorists in the field recognize this problem. Poldrack and Yarkoni (2016), for instance, note that there is "no algorithmic way" to approach ontology revision in light of specific mapping results. They seem to suggest, however, that more analysis and case-by-case thinking will allow for sufficient explication. The individuation and seeding problems should raise concerns for that approach.

4.3. Abandonment?

One also finds more-or-less explicit discussion of the explanatory ideals of the standard framework in the literature. The clearest cases of these are Yarkoni and Westfall (2017) and Anderson (2014). Yarkoni suggests explicitly that results from databasing and brain mapping projects suggest abandoning *explanation* altogether, in favor of a purely predictive neuroscience. Anderson's view does not cite prediction per se, but does suggest that we need to change to a *dispositional* approach to brain organization, wherein we do not understand a part of the brain as contributing a specific causal influence at a specific time, but instead as exhibiting dispositions to contribute to a range of functions.

I lack the space to assess these proposals in detail, but for my purposes it suffices to note that they both, more-or-less-explicitly, move away from the mechanistic kind of explanation inherent to the standard framework. Much has been said about the relative merits of mechanistic explanation versus prediction in explanation (Craver 2006), and now is not the time to re-adjudicate these issues. What I want to argue for in the remainder of the paper is that abandoning the standard framework is not itself equivalent to abandoning mechanistic explanation. Instead, we can abandon the standard framework by abandoning the central explanatory role it affords to mentalistic constructs.

5. An Alternative View.

5.1. Mental constructs as heuristics.

My proposal is based around the following negative claim: *posits of psychological faculties do not explain behavior*. Individuation problems arise from the notion that posits of psychological faculties are explanatory. That is why they must be distinct from each other;

that is why they must be discretely realized; and that is why causal interactions between them need to be established. Get rid of their explanatory status, and all of those problems go away – it's not particularly worrisome of psychological kinds are not clearly distinct from each other, if specific behavioral tasks don't measure only one of them at the expense of others, or even if they massively "crosscut" the causal patterns we measure in the brain (Hochstein, 2016; Weskopf, 2011)

This leaves us with two questions. First, what does explain behavior? And second, do psychological posits play any role in understanding it? These questions can be asked within the context of brain mapping projects as well. What sort of explanation should these projects be seen as working towards? And should mental concepts play any role as they seek them?

My answers are as follows. First, information processing in the brain explains behavior directly, and not in virtue of instantiating some particular mental function. As I will attempt to show, this proposal is compatible with each brain area being multifunctional, and with function generally being distributed (Author's papers). Second, mentalistic concepts are best understood as playing a *heuristic* role (cf. Feest, 2010). Rather than serving as explanantia, we should view mental categories as helping us parse behaviors into rough, and revisable, similarity classes. They provide traction on an otherwise impossibly complex space of behavioral abilities, and guide search for important distinctions between types of behaviors, so that we can *then* investigate how the brain implements those differences. Importantly, they can play this role even if *there is no fact of the matter* about which neural processes implement which psychological constructs. Hence, no individuation problems are faced.

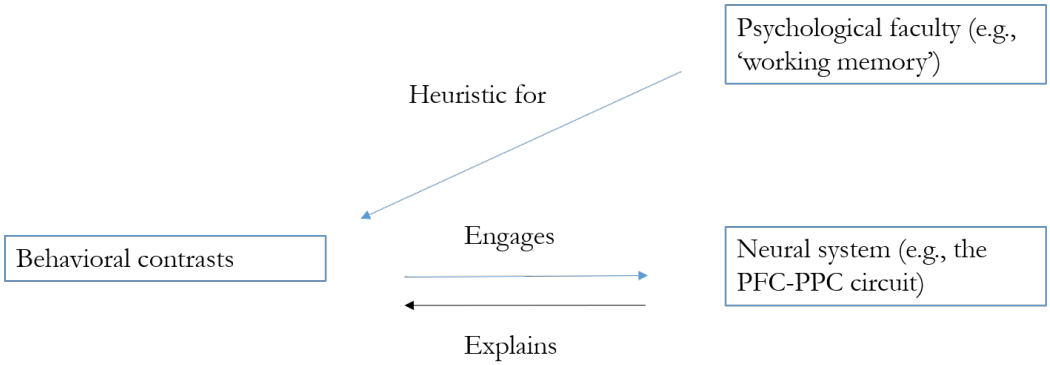

Figure 2. The heuristic approach to mental constructs.

