Forthcoming in *Philosophical Psychology*

Title:
A Property Cluster Theory of Cognition

Author:
Cameron Buckner

Visiting Assistant Professor
University of Houston
513 Agnes Arnold Hall
Houston, TX 77204-3004
Phone: 713-743-3010
Fax: 713-743-5162

Humboldt Postdoctoral Fellow
Ruhr-University Bochum

cjbuckner@uh.edu

A Property Cluster Theory of Cognition

Abstract: Our prominent definitions of cognition are too vague and lack empirical grounding. They have not kept up with recent developments, and cannot bear the weight placed on them across many different debates. I here articulate and defend a more adequate theory. On this theory, behaviors under the control of cognition tend to display a cluster of characteristic properties, a cluster which tends to be absent from behaviors produced by non-cognitive processes. This cluster is reverse-engineered from the empirical tests that comparative psychologists use to determine whether a behavior was generated by a cognitive or a non-cognitive process. Cognition should be understood as the natural kind of psychological process that non-accidentally exhibits the properties assessed by these tests (as well as others we have not yet discovered). Finally, I review two plausible neural accounts of cognition's underlying mechanisms—one based in localization of function to particular brain regions and another based in the more recent distributed networks approach to neuroscience—which would explain why these properties non-accidentally cluster. While this notion of cognition may be useful for a number of debates, I here focus on its application to a recent crisis over the distinction between cognition and association in comparative psychology.

1. Introduction

What is cognition? Thus far, cognitive science has gotten by with highly vague answers to this question. At the onset of the cognitive revolution, cognitivism could be distinguished as the approach to the mind that appeals to the internal psychological states forbidden by radical behaviorism. Successive research programs have since defined cognitive processes as those that involve rules and representations, behavioral flexibility, skillful coping, and generally intelligent behavior. Regardless of doctrinal affiliation, we should all admit that none of these definitions were ever particularly good at issuing verdicts about the numerous non-paradigm cases—from animats to zombies—that have cropped up in internal disputes amongst the cognitivists. The problem has progressively worsened, as each new revolution expanded the reach of cognitive explanation by offering an account even vaguer than the last.

The increasing unclarity of 'cognition' has not prevented theorists from engaging in entrenched debates as to whether associative learning, perception, emotion, extended processes, or purportedly non-representational dynamical agents could count as cognitive. Granted, the increasingly toothless definitions opened up space for an explanatory pluralism in which different research methodologies could be explored (Chemero & Silberstein 2008). However, this methodological pluralism suggests a corollary taxonomic pluralism: that different sub-areas of cognitive science now presume distinct notions of 'cognition', and participants in these debates are talking past one another. The problem is that existing

theories are too nebulous to determine whether these borderline disputes are empirical or merely terminological disagreements. To do so, our notion of cognition must acquire more substantial empirical bite. The trick is to turn the question "Is *X* cognitive?" into a question with specific empirical consequences without repeating the mistakes of behaviorism by defining 'cognition' in terms of particular operational criteria.

To complicate matters, philosophers of science increasingly reject the deductive-nomological approach to explanation presumed by prior functionalist accounts of cognition. Instead, the critics argue, explanation in cognitive science appeals to underlying mechanisms or structures that cause the behaviors that result from cognitive processing (Bechtel 2009; Piccinini & Craver 2011). If this story is correct, then cognitive explanation requires a specific account of cognition's underlying nature—a challenge to which the vague conceptions are unequal. In short, not only has the need for a precise account of cognition never been more pressing, but the bar for the adequacy of such an account has never been higher. Perhaps what is now needed is not yet another shiny new paradigm, but rather a more careful examination of what has come before.

In this paper, I articulate one such worked-out account of cognition. I will not argue here that this is the only defensible notion of cognition, or that this notion should be applied to all extant debates—theses that are both well beyond the scope of a single paper, and moreover probably false. In some cases, one means little more by *"X* is cognitive" than that cognitive scientists should be allowed to study *X*. I take it, however, that many participants in these debates believe themselves to be engaged in genuine disagreements over the degree of underlying similarity between *X* and paradigm cases of human and animal cognition. If these debates are worth pursuing, then they must be framed with renewed clarity and precision. Here, I will focus exclusively on the implications my theory holds for one such debate, a recent crisis in comparative psychology over the distinction between cognition and association. Nevertheless, I hope the example both demonstrates that the notion articulated here has much to offer other debates, and illustrates what form other empirically adequate theories might take.

Devising any suitable theory of cognition is a daunting challenge, for by now many different capacities have been called 'cognitive', including at least conceptual abilities, cognitive mapping, transitive inference, episodic memory, numerical competence, sequence and serial position learning, causal inference, analogical reasoning, language use, imitation, mindreading, and metacognition. On the surface, these capacities appear diverse. Nevertheless, a minimally adequate account must show what they have in common in virtue of being cognitive—in short, that cognition is a natural kind. There is hope that this challenge can be answered, but only if we abandon the goal of analyzing cognition into a simple set of necessary and sufficient conditions. Some have argued that searching for a simple "mark" of cognition is wrong-headed, as cognitive processes are united only by family resemblance rather than by a shared essence (e.g. Allen 2006). But this worry does not entail eliminativism about cognition—for we may still apply an account of kindhood which allows cognition to be unified by a "cluster of marks" that establish the family resemblance.

The most promising and developed such account is the homeostatic property cluster (HPC) approach to natural kinds, a type of view endorsed by many but most associated with the work of Richard Boyd (1991; 1999). The HPC approach provides a way to distinguish categories suitable for empirical study from those that are not without relying upon essences. "Objects currently in the trunk of my car," "people whose surnames start with the letter 'N'," and "things about the size of a toaster" are probably not promising candidates for scientific study, whereas glial cells, episodic memories, earthquakes, chimpanzees, and uranium probably are. The latter sort of category are precious because they are (purportedly) loci of inductive potential, given that a large number of scientifically interesting properties non-accidentally cluster in instances of the kind. The HPC approach distinguishes itself by imposing the requirement that to count as a natural kind, a category must be the maximal class of items in which a significant number of scientifically interesting properties cluster due to the operation of at least one shared causal mechanism.

On my theory, cognition is identified by eight specific types of representational flexibility that enable agents to pass typical tests for cognition.[1] These properties constitute a scientific "stereotype"

which comparative psychologists use to distinguish cognitive from noncognitive processes. To ensure that the properties attributed to cognition are scientifically important, the cluster has been reverse-engineered from empirical practice in the area where this kind of diagnostic enterprise has reached its highest degree of sophistication: comparative cognition research. For if there is a problem of "demarcating" cognition, comparative psychologists face it every time they walk into the lab.

In Section 2, I sketch the structure of an HPC approach to cognition and distinguish it from less-promising alternatives such as purely behavioral or model-based strategies. In Section 3, I review the crisis over the distinction between cognition and association. In Section 4, I explain cognition's current scientific stereotype, reverse-engineered from influential experimental paradigms in comparative psychology. In Section 5, I review evidence that the properties comprising this stereotype non-accidentally cluster due to shared neural mechanisms, and thus mark out a natural kind, and in Section 6 I return to the crisis between cognition and association to show how the completed theory can help arbitrate difficult cases.

Before proceeding, two dialectical comments. First, in the interests of space I will not here defend the commonly-held assumption that cognition must be a natural kind to be empirically significant. Second, important questions have recently been raised about the HPC approach to natural kinds (e.g. Craver 2009); I adopt this approach despite these outstanding worries because putting the view in escrow until they can be answered in the abstract is the wrong way to proceed. Rather, the approach's viability should be assessed by seeing whether it can deliver verdicts on difficult cases, as cognition surely is. Thus, to work.