5.2. An exemplar.

The heuristic view makes a number of invocations about successful explanations in neuroscience. First, overlap between mental constructs across tasks and contexts should be just as important separation between them. Second, understanding the differences in structure between tasks is paramount for understanding neural function. Third, assuming spatial decomposition between distinct purported mental faculties would *limit*, rather than enabling, understanding of how the system works.

I will discuss one example in detail. Murray, Jaramillo, and Wang (2017) pursued a modeling study of the interaction between the prefrontal cortex and the posterior parietal cortex. The initiating motivation for their study is that both the PFC and the PPC have been shown physiologically to be involved across a wide range of both working memory (WM) and decision-making (DM) tasks. The question, then, is what their distinct contributions are.

Murray et al.'s approach was as follows. They modeled each area as a fully recurrent neural network, and each area had distinct sub-populations selective for distinct perceptual stimuli. The PFC and PPC networks were bi-directionally connected via long-range projections. The main difference between the two populations was a difference in local structure. In particular, the PFC population was modeled as having a higher degree of internal influence – both in self-excitation of each subpopulation, and in inhibitory connections between them, than the PPC. Given this network structure, the investigators could model the dynamics of the system in a range of task types, and think about how the network responded in each.

Murray et al. posited that one key factor involved in working memory tasks is *multi-stability*. That is, the network can represent a range of possible stimuli, but given that it has already represented one, it must maintain that information across a delay, perhaps in the presence of distractors. So, they modeled the presentation of a stimulus and whether its representation could be maintained in the network even as other modeled stimuli were presented. What they showed is that a particular dynamics occurred during "successful" working memory trials, in which both PFC and PPC populations represented the stimulus during presentation. During delay, presentation of a distractor would "switch" the PPC representation to representing the distractor, but PFC would not switch. After presentation of the distractor, the PFC → PPC long range connections would enforce the PPC populations to "switch back" to representing the remembered stimulus. This model predicted a range of physiological results found in PFC and PPC during these kinds of tasks, as well as predicted types and durations of distractor-presentation that would cause errors. Importantly, this explanation relied on the degree of internal connection in each area, and the feedback connections from the PFC to the PPC. If the internal connections within the PFC were not sufficiently strong, then it would not maintain the representation during distractor presentation.

For decision tasks, the investigators asked whether the network could produce an evidence-accumulation-to-threshold kind of process. These processes have been shown to be important for a range of decision-making processes including perceptual decision and multi-attribute choice (Teodorescu and Usher, 2013). They modeled a perceptual decision-making task, in which one out of a range of possible perceptual outcomes must be decided on in the presence of a noisy signal. They showed that the network could implement an evidence-accumulation process, in which buildup of evidence occurred primarily in the PPC population and selection of outcome in the PFC population. Intriguingly, these dynamics also were dependent on the degree of internal structure in the populations. Specifically, if the PFC population had a lesser degree of internal recurrent influence, the network would evolved towards a decision more slowly, whereas if it was strongly internally connected it would evolve very quickly. As predicted, at a higher degree of internal influence the network "decided" faster, which in turn contributed to more errors when the stimulus was noisier, and more time to integrate evidence would have been helpful.

So, the *very same* network could implement both the kind of information processing required in a WM task, and the kind required in a perceptual DM task. One of the most intriguing results, however, is that these two kinds of information processing *trade off* in a network. Greater resistance to distraction in the PFC network required a high degree of internal influence in that module. But a high degree of internal influence also shortened the timeline over which perceptual evidence could be accumulated. They posited that the particular structure of the PFC-PPC circuit helps ameliorate this tradeoff. In particular, if one removed the recurrent connections from the PFC to the PPC, then high performance in the decision-making task would result in lower performance in the working memory task.