2. Preliminaries: How It All Hangs Together

Any adequate account of cognition must incorporate psychology's many moving parts—behavioral criteria, models, and underlying mechanisms—and show how they fit together. My goal in this section is to briefly explain how an HPC approach to cognition integrates these pieces by distinguishing it from two less promising alternatives: purely behavior- or model-based approaches (both of which are represented in the previous vague accounts).

A purely-behavior-based approach to cognition explicitly operationalizes cognition by defining it in terms of specific behavioral criteria, and anything that satisfies those behavioral criteria—regardless of how it satisfies them—counts as cognitive. There are a variety of well-known problems facing such approaches (Buckner 2011, p. 338-342; Adams & Aizawa 2008, p. 79-81), the general moral being that an adequate theory of cognition must specify not only what cognition enables agents to do, but also how it enables them to do it. A second type of model-based approach may thereby initially seem more plausible, where cognition would be defined in terms of specific styles of model, such as those built from rules and representations, links and nodes, or Bayesian networks. However, highly abstract model-based approaches offer little guidance on how to empirically assess the presence of cognition, for they are typically notational *lingua franca* with enough representational power to model any nonpathological inference (Buckner 2011, 326-329; Penn & Povinelli 2007, p. 105).

The solution to this dilemma is a proper integration of both behavioral and modeling criteria. The HPC approach is well-suited for this task, for HPC kinds are defined by an accommodation between our inductive and explanatory practices and the underlying mechanistic structures that explain how those practices could be at least approximately true or successful. In short, particular behavioral benchmarks grant us (imperfect) epistemic access to kinds of real phenomena, but only when constrained by specific models of how systems generate the relevant activity; and models can achieve the right degree of specificity by being selected as (plausible) minimal models of the underlying mechanisms that pass those specific behavioral benchmarks.

Applied to cognitive science, the idea goes something like this (Figure 1): cognitive scientists should collect the behaviors that they are interested in explaining as the result of cognition. They should then theorize about a minimal set of capacities that would allow systems to display these behaviors, and see whether agents possessing capacities that allow them to pass one set of behavioral tests also tend to possess the others. If it is plausible that they do, then scientists should attempt to develop a model of the underlying mechanisms that could produce those capacities and explain why they would tend to cluster together. The attempt to do so will spur an iterative process of multi-level theory revision which is robust

against limited numbers of counterexamples; scientists may find, for example, that one of the capacities produced by distinct mechanisms, and so should be excluded from the cluster (a "modest embarrassment"—Boyd 1999, p. 74); and having so reformed the cluster, they can in turn develop a more accurate view of the underlying mechanisms responsible for that flexibility. Cognition is defined by the limit of this iterative process of accommodation, by the minimal model depicting the kind of actual shared mechanism(s) that could satisfy the relevant behavioral tests.

Whether cognition can be associative[3], extended, or non-representational[4] on this approach becomes the question of whether an associative, extended, or non-representational system could achieve accommodation with a relevant property cluster. Whether one adopts the specific theory offered here, precisely articulating such property clusters and models of underlying mechanisms remains an essential step in determining whether borderline disputes amount to merely terminological or genuinely empirical disagreements.[6] To demonstrate the challenges facing such an account, I move to consider one of cognitive science's most difficult recent crises: the breakdown of the traditional distinction between cognition and association in comparative psychology.

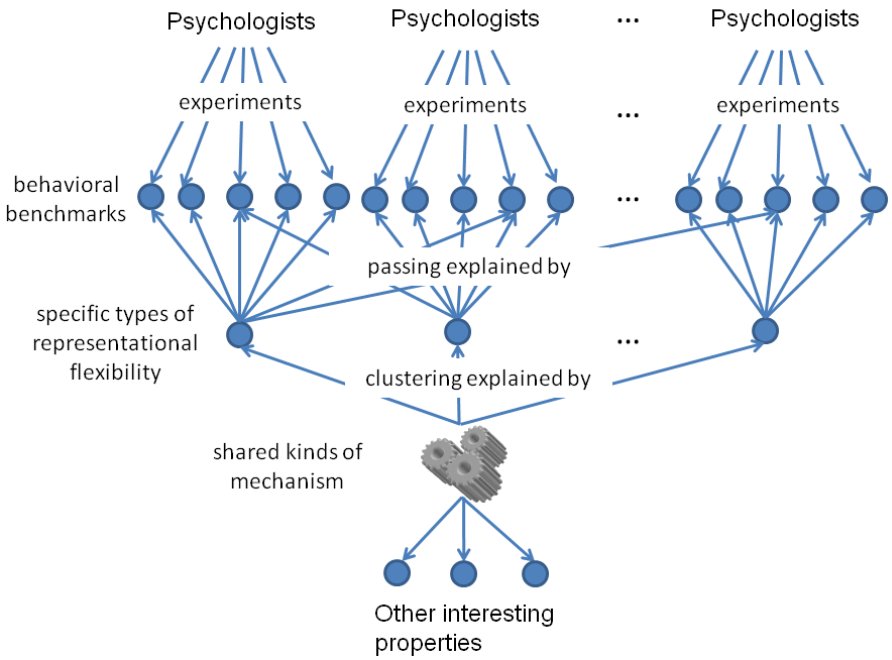

Figure 1. A schematic view of relationships between experiments, behavioral tests, representational flexibility, and underlying mechanisms. Typically, multiple tests exist to assess the same type of flexibility; and any one benchmark may require multiple forms to pass.

3. Cognition and Association: The Recent Crisis

Since the inception of comparative psychology, its methodology has relied upon a distinction between cognitive processes and those that stand "lower in the scale of psychological evolution and development"[7]—with the latter category including reflexes, innate-releasing mechanisms, imprinting, and, especially, associative learning (Wasserman & Zentall, 2006). This methodology can be characterized by three features. First, there is a default preference for associative explanations: producing a plausible associative explanation of some behavior is seen as a trump card which calls into question any cognitive interpretation of the results. Second, there is a default concern for associative explanations: since they are presumed to be ubiquitous in the Animal kingdom, associative processes must always be considered and ruled out for a cognitive explanation of some behavior to be regarded as legitimate. Third, these practices are only cogent if cognitive and associative explanations of behavior are mutually-exclusive alternatives—for if they are not, attempting to experimentally distinguish them would be confused. Let us call the approach to comparative psychology characterized by these three features its "Standard Practice."

Recently, some heavyweights of Standard Practice have challenged the viability of the distinction presumed by this methodology (Papineau & Heyes, 2006; Penn & Povinelli, 2007). For present purposes, it is enough to note that the tensions now surrounding the distinction largely stem from its supposed inability to neatly classify the most recent generation of psychological models. A number of these models are based in "associative principles" like spatiotemporal contiguities between stimulus cues, but can purportedly account for a variety of paradigm cognitive capacities like transitive inference and causal reasoning (e.g. Frank et al., 2003; Denniston et al., 2001). If these models are deemed 'associative' in the sense mutually-exclusive with 'cognition', then their expanded reach threatens the explanatory role of cognitive hypotheses in comparative psychology.

To consider a specific example, Penn & Povinelli wonder whether Denniston et al.'s Extended Comparator Hypothesis (ECH) should be classified as associative or cognitive. The ECH appears able to model higher-order backward blocking, an effect once thought to be a clear indicator of causal cognition. In the first phase of a backward blocking task, an animal is conditioned to learn that a compound of two cues (represented as 'AB') will be followed by another stimulus such as a reward (with the cue compound followed by the reward represented as 'AB+'). In the second phase, the animal is exposed to only one of the cues with the reward (i.e. A+). The backward blocking effect occurs when the exposure to the individual cue with reward (A+) in the second phase leads to diminished responding to the other cue in the original compound (B) during the final, test phase of the experiment. The cognitive interpretation of backward blocking supposes the animal to have learned that only one of the cues (A) was the true cause of the reward. Previous associative models could not exhibit backward blocking because they only allowed the modification of associations between cues presented together on a trial, whereas in backward blocking, learning in the second phase affects the strength of association to cues not present on that trial.

Several associative models were eventually able to exhibit the backward blocking effect, by exploiting the fact that A and B had spatiotemporally co-occurred in the past. Cognitivists upped the ante by demonstrating that these same animals exhibited the even more sophisticated higher-order backward blocking involving a third intermediary cue 'X' (e.g. AX+, XB+, B- | A?), feeling certain that it would not in turn fall to an associative analysis, since the blocked cue (A) never actually co-occurs with its blocker (B). The ECH, however, can model this effect, for it takes an animal's response to a given stimulus to be computed via a series of higher-order comparisons between associations amongst present cues and associations amongst other cues with which the reward has co-occurred in the past. This result might lead some associationists to triumphantly conclude that, once again, an apparently cognitive psychological process has been revealed to be instead associative in nature.

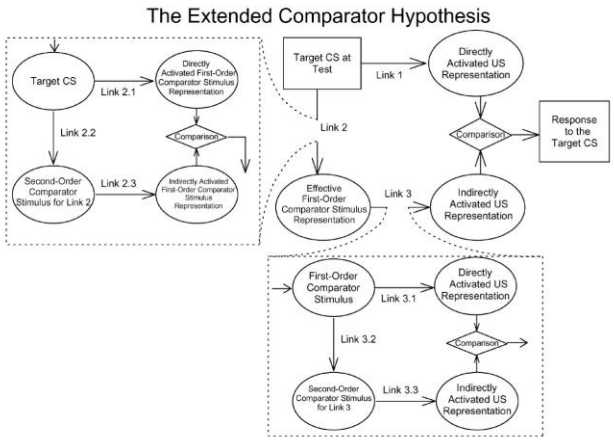

Figure 2. The Extended Comparator Hypothesis (Denniston et al., 2001). At the time of action, the likelihood that a given conditioned stimulus (e.g. a light) is the cause of an unconditioned stimulus (e.g. a reward) is computed via a higher-order comparison involving associations with other cues which have co-occurred with the unconditioned stimulus in the past.

The problem is that the ECH seems to satisfy criteria for both cognition and association. Penn & Povinelli note that on the associative side, the model is "based firmly on traditional associative principles like spatio-temporal contiguity and 'semantically-transparent associations'"; on the cognitive side, the model also "posits the kind of performance-focused, structured information processing that associationists have traditionally eschewed," such as an ability to understand the "'web of possibilities' that connects causes to effects" (2007, p. 100). While some may urge that the model's associative basis in contiguities between cues ought to settle the matter, it has long been accepted that cognitive processes may be accurately described by associative models at a "lower level of analysis" without this threatening their cognitive status (Fodor & Pylyshyn, 1988). Moreover, any model offers only a partial sketch of real mechanisms, and distinguishing "how possibly" from "how actually" models requires reference to real mechanisms (Craver, 2006). Associationists may hold that the ECH demonstrates that animals possessing only simple associative processing can clear a typical behavioral benchmark for cognition; but the real question is whether animals actually implement the ECH using a mechanism that is unable to satisfy other characteristic benchmarks for cognition.

In short, if the mechanisms described by this new generation of associative models can only pass behavioral tests for cognition by implementing cognition, then they do not impugn those tests' status as evidence for cognitive processing. Working out this compatibility as a conceptual possibility provides little methodological guidance for comparative psychologists, however, who must actually decide whether a proposed associative model ought to count as a deflationary alternative to some cognitive hypothesis or rather a description of a cognitive process at a lower level of analysis. In short, comparative psychologists need criteria that can be applied to determine whether a candidate associative model implements a cognitive process. Criteria have famously been offered in this context—Fodor & Pylyshyn champion compositionality and systematicity—but these criteria have found little uptake in comparative psychology as distinguishing marks of cognition. Moreover, it is not clear how to relate these criteria to those on which the scientists do rely—such as those described in the next section.

4. Meet the Cluster

In this section, I review eight properties commonly attributed to cognitive processes by comparative psychologists in their experimental practice. The general characteristic that nearly every test for cognition is meant to elicit is behavioral flexibility. The notion of behavioral flexibility simpliciter is too vague for our purposes, however, so I review the specific forms popularly assessed by comparative psychologists when they want to determine whether a behavior is caused by a cognitive or non-cognitive process. These tests assess not just any forms of flexibility—in some sense, random behavior is flexible—but rather forms that demonstrate an ability to adaptively and efficiently shape behavior to fit environmental contingencies.

Three orienting comments before reviewing the cluster: First, the cluster is offered not as a radical break with previous accounts of cognition, but rather as an explication of them. These properties are what enable a cognitive agent to master rules, flexibly satisfy goals, and cope skillfully. Second, it is a feature of the HPC view that not every property in the cluster need be present for a process to count as cognitive, nor will a process count as cognitive in virtue of exhibiting only one or two forms of flexibility in isolation. For example, basic forms of associative learning like conditioned taste aversion can occur

very rapidly. These processes will not count as cognitive on my scheme, however, for the range of cues eligible for contextual variation and rapid learning appear to be highly-constrained (taste-nausea links), and they appear to lack other forms of flexibility such as the capacity to be rapidly inhibited (cf. property 4.3). Moreover, the properties need only manifest on tasks which offer the right raw materials; so we should not expect to find e.g. monotonic integration (cf. property 4.7) on tasks that offer no monotonic orderings. Third, this list is not offered as an exhaustive account of the properties which cluster in cognitive processes; indeed, it is a corollary of the HPC approach to kinds that there be many more interesting inductive generalizations "waiting to be discovered" for a category to count as a natural kind. The current list is rather an inventory of cognition's current scientific stereotype—tentative, but sufficient to mark out a kind.

4.1 Context-sensitivity

One of the most important ways that a flexible strategy can differ from a stimulus-bound approach is by enabling its bearer to react differently to the same cue in different contexts. In contemporary psychology, context-sensitivity is a prerequisite for many of the most-studied cognitive capacities. Episodic memory is precisely a matter of putting experienced events in a what-when-where, autobiographical context. Cognitive mapping requires an organism to know where it is going, where it has been, and to be able to place its location in the context of visible landmarks. Transitive inference and serial position learning require attunement to contextual information of what comes before and after in the sequence. Complex tool use and manufacture often require the repetition of particular movements, in context, until a sub-goal has been achieved. Social cognition is especially context-sensitive, for the attention and mood of conspecifics can comprise a context of life or death importance.