I suggest that this kind of modeling project results in a mechanistic understanding of the network, but only by exhibiting the three properties I discussed above. First, the understanding of the circuit developed in the study *starts out* from the data point that both working memory and decision employ overlapping circuits. Second, understanding the informational requirements that are in common and differ across tasks is central to the explanation. In particular, there is something in common between working memory and decision-making tasks, namely that it is useful to have both a population that is multiple in its responses paired with a more categorically responding population. The difference in internal structure between the PFC and the PPC leads to the former having more univocal responses, which lead to both its robustness in WM contexts and its thresholding behavior in DM contexts. However, the *differences* between the tasks are also vitally important, because they illustrate the tradeoff in the network. WM contexts benefit from stronger interconnection, since it increases resistance to distractors. But DM contexts benefit from weaker interconnection, since it allows for increase in evidence-gathering. This in turn

leads to a mechanistic hypothesis, namely that the distinction between the PFC and PPC circuits in their degree of self-influence, and the feedback connection from the former to the latter, help ameliorate this tradeoff.

Importantly, because the explanation takes this form, it would be a *mistake* to attempt to spatially map WM and DM to distinct brain systems. There is not one part "doing" working memory and another part "doing" decision-making, and therefore there is not a causal relationship between so-individuated parts. There is one distributed circuit underlying those intuitively distinct functions. Given this, I submit, there is no metaphysically important distinction between working memory and perceptual decision-making. What there are are *distinct task demands*, and the ways in which those demands are implemented by a distributed system.

The heuristic view, on the other hand, simply doesn't assume that there is a fact of the matter about (i) whether WM is *really* distinct from DM, (ii) which tasks measure one versus the other, or (iii) whether a brain part really performs one rather than the other. What it suggests, as seems to be the case, is that there are deep commonalities in the brain systems performing these functions. The explanation also does not require that there be any firm division between the ultimate set of tasks that are WM, versus those that are DM, tasks. There are simply tasks with different informational requirements that are implemented differently in the network.

This is compatible with working memory and decision-making having played important heuristic roles in the understanding of this system. It was not, perhaps, initially obvious what the relationship between WM and DM might be. The concepts were operationalized differently. However, the persistent discovery of overlapping involvement in each of these tasks by the distributed PFC/PPC circuit led to the question of exactly how these functions are implemented. This led to a modeling project which uncovered both the commonalities and the differences between the informational requirements of, and the neural processing instantiating, tasks that correspond more-or-less closely to each of these categories.

I have only discussed one example, but I take this to be an exemplar of how multifunctional distributed circuits might be decomposed. I discuss a variety of other examples in other venues. If this case *is* exemplary, however, then it stresses the normative bit of the heuristic approach, as opposed to that of the standard framework.

6. Normative upshot.

6.1 For ontologies.

Whether each process instantiates attention, some form of memory, or something else, I contend, is not as important for explaining how the brain works as the informational demands of the particular task. The normative upshot for databasing projects is that more explicit attention needs to be paid to the kinds of behavioral paradigms at work and the way that they tend to vary (Figdor, 2010; Sullivan et al., 2021)

This is not to say that databasing projects have not paid attention to tasks. Projects such as the "Cognitive Paradigm Ontology" (Turner and Laird, 2012) are specifically intended to index the different types of behavioral experiments involved in studying cognition, and Sochat and colleagues (2017) have recently argued that standardization and publicity of experimental design is vital for coordinating study of cognition between laboratories. But the particular way in which tasks are approached in the field tends to strongly mirror the standard framework. For instance, Sochat et al. (2016) argue that the Cognitive Atlas should be "integrated" with a task ontology, but which they means by this is that each type of behavioral experiment should be categorized according to the kind of psychological function it is used to study. Standardization is important so that neuroscientists can "select paradigms based on the specific cognitive functions that they are thought to measure" (2016, p. 7).

The heuristic approach views the situation differently. While an investigator may "base" their search for behaviors to study on associations with cognitive functions that interest them, this basis is a heuristic one and not a measurement one. That is, a neuroscientist interested in "working memory," broadly speaking, should not find particular behavioral paradigms because they *measure* that explanatory construct, but instead as a way of looking for behavioral distinctions that may make a difference in how the brain processes information.