Thus, experiments designed to distinguish cognitive from non-cognitive strategies often rely on tests of context-sensitivity. For example, consider Cheney and Seyfarth's well-known body of work on the predator calls of vervet monkeys (for a review, see Radick, 2007). Vervet monkeys emit distinct calls in the presence of pythons, eagles, and leopards. Cheney and Seyfarth argue that the emission and comprehension of these calls are governed by a genuinely semantic, cognitive competence, rather than

being purely instinctual or associative. Their argument is based on evidence that the calls—rather than being emitted invariantly to detected predators and producing rigid behavioral responses in listeners—are produced and consumed in contextually-appropriate ways. Using hidden loudspeakers, they played recorded predator calls in a variety of contexts. They found that responses to alarm calls depend on the current location of the monkey; an eagle call heard while in the treetops prompts running out of the tree, whereas monkeys located in the bushes hunker down for deeper cover. Vervets take into account the reliability of the signaler, generalizing appropriately to other contexts—repeatedly unreliable signalers of pythons (i.e. the "monkey who cried snake") are ignored not only when emitting that particular call, but also when emitting acoustically dissimilar calls with the same purported referent. Emission also depends on social context—isolated monkeys tend not to emit predator calls, high-ranking troop members are more likely to emit calls, and calls are more likely to be emitted when immature kin are nearby. All of these comparisons assess whether the vervets are able to vary the emission of and response to alarm calls appropriately by contextual cues.

4.2 Speed

The rapidity of learning and problem solving on novel tasks is also treated as an indicator of cognition. In contemporary psychology, rapidity criteria especially feature in problem-solving and tool-use experiments; the sudden emergence of a successful strategy on a complex task is still taken as a mark of "insight," "innovation," and "creativity" (Taylor, 2007). Hihara et al. (2003), for example, take a rapid spike in learning curves on complex tool-use tasks as evidence that their macaques solved a task through cognition rather than trial-and-error. The assumption is that trial-and-error and rote conditioning tend to be slow, whereas cognitive strategies make more efficient use of available evidence. Speed alone is not decisive of cognition—less flexible forms of learning like imprinting, conditioned taste aversion, and fear conditioning can be very rapid. But in conjunction with other forms of flexibility and on novel problems, speed can be the key to dissociating cognitive from non-cognitive strategies.

A venerable appeal to rapidity criteria is found in the "learning set" paradigm—which assesses an increasingly rapid error-rate reduction over a series of problems with different stimuli but governed by

similar logic. The increasingly rapid mastery of each new problem is taken as evidence the subject has "learned how to learn" by forming a "learning set," or category of problems sharing a similar structure. For example, Harlow showed that monkeys presented with successive trial blocks of two-object visual discrimination tasks improved their error-reduction rate over trial blocks, eventually achieving nearly effortless learning on new problems (Harlow, 1949). Though the cues which predicted reward varied from block to block, the monkeys learned to use the first trial on a new problem to figure out which features were important for that task. This improvement was taken to distinguish the subject that "adapts to a changing environment by trial and error" from one that achieves goals by "seeming hypothesis and insight." Such improvement in error-rate reduction is thought unavailable to simple associative and other strategies because the cues diagnostic of correct discrimination differ on each problem. For years, learning set formation was considered by many psychologists to be the strongest evidence of cognitive ability in non-human animals, and the error-reduction effect has since been observed in a wide variety of species, including cats, rats, rooks, and jays, to name a few (Slotnick, Hanford, & Hodos, 2000).

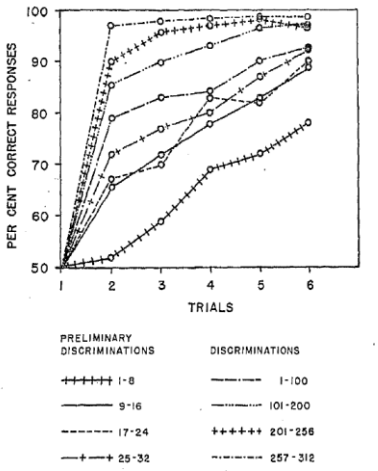

Figure 3. Improved rate of error reduction over a series of trials in a learning set task; different lines indicate different problem sets (Harlow, 1949).

4.3 Class formation

Another characteristic commonly attributed to cognition is the ability to group objects and situations into classes governed by a firm member/non-member distinction rather than by superficial perceptual

similarity. In comparative psychology, the persistent challenge has been to distinguish genuine conceptual abilities from associative stimulus generalization. Even the most radical behaviorists predict generalization from learned S-R links to novel stimuli, but this generalization should occur only along a gradient defined by psychophysical similarity of the stimuli. An empirical technique used to distinguish genuine categorization from stimulus generalization is thus to look for "rectangular" generalization gradients. Whereas subjects categorizing by stimulus generalization will show a gradual diminishment in response rate based on raw perceptual similarity of exemplars, the response rate of subjects imposing a member/non-member distinction should show a sharp drop-off when faced with stimuli that lack threshold values for core features. Moreover, animals that have learned a member/non-member distinction should fully generalize information learned about one category member to others, but not at all to perceptually similar non-members.

As an example, Watanabe tested whether pigeons could learn the concept of a triangle (Watanabe, 2006). Pigeons trained to discriminate a single triangle from three randomly placed lines showed a gradual diminishment of responding to exemplars along a perceptual similarity gradient, and thus did not provide evidence of concept acquisition. The response rates of pigeons trained on multiple exemplars, however, did show a sharp drop-off. In extensions of this research, Watanabe found that once trained on sufficient exemplars, pigeons could sharply discriminate the styles of particular artists (e.g. Picasso or Van Gogh), unfocused from focused pictures, and multi-chromatic from monochromatic canvasses—categories that require a complex integration of multiple cues to master. Pigeons also appeared capable of learning functional categories by demonstrating a rectangular generalization gradient in distinguishing "food" and "non-food" items, categories which are not united by any specific perceptual similarity but rather by core functional features like edibility.

Figure 4. Rectangular generalization gradient on pigeon "triangle concept" task (from Watanabe, 2006).

4.4 Higher-order and Abstract Learning

Another property used as a key indicator for cognition is the ability of subjects to master higher-order or abstract relationships in stimuli. In current psychology, such abilities have again been thought unavailable to non-cognitive strategies, since higher-order and abstract relationships cannot be learned by mechanisms sensitive only to first-order contingencies. A classic example involves a series of experiments in which Premack and colleagues attempted to teach Sarah the chimpanzee the concepts *same* and *different* (Gillian, Premack, & Woodruff, 1981). Sarah was given a series of analogy problems where correct choice was based on abstract relations between stimuli; examples included "banana is to banana peel as orange is to orange peel" (rather than peeled orange) and "marked paper is to marker as painted, closed can is to paint brush" (rather than can opener). In a modification of the experiment, Sarah was given two sets of objects, and correct responding required her to determine whether the objects in the two sets bore the same or a different relation to one another—another kind of task which cannot be solved by focusing on only first-order perceptual similarity. Sarah did well on both sets of experiments, leading Premack and colleagues to conclude that she possessed sophisticated cognitive abilities including analogical reasoning and the higher-order concepts of *same* and *different*.

To further control for associative mechanisms, researchers commonly rely upon a transfer test in conceptual discrimination problems, in which animals are trained on one set of stimuli and then tested on another which shares no systematic first-order perceptual similarities to the training stimuli. In influential work on concept-learning in pigeons, for example, Cook and colleagues trained pigeons to discriminate between arrays of six icons which were either all identical or on which one icon was different (Cook 2002). Pigeons routinely passed transfer tests involving same and different arrays containing entirely

novel icons, leading Cook to argue that they possess cognitive abilities which allow them to learn about higher-order relations amongst stimuli.

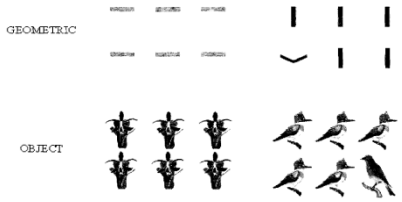

Figure 5. Example icon arrays used in same/different task by Cook (2002).

4.5 Multi-modality

Multi-modality is the capacity to integrate and process cues across different sensory channels. Philosophers and psychologists have often treated multi-modality as an indicator of flexible, centralized processing which occurs in a "common code," rather than peripheral processing bound to a particular type of cue. Contemporary psychologists have often agreed, taking multi-modal navigation strategies (e.g. combining vision, touch, and smell) as evidence of the true comprehension of space required for cognitive mapping (Arleo & Rondi-Reig, 2007). Integrating visual and haptic information is also required for complex tool manufacture and use, to monitor and coordinate subgoals. Developmental psychologists have long viewed cross-modal transfer as a crucial cognitive milestone in the understanding of space, number, and object location, and early multi-modal impairments predict broader cognitive deficits in both humans and macaques (Gunderson et al., 1990). These lines of thought all suggest that cognition is at least potentially multi-modal, and as such cross-modal transfer has been used as a distinguishing mark.

For example, Proops et al. recommend a cross-modal paradigm to distinguish true recognition of individual conspecifics from associative conditioning (Proops, McComb, and Reby, 2008). Proops et al. conducted an experiment on horses in which a familiar conspecific was lead past the subject and away from view. A vocalization was then played from a loudspeaker in the direction where the observed horse had exited. In the congruent condition, the vocalization was recorded from the horse that had been seen, and in the incongruent condition, the vocalization was from a different horse. Horses looked significantly more quickly, longer, and more often in the incongruent condition than in the congruent condition,

leading Proops et al. to suggest that the data are better explained by a cross-modal faculty of individual recognition than mere uni-modal association of cues. Whereas cross-modal recognition was long thought to be a uniquely human capacity, this paradigm has since been successfully applied to several primates, to crows, and is even being investigated in octopods (Kondo, Izawa, and Watanabe 2012; Tricario et al. 2011).

4.6 Inhibition

The ability to inhibit behavioral strategies in the face of changing environmental circumstances also suggests a strategy characterized by cognitive flexibility. Today, psychologists suppose that behaviors under cognitive control can be flexibly inhibited or reversed on the fly in response to evidence of changing environmental contingencies. As a result, good performance on "reversal tests"—in which a creature is trained to criterion on some discrimination problem, and contingencies governing correct response are then reversed—has been taken to indicate that a behavior is under cognitive control (Watanabe, 2006).

Heyes & Dickinson explicitly make the case for reversal criteria to determine whether behavior is truly 'intentional' (in the "caused by contentful psychological states" sense) or has merely appeared consistent with such an interpretation (Heyes & Dickinson, 1990). Specifically, they consider the behavior of Fodor's anecdotal "Graycat," who reliably approaches his food bowl during the usual feeding time in a way that facilitates digestion. To determine whether this behavior is actually caused by a desire for food together with the belief that the food bowl (at that time) is the place to find it, they recommend what they call the "looking glass" criterion:

> Suppose we placed Graycat in the world faced by Alice when she went through the looking glass; a world in which goals recede when you walk towards them, but draw nigh when you attempt to retreat from them…if Graycat's action is intentional he should, like Alice, adapt…by no longer attempting to walk towards his food bowl. (p. 90)

For desire, Heyes and Dickinson recommend an "irrelevant incentive" criterion, which assesses the diminishment of the target behavior when the purported desire responsible for the behavior has been

sated. The ability to adapt to changing circumstances by reversing previously learned responses in these ways, Heyes & Dickinson argue, would supply grounds for asserting that the behavior is mediated by causal beliefs and motivational states sensitive to environmental contingencies, rather than being a hardwired response.

4.7 Monotonic Integration

One of the most sophisticated tests for cognition relies upon the ability to distinguish elemental conditioning from strategies that organize cues along monotonic dimensions. A monotonic dimension is one along which stimuli can be ordered by increasing value—whether a "natural" dimension like size, intensity, distance, order, or duration, or a more dynamic one like a social dominance hierarchy. The ability to master such orderings plays a key role in a variety of cognitive competences. For example, the "short-cutting" test for cognitive mapping relies on the ability of animals to flexibly find a novel route to a goal in a maze when a familiar path has been obstructed; the information required to flexibly re-route is supposedly derived from a spatial ordering of locations and landmarks in an integrated, map-like representation of the environment. Monotonic integration also underlies a variety of timing and sequence-learning capacities, which depend upon the ability of the animal to detect patterns in temporal orderings of stimuli (Fountain & Doyle, 2011). Since monotonic ordering of cues across learning episodes has been thought beyond the reach of elemental associative learning, sensitivity to task-relevant orderings has been taken to be prima facie evidence that a psychological process is cognitive.

Work on transitive inference has especially focused on monotonic integration. In transitive inference tasks, animals are trained to make discriminations on successive sets of stimuli ordered along a monotonic dimension. The test phase of the experiment then presents the animal with a novel choice option and the experimenters assess whether the animal's choice is consistent with the ordering. For many years, the gold standard test for transitive inference was the five-element series. In the 5-element series, subjects are trained on equal numbers of four discrimination problems involving five successive stimuli (i.e. A, B, C, D, and E), each time with the stimulus prior in the sequence being rewarded. In other words, the animals are presented with equal numbers of AB, BC, CD, and DE discriminations, with A rewarded over B

(represented as ‘A+B-’), B over C (B+C-), and so on. Once reaching criterion on these four dyads, the subject is then tested on the novel BD dyad. Since B and D have been rewarded an equal number of times in the past, the organism responding on the basis of elemental associative conditioning should show no preference for B over D—and so an above-chance preference for B is taken as evidence that animals are capable of transitive inference.

Attention to monotonic integration highlights even more subtle concomitants of cognition such as the Symbolic Distance Effect (or SDE). The SDE is demonstrated when subjects can more rapidly and accurately make discriminations between exemplars which are distant along a monotonic ordering than between those that are nearby. For example, in the five-element series, animals demonstrating the SDE will make discriminations amongst A and E more quickly and accurately than those between B and D, even though the animal had not previously evaluated either pair, and even when there are no natural orderings amongst stimuli. SDE stands out as one of the few protocols to test an indirect “side effect” of the way information has been represented by an organism, and it has been demonstrated in humans, several primates, and corvids (Matsuzawa, 2009).

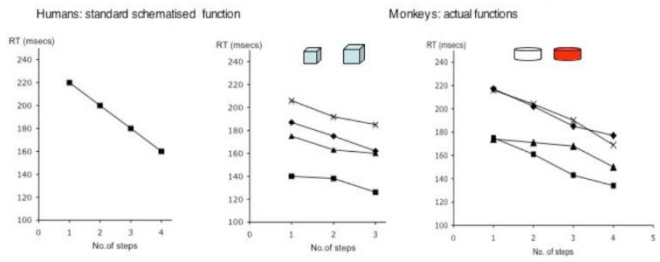

Figure 6. Reaction time evidence for the SDE; subjects demonstrating the SDE make discriminations amongst stimuli further apart in the monotonic sequence (larger no. of steps) more quickly and accurately than those amongst stimuli nearby in the ordering (McGonigle and Chalmers, 2010).

4.8 Expectation generation and monitoring

I round out cognition’s stereotype with the formation and monitoring of robust expectations regarding complex environmental regularities. Today, psychologists assess surprise by monitoring the amount of time a subject looks at an outcome, on the assumption that subjects will look longer at stimuli they

perceive as novel or unnatural. Preferential looking experiments are thought to dissociate cognitive from associative processes because learning takes place in the absence of any reward or unconditioned stimulus, and because the environmental regularities learned are thought too complex to be acquired by associative learning.