This has important upshot for how databases are constructed. In Neurosynth and the Cognitive Atlas, behavioral tasks are given abstract definitions, the assumption being that their role is to measure cognitive functions rather than themselves serving as explananda. In the Cognitive Paradigm Ontology and Brain Map, importantly, there are entries for experimental manipulations like epoch, stimulus, and response, but as far as I can tell these categories have not been standardized, organized for comparative analysis, or as rigorously codified as the mental constructs have been.

From the standpoint of the heuristic approach, this is generally insufficient. In the exemplars above, what makes a difference for explanation is understanding how the brain responds differentially to variations in task demands, from the presentation of the stimulus, often through a delay or across learning, to an eventual behavior. Differential responses to *these* variations are what explain how the brain implements the behavior. Other sub-fields of neuroscience employ different variations – for instance, other areas of decision-

neuroscience specifically vary reward types and regimes in conjunction with changes in stimuli and behavioral requirements. But entries in ontologies almost never contain detailed information of this sort, or at least that information is not standardized and codified.

Now, given the potentially infinite ways that behavior can be varied, codifying the kinds of behavioral and stimulus changes, the durations of temporal epochs and how they relate, etc., will be both a significant and a difficult task. But it is exactly the *point* of databasing projects to codify large amounts of data and make it accessible across individuals trying to explain things in a field. The normative bite of the novel view is that this project is more important than trying to find the "true specificity" between abstract mental constructs and brain activity.

This is not to *deny* that psychological concepts should be included in databasing projects. What it does suggest though is that the current state of the situation, on which most tasks are associated with a range of cognitive concepts, is neither surprising nor problematic. If the goal of these concepts is to serve as heuristics in search for behavioral paradigms is to index potential behavioral differences, rather than the reverse, then it is not surprising that these constructs should overlap both with each other and in the brain, and there is no need to try to theorize that overlap away.

So, suppose a neuroscientist is interested in 'working memory' or 'decision'. They approach a databasing tool in trying to devise research questions and develop experiments. What they find is that their construct of interest overlaps with other related constructs, and they find a huge number of extant behavioral distinctions that have been shown to make a difference in how the brain operates in, broadly, those cognitive contexts. This allows them to understand the behavioral distinctions that have been done, where brain activity has been measured in these conditions, etc. If they are interested in a particular part of the brain, they may find related behavioral measures that might clarify its function. If they are interested primarily in a construct, they may find a range of areas and distinctions that they could investigate, or use in the backdrop of forming new ones. The use of the psychological construct is a *heuristic for investigation*, rather than an explanatory claim.

This is, of course, an ideal, but I think it is an ideal that is importantly distinct from the current focus in databasing and brain mapping projects.

6.2. For functional connectivity analyses.

An exciting series of recent functional connectivity studies have begun the attempt to uncover dynamic principles for the entire brain during the course of behavior. Functional connectivity measures the co-activation of brain regions, with the assumption being that co-

active regions are coordinating in producing the relevant behavior. One can track changes in functional connectivity with changes in task or even across temporal epochs of the same task. One can then ask a variety of questions about the functional principles underlying these changes.

Here are just a few examples. Shine et al. (2016) hypothesized that network interconnectivity scales with *task complexity*. So, they measured functional connectivity in a range of tasks, and compared the overall degree of connectivity in the brain during each. They showed that "language" tasks or "working memory" (in this case, N-back) tasks resulted in higher levels of connectivity then more "simple" motor tasks.

Other studies have attempted to spatially decompose the brain into parts that help *organize* its dynamic changes across tasks, versus those that are in charge of performing those tasks. So, Shine et al. (2018) took whole-brain data across a range of tasks and applied principal components analyses to both the *spatial* and *temporal* data. They showed that the first principal component in the temporal domain correlated with activity in the spatial regions associated with the "rich club" network, and hypothesized that these regions underlie an across-task organizing function, whereas regions more closely responding to the other components were more task-specific.

This is cutting-edge work, and I do not wish to speculate too much on its eventual upshot. What I want to suggest, however, is that the heuristic approach offers distinct normative prescriptions than the standard framework for how to pursue further investigation. The standard framework suggests that these dynamical processes must be divided up, both spatially and temporally, according to distinct psychological faculties, and the causal relationship between those faculties explained. The heuristic view denies this necessity. Instead, it invocates the need to understand the structure of the tasks further – what distinctions in stimuli or behavioral requirements drive the different dynamic shifts, and what about the function of the brain networks involved enables them to enact those requirements specifically?