One area of research which has used looking times to assess causal understanding pertains to the "support principles" governing physical objects. Such principles include the idea that only certain types and amounts of contact between the surfaces of objects are sufficient for one to support the other. A good example in this line of work was conducted on rooks by Bird and Emery (2010). In these experiments, rooks could look through a hole at images which depicted a cylindrical plastic container standing in various spatial relations to a wooden platform. Bird and Emery varied the presence, type, and amount of contact between the cylinder and platform, exploring a variety of possible and impossible combinations. Images depicting the cylinder floating stationary above the wooden platform or hugging the side of the container should be perceived as impossible, whereas the cylinder resting on its long or short edge on top of the platform should be perceived as possible. Rooks looked significantly longer and more often at stimuli which showed physically impossible types and amounts of physical contact for support, leading Bird and Emery to conclude that their birds possess sophisticated cognitive understanding of support principles.

5. But Do They Cluster?

In the previous section, I characterized cognition by extracting a set of characteristics that comparative psychologists rely upon to distinguish cognitive from non-cognitive causes of behavior. If these practices are cogent, cognition will tend to produce behaviors exhibiting these forms of flexibility in relevant circumstances. Even if this is cognition's scientific stereotype, however, we have not yet secured the conclusion that this notion of cognition picks out a natural kind. The experiments canvassed so far provide little reason to suppose that these properties cluster. Given the overhead required to master a new experimental paradigm, most research groups have specialized in the study of only one or two kinds of test for cognition. This leaves little direct evidence of correlations between properties.

As a result, here is a skeptical view that has initial plausibility upon surveying the work discussed thus far: the term 'cognitive' is now applied to a motley crew of at least a dozen distinct capacities that share few commonalities. These capacities operate on different kinds of information and according to different principles; for example, a transitive inference problem requires the subject to order the relative values of multiple stimuli, whereas a concept learning task requires the identification of abstract features governing category membership. Furthermore, these capacities have been studied across a wide variety of species with different underlying neuroanatomy and facing different ecological pressures. Surveying all this apparent diversity, we might conclude that these capacities have only been grouped together for social or historical reasons—e.g., because none could be predicted by behaviorist theory circa 1970. While 'cognition' might have played an important role in liberating psychologists from the constraints of radical behaviorism, those strictures are now dead and gone, and we no longer need the label to protect us from them. Thus, scientists should eliminate 'cognition' from psychological theorizing—together with debates about its penumbra—and simply theorize directly about the various psychological capacities themselves.

To rebut this skepticism, we must provide evidence that, by and large, the properties grouped together in cognition's current scientific stereotype do in fact cluster, while at the same time making sense of the obvious diversity amongst capacities studied as cognitive. This is a challenge that even staunch proponents of cognitive approaches have not yet adequately appreciated. To discharge this burden on an HPC-style view, we must show both that the properties in the cluster are in fact mutually correlated, and identify a mechanism that explains why these properties non-accidentally co-occur (i.e. by reliably causing them to cluster).

In short, what is needed is an account of cognition as a "superordinate" natural kind that groups together a variety of more specific natural kinds of psychological process. Such superordinate kinds are common and can be useful to retain in theorizing and methodology even in mature sciences. Consider the classificatory term 'metal' in chemistry—a term that groups together many more specific natural kinds (gold, nickel, iron, lithium, etc.). Metals are distinguished from non-metals and semiconductors by their

tendency to display a core set of characteristic properties: they are good conductors of electricity and heat, and tend to be solid, opaque, lustrous, dense, and ductile at room temperature. Though metals were studied as a category long before the underlying explanation for these properties was known, we now know these properties cluster due to the operation of common causal mechanisms (Mizutani, 2001). Specifically, partially-filled outer electron shells cause atoms of metallic elements, at room temperature, to form a lattice governed by metallic bonds in which the outer electron energy bands of neighboring atoms overlap (e.g. allowing electrons to flow freely about the lattice). Despite the fact that different metals vary widely in their specific nuclear configurations, their outer electron shells all instantiate a more abstract kind of mechanistic structure that explains why they all tend to exhibit the characteristic properties of metals. At the same time, the theory predicts that metals will differ in the degree to which they exhibit these characteristic properties—for example, some metals are better conductors than others, and mercury is a liquid at room temperature. However, these variations and exceptions enhance the utility of the category—for the differences in e.g. conductance and melting point systematically arise from differences in these shared underlying causes (i.e. outer electronic configuration). In other words, the generality of a superordinate kind can enhance our ability to learn about the differences between its more specific children by generating hypotheses about the ways a child will differ from its siblings and how modulation of shared underlying mechanisms explains that variation.

So, the rebuttal to this skepticism is as follows: 'cognition' is a superordinate natural kind term that applies to a variety of more specific kinds of psychological capacities such as transitive inference, cognitive mapping, and conceptual abilities. Cognitive capacities are distinguished from non-cognitive capacities by supporting the specific forms of behavioral flexibility discussed above. They enable behavioral flexibility by operating on a distinctive kind of representation—one that is potentially context-sensitive, configural, abstract, multi-modal, highly-plastic, and responsive to perceptual evidence. The common thread here is the rapid and efficient encoding of interstimulus relations—whether contextual, higher-order, temporal, spatial, or multi-modal—to predict environmental outcomes. These representations are to be contrasted with the more hardwired, elemental, stimulus-response and stimulus-

stimulus representations which drive innate-releasing mechanisms, imprinting, and elemental forms of associative conditioning. There will of course be variation in the degrees of flexibility exhibited by cognitive processes across tasks, species, and individuals, but these differences can be explained by systematic variation in these underlying representational structures.

Let us begin with the evidence that these properties reliably co-occur. Meta-analyses have investigated correlations between the various forms of behavioral flexibility in the cluster, producing a body of knowledge is both vast and, at present, inconclusive. However, there are already enough data to justify optimism that the different forms of behavioral flexibility associated with cognition do in fact cluster and are enabled by shared mechanisms, for this hypothesis generates predictions that appear to be confirmed by the data. In a review of this literature, Lefebvre and Sol (2008) note that capacities for tool-use, innovation, imitation, learning latency, and reversal learning correlate across a variety of avian and mammalian species; test batteries run within species also show within-subjects correlations amongst different measures of cognitive ability. Moreover, if cognition were not a kind, one would expect to find a variety of inverse correlations between cognitive abilities, based on the assumption of biological tradeoffs between distinct capacities competing for resources. Such negative correlations appear to be rare (Lefebvre & Bolhuis, 2003).

Though suggestive, these behavioral correlations are agnostic on the question of whether there are shared underlying mechanisms that enable the various forms of representational flexibility. It could still be that organisms utilize entirely unrelated mechanisms to represent e.g. spatial and transitive relations, and the apparent clustering of spatial and transitive abilities may be due to statistical coincidence or definitional gerrymander. Indeed, the correlational studies have been dogged by just such worries (Giraldeau, Lefebvre, & Morand-Ferron, 2007). If this were the case, the different forms of behavioral flexibility might yet be supported by entirely distinct mechanisms, and cognition's current scientific stereotype would fail to mark out a natural kind.

Unsurprisingly, there has thus been intense interest in the neuroanatomical correlates of cognitive flexibility. Candidates include a variety of volumetric and connectivity ratios between body, whole brain,

pallium, cortex, and hippocampus. Many of these ratios offer impressive correlations with cognitive flexibility, but no clear winner has yet emerged. Moreover, mainstream empirical work on the neural correlates of cognition is currently divided between two dissenting camps: the traditional "lesion and localize" approach to neuroscience and the more recent "distributed networks" approach. As such, at present I sketch two optimistic lines of reasoning supported by current evidence that would rebut the skeptical hypothesis by homing in on the underlying mechanisms responsible for cognitive flexibility in animals. While readers may strongly prefer one or the other of these two views depending upon their attitudes towards this emerging controversy in neuroscience, I ultimately conclude this section by arguing that, when the details are filled in, there is less disagreement between them on the central question of this paper than may initially seem.

5.1 Proposal 1: The MTLs

The first approach bases the distinction between flexible and inflexible representations in the theory of multiple memory systems in neuropsychology—a linkage which has recently grown in popularity in comparative psychology (Smith et al., 2010; Fountain & Doyle, 2011). On this view, most animals possess dissociable memory systems which are specialized for the acquisition of representations of different degrees and kinds of flexibility. Procedural and declarative memory have long been dissociated at the behavioral level, with the former thought a more inflexible form of memory supporting habits and motor skills and the latter specialized for the rapid encoding of more abstract, fact-based knowledge. More recently, different memory systems have been localized to different neuroanatomical circuits, with architectural and neurobiological features adduced to explain the functional differences between systems (Squire & Schacter, 2002; Eichenbaum & Cohen, 2001; Rolls, 2000).

Thus, perhaps cognition's shared underlying mechanism—required by the HPC approach to natural kinds—can be identified with a particular neural circuit. In mammals, the medial temporal lobes (MTLs)—which include the entorhinal cortex, hippocampus, dentate gyrus, and subiculum—are repeatedly singled out as the crucial component of the neural system responsible for representations that enable cognitive flexibility. In one of the most influential such views, Eichenbaum and Cohen (2001)

argue that while many different brain regions are capable of elemental associative learning, the MTLs are uniquely specialized for detecting and encoding interstimulus relations. Scores of lesion studies support this hypothesis; since Hirsh (1974) first suggested that hippocampal lesions convert cognitive rats into stimulus-response automata, MTL structures have specifically been implicated in context-sensitivity, spatial learning, knowledge of temporal orderings and seriations, behavioral inhibition, relational learning, multi-modal integration, and configural learning (Gluck, Meeter, & Myers, 2003; Sweatt, 2004; Toates, 1998). Where the relevant neuroscience is known, these structures have also been directly implicated in a variety of cognitive capacities, especially transitive inference, declarative memory, conceptual abilities, and cognitive mapping (Dusek & Eichenbaum, 1997; O'Keefe & Nadel, 1978; McClelland, McNaughton, & O'Reilly, 1995). Strong evidence in favor of the anatomical memory systems view also comes from findings of competition between different responses on tasks that simultaneously activate different systems (Poldrack & Packard, 2003).

A focus on the MTLs, however, raises the question of whether we have adopted an unduly "mammalocentric" criterion for cognition. But here comparative neuroscience provides some solace; while much of the preceding evidence has been obtained from mammals, the basic framework may generalize to other classes. The best established extension is expressed in the verdict of homology between the mammalian and avian hippocampus, which reflects a recent sea change in thinking about the neural substrates of cognitive flexibility in birds (Columbo & Broadbent, 2000). Further bolstering this extension, a body of excellent, interdisciplinary comparative work has been conducted by Kamil, Balda, Bond, and associates investigating correlations between multiple forms of behavioral flexibility and hippocampal volume in corvids (Kamil, Balda, & Olsen, 1994; Bond, Balda, and Kamil, 2007). A recognizable form of the hippocampal formation dates back to fish, and Day presents a cladogram summarizing evidence that the formation supports a cluster of cognitive capacities across a swathe of vertebrates (Day, 2003). Where relevant forms of behavioral flexibility are found in even more phyletically distant species, such as the sophisticated problem-solving abilities of octopods or configural learning of bees, multiple forms of behavioral flexibility co-occur and have been found to depend upon

neural structures with architectures and patterns of connectivity explicitly declared similar to the mammalian MTLs by the researchers who study them (e.g. the octopod vertical lobe and the bee mushroom bodies—see Hochner, Shomrat, & Fiorito, 2006; Giurfa, 2003).

Whether these neural structures are similar enough to those of more paradigmatic cognition-capable organisms, remains, to be sure, an open empirical question. Perhaps octopods and insects will remain, like mercury with respect to metals, persistent but intelligible borderline cases. The challenge is to find the appropriate grain of detail to describe an underlying mechanism that could plausibly be shared across these diverse organisms. Popular connectionist or mathematical models of MTL function, such as that of Gluck & Myers (1993) or Howard et al. (2005) may offer the right grain. Nevertheless, given that neuroscientists are already making progress on these comparisons—and especially have already reached a consensus regarding the case of mammals and birds (Reiner et al., 2004)—optimism here is reasonable.

5.2 Proposal 2: A Cognitive Connectome?

The second type of optimistic view arises out of a recent paradigm shift away from the "lesion and localize" methodology favored by mainstream neuropsychology and towards a "distributed networks" approach to neuroscience (Anderson, 2010). Proponents of the latter approach are particularly unimpressed with the double dissociation criteria used as neuropsychology's stock and trade (Van Orden, Pennington, & Stone, 2001). A double dissociation occurs when a lesion in area *A* diminishes psychological capacity *X* but leaves *Y* intact, and a lesion in area *B* diminishes *Y* but leaves *X* intact. This is often assumed to provide evidence that capacity *X* can be localized to *A* and *Y* to *B*. But as even proponents of double dissociation concede, this reasoning only goes through if we can already presume that *X* and *Y* are distinct psychological modules which could be localized in distinct neural circuits—an assumption network theorists are typically not willing to grant. This creates an apparent problem for Proposal 1, for by calling into question the ability of lesion and localize to identify neural circuits responsible for cognitive processing, the network theorists would challenge much of the evidence that traditional neuropsychology could offer in support of the thesis that cognition is an HPC kind.

The alternative view they recommend—supported by meta-analyses of neuroimaging studies—holds that the neural substrates of most cognitive capacities are widely-distributed throughout the brain. Rather than expecting a one-to-one correspondence between psychological capacities and neural regions, any particular neural circuit may underlie a variety of distinct, domain-specific psychological systems. This thesis may appear to entail eliminativism about cognition because it denies that a simple neural region could serve as the shared underlying mechanism uniting cognitive processes. However, this view does not necessarily entail that there is no underlying mechanism responsible for the clustering of various forms of cognitive flexibility, for its proponents might identify this type of mechanism with an abstract network structure instantiated by many different networks throughout the brain. In other words, "cognitive process" might be an abstract superordinate kind term (cf. the discussion of "metal" above) that applies to the dynamics exhibited by many distinct distributed networks because they all instantiate a shared higher-order network structure that enables the forms of flexibility in the cluster.

Though our knowledge of the computational principles governing distributed networks is still in its infancy, there may be a type of network connectivity (a "motif" or "connectome" signature) distinguishing cognitive from non-cognitive distributed networks. For example, the networks supporting cognitive processing might be more widely distributed, involve more hierarchical integration of segregated resources, and feature more coordinating "hub" nodes than those supporting non-cognitive capacities (Sporns 2010, p. 179-206). Perhaps they offer the optimal distributions of in-degree and out-degree nodes—"talkers" and "listeners" (Zhao et al., 2011)—to facilitate synchronization between distant cortical regions required to learn the kinds of complex interstimulus relations involved in the cluster of representational abilities described above. Whatever the critical feature, the brain-body correlations discussed above would seem on the surface to offer stronger support for this kind of hypothesis compared to Proposal 1, given that hippocampal volumes are not nearly as well correlated with cognitive flexibility (predicting only 11.5% of variance in cognitive abilities) as are more inclusive ratios involving the entire cortex (which capture 99.5% of variance in birds and 98.5% in humans – Lefebvre, 2010).