The authors of these studies seem to view them, at least in part, as stepping beyond the standard framework. As Shine and Poldrack note: "these results shifted the focus from where in the brain a particular function resides to how the coordinated recruitment of segregated specialist neural regions works together to accomplish the challenges associated with complex behavioral tasks" (2018, p. 396).

But it is important to note that, at least as of now, the invocations of the heuristic approach have not been followed. For instance, there is no analysis of what "complexity" of tasks amounts to in the Shine et al. (2016) paper. The notion is left intuitive. Nor is there any

analysis of what exactly makes tasks "language" tasks versus "memory" tasks in the other studies. This is a lacuna in these projects, according to the heuristic approach.

7. Conclusions.

A number of years ago, it was common for textbooks in philosophy of mind to teach the following: either our intuitive conception of the mind, with its commitments to intentional attitudes, etc., is true, or behaviorism is. I hope this strikes the modern reader as almost charmingly anachronistic. One can find newer versions of the dichotomy, however. Uttal (2001), in his famous criticism of fMRI research, argues that the alternative to discovering discrete localizations for distinct cognitive faculties is to view the mind as "unanalyzable," by which he means indivisible into distinct parts. More recently, Hutto et al. (2017) have suggested that the way to react to the "protean" – by which they mean dynamically reconfigurable – functionality of the brain is to embrace enactivist views of cognition, with their attendant rejection of mental representation and computation (Anderson, 2014; Silberstein & Chemero, 2013).

I have tried to argue that one can abandon the standard explanatory framework of cognitive neuroscience, and its attendant commitments about psychological constructs, without abandoning mechanistic explanation in the brain. And, while I haven't argued for it here, I claim elsewhere (Author's paper) that this general approach extends to representational explanation as well. The heuristic approach to cognitive ontology is a *very* different stance on explanation than is currently assumed in the literature, and I believe it deserves to be taken as a realistic option in this emerging field.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Anderson, M. L., Kinnison, J., & Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *Neuroimage, 73*, 50-58.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Bechtel, W. (2017). Using the hierarchy of biological ontologies to identify mechanisms in flat networks. *Biology & Philosophy*. doi:10.1007/s10539-017-9579-x

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account* (Vol. 2): Springer Science & Business Media.

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese, 193*(5), 1509-1534.

Bunzl, M., Hanson, S. J., & Poldrack, R. A. (2010). An exchange about localism. *Foundational issues in human brain mapping*, 49-54.

Burnston, D. C. (2019). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*. doi:10.1093/bjps/axz039

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *Neuroimage, 180*, 88-100.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355-376. Retrieved from http://rd.springer.com/article/10.1007/s11229-006-9097-x

Cummins, R. C. (1983). The nature of psychological explanation.

Darden, L., Pal, L. R., Kundu, K., & Moult, J. (Forthcoming). The Product Guides the Process: Discovering Disease Mechanisms. In E. Ippoliti & D. Danks (Eds.), *Building Theories*. Dordrecht: Springer.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese, 191*(2), 155-185.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences, 41*(2), 131-149. doi:10.1002/jhbs.20079

Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations: A Journal for the History and Philosophy of Science, 4*(1), 173-190.

Figdor, C. (2011). Semantics and metaphysics in informatics: Toward an ontology of tasks. *Topics in Cognitive Science, 3*(2), 222-226.

Fox, P. T., & Lancaster, J. L. (2002). Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience, 3*, 319. doi:10.1038/nrn789

Gomez-Lavin, J. (2020). Working memory is not a natural kind and cannot explain central cognition. *Review of Philosophy and Psychology*. doi:10.1007/s13164-020-00507-4

Griffiths, P. (2002). Is emotion a natural kind?

Hastings, J., Frishkoff, G. A., Smith, B., Jensen, M., Poldrack, R. A., Lomax, J., . . . Martone, M. E. (2014). Interdisciplinary perspectives on the development, integration, and application of cognitive ontologies. *Frontiers in Neuroinformatics, 8*.

Hutto, D. D., Peeters, A., & Segundo-Ortin, M. (2017). Cognitive ontology in flux: the possibility of protean brains. *Philosophical Explorations, 20*(2), 209-223.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc Natl Acad Sci U S A, 107*(25), 11163-11170. doi:10.1073/pnas.1005062107

Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an ontology of cognitive control. *Topics in Cognitive Science, 2*(4), 678-692.

Leonelli, S. (2012). Classificatory theory in data-intensive science: the case of open biomedical ontologies. *International Studies in the Philosophy of Science, 26*(1), 47-65.

Machery, E. (2009). *Doing without concepts*: Oxford University Press.

Murray, J. D., Jaramillo, J., & Wang, X. J. (2017). Working Memory and Decision-Making in a Frontoparietal Circuit Model. *J Neurosci, 37*(50), 12167-12186. doi:10.1523/JNEUROSCI.0343-17.2017

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese, 183*(3), 283-311. doi:10.1007/s11229-011-9898-4

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., . . . Bilder, R. M. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front Neuroinform, 5*, 17. doi:10.3389/fninf.2011.00017

Poldrack, R. A., Mumford, J. A., Schonberg, T., Kalar, D., Barman, B., & Yarkoni, T. (2012). Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Comput Biol, 8*(10), e1002707. doi:10.1371/journal.pcbi.1002707

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual review of psychology, 67*, 587-612.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2016). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*.

Robins, S. K. (2016). Optogenetics and the mechanism of false memory. *Synthese, 193*(5), 1561-1583.

Rubin, T. N., Koyejo, O., Gorgolewski, K. J., Jones, M. N., Poldrack, R. A., & Yarkoni, T. (2017). Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *PLoS Comput Biol, 13*(10), e1005649. doi:10.1371/journal.pcbi.1005649

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of learning and memory, 117*, 14-21. doi:https://doi.org/10.1016/j.nlm.2013.12.008

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., . . . Poldrack, R. A. (2016). The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance. *Neuron, 92*(2), 544-554. doi:10.1016/j.neuron.2016.09.018

Shine, J. M., Breakspear, M., Bell, P. T., Martens, K. E., Shine, R., Koyejo, O., . . . Poldrack, R. A. (2018). The low dimensional dynamic and integrative core of cognition in the human brain. *bioRxiv*, 266635. doi:10.1101/266635

Shine, J. M., & Poldrack, R. A. (2017). Principles of dynamic network reconfiguration across diverse brain states. *Neuroimage*. doi:https://doi.org/10.1016/j.neuroimage.2017.08.010

Silberstein, M., & Chemero, T. (2012). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences.

Sochat, V. V., Eisenberg, I. W., Enkavi, A. Z., Li, J., Bissett, P. G., & Poldrack, R. A. (2016). The Experiment Factory: Standardizing Behavioral Experiments. *Front Psychol, 7*, 610. doi:10.3389/fpsyg.2016.00610

Sullivan, J. A. (2010). Reconsidering 'spatial memory'and the Morris water maze. *Synthese, 177*(2), 261-283. Retrieved from http://rd.springer.com/article/10.1007/s11229-010-9849-5

Sullivan, J. A. (2014). Stabilizing mental disorders: prospects and problems.

Sullivan, J. A., Dumont, J. R., Memar, S., Skirzewski, M., Wan, J., Mofrad, M. H., . . . Prado, V. F. (2021). New frontiers in translational research: Touchscreens, open science, and the mouse translational research accelerator platform. *Genes, Brain and Behavior, 20*(1), e12705.

Tekin, Ş. (2016). Are Mental Disorders Natural Kinds?: A Plea for a New Approach to Intervention in Psychiatry. *Philosophy, Psychiatry, & Psychology, 23*(2), 147-163.

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological review, 120*(1), 1.

Turner, J. A., & Laird, A. R. (2012). The cognitive paradigm ontology: design and application. *Neuroinformatics, 10*(1), 57-66. doi:10.1007/s12021-011-9126-x

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*: The MIT Press.

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J. B., & Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS Comput Biol, 14*(11), e1006565. doi:10.1371/journal.pcbi.1006565

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods, 8*(8), 665.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci, 12*(6), 1100-1122. doi:10.1177/1745691617693393