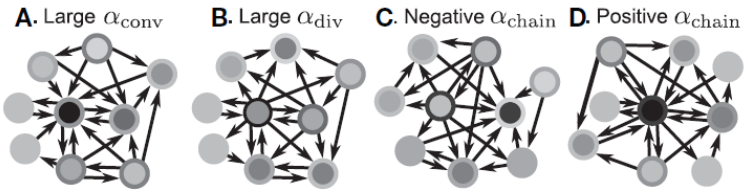

Figure 7. Different types of network connectivities (from Zhao et al., 2011). Arrows indicate synapses, and α_{div} and α_{chain} indicate the frequencies of high in-degree and longer multi-node synapse chains, respectively. The authors found that synchrony tends to increase with the relative frequency of chains and decrease with the relative frequency of convergence.

An apparent problem with this proposal, however, is that the connectivity factors (like α_{div} and α_{chain}) feature in these models as continuous variables. Rather than marking out the boundaries of cognition as a natural kind, they suggest a gradual progression from inflexible to flexible processing, and thus the failure of any neat accommodation between a cluster of cognitive properties and a unitary type of underlying neural mechanism.

Perhaps this is the correct perspective—whether it is remains an empirical question—but before giving up on this approach to cognition, we should remember that network models such as Zhao et al.'s are pitched at a high level of abstraction from actual physical systems. Notably, these models do not explain how certain neurons or neural assemblies might exhibit higher or lower in-degree and out-degree, and are silent on the constraints that different connectivity patterns impose on real physical systems that instantiate them. Viable widely-distributed networks must address a variety of challenging implementation constraints, including sufficiently powerful signaling mechanisms (e.g. why the MTL's perforant path contains mossy fibers), interference from nearby networks (e.g. why many axons are myelinated), and accessibly located coordinating nodes (e.g. why integration zones tend to be located in spatially central brain regions). Moreover, evolved agents can only satisfy such constraints by iteratively building upon or rewiring the architectures of their forebears. The next step, then, is to flesh out the neurobiological mechanisms that could feasibly implement such networks—an important possibility

being that the only networks capable of such flexible coordination will crucially involve the same areas of the brain highlighted by traditional localization methods.

5.3 Proposal 1 & 2: Ecumenical Coda?

Despite appearances, then, Proposals 1 and 2 may not really be at odds. Indeed, the MTL-centric theories of cognition already predict that the neural substrates of episodic memories and cognitive skills will be distributed, through the process of consolidation, to distant areas of the brain (McClelland, McNaughton, & O'Reilly, 1995). Thus, the two views could be mutually consistent to a large degree if there were a behaviorally-relevant difference in connectivity patterns between distributed networks which involve at least initial participation of the MTLs and those that do not.

Indeed, the various "distributed networks" theories should not and need not deny the behavioral implications of the distinctive neuroanatomy found in MTL tissues. Anderson's reuse theory, for example, is based on the idea that while a particular anatomical circuit may have many different domain-specific functions, it has only one common domain-general "working." These low-level workings are best uncovered through anatomical studies, and the data in this area suggest that the type of neurochemical dynamics in the MTLs uniquely predispose the area to rapidly attune to coactivations of input synapses along a variety of temporal scales (Nicoll & Schmitz, 2005). This working would leave the region uniquely well-suited to the discovery of complex interstimulus relations, which could be packaged into flexible representations and consolidated to diverse cortical networks.

Moreover, reuse may be required to explain why so many different domain-specific cognitive capacities have been associated with the same neural circuit (Petrov, Jilk, & O'Reilly, 2010). For example, suppose the hippocampal formation originally evolved to learn spatial relations amongst stimuli for the purposes of cognitive mapping. The neural workings required for such mapping would leave the system well-placed to learn temporal, transitive, and sequential mappings as well. These relations would in turn bootstrap a further reuse in episodic memory, since such memory is characterized by the ability to place events in spatial and temporal context. Indeed, such a sequence of reuse is now espoused by prominent members of the old guard (Nadel & Hardt, 2004). Finally, if this were the right story, it is

unclear that improved cognitive ability would require larger MTLs. Better or more representational abilities may not require bigger MTLs so much as better wiring (Roth et al., 2010), and at any rate studies of amnesia suggest that cognitive representations (such as episodic memories and concepts) only require MTL resources until consolidated to distributed cortical networks.[8]

6. Application and Conclusion

Supposing the view of the preceding sections to be coherent, let us apply it to the crisis over the distinction between cognition and association. This crisis was precipitated by our inability to determine whether recent psychological models ought to count as depicting implementations of cognitive processes, or rather deflationary associative alternatives to them. My proposal offers a way to answer this question: we should assess whether the new models depict mechanisms that can acquire and manipulate the kind of interstimulus representations that would enable the forms of flexibility in the cluster described above.

Consider the model of transitive choice developed by Frank et al. (2003) mentioned above in Section 3 (Figure 8). The model's creators offer it as a deflationary associative alternative to cognitive accounts of transitive inference because it construes animals in transitive inference tasks as making choices based on gradients of elemental associative strength rather than rule-based inferences. However, the model only exhibits transitive-like behavior on novel options because its network structure—inspired by the neuroanatomy of the MTLs—enables it to encode conjunctive representations of stimuli. This allows some of a stimulus' associative value to "transfer" to absent stimuli with which it has co-occurred in the past (Von Fersen et al., 1991). Granted, Frank et al.'s model makes different predictions than some other specific cognitive models based on SDE (e.g. on a 6-element series), and that is a good thing. But in implicating the MTLs, and especially by trafficking in conjunctive stimulus representations, their model predicts that organisms capable of transitive-like choice would also be capable of many of the other forms of flexibility in the cluster. So, while the model may offer a genuine advance over other cognitive models, it is not clear that it should be counted as associative in the deflationary sense. It rather counts as a description of a (perhaps distinct) cognitive strategy pitched at a lower level of analysis (see also Howard et al. 2005, p. 35). The same goes for the other problematic models such as the ECH (recall its higher-

order comparisons), which all distinguish themselves by trafficking in more complex interstimulus relations than traditional elemental models of associative learning (e.g. Estes, 1950). Thus, rather than demonstrating that characteristically cognitive tasks can be solved by less flexible, non-cognitive mechanisms, these models should be regarded as (partial) sketches of underlying mechanisms capable of satisfying the other common benchmarks for cognition.

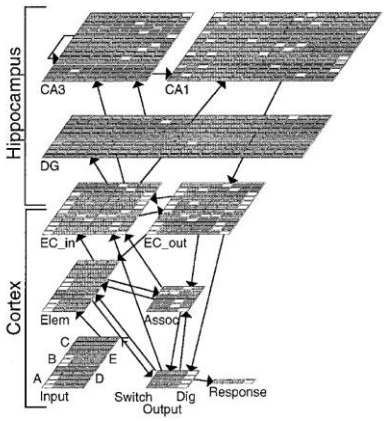

Figure 8. Frank et al.'s model of rat performance on transitive choice tasks. The hippocampal network forms conjunctive stimulus representations that allow the model to build an ordered associative gradient of stimulus values.

I conclude by noting that while the current account holds that cognitive processes are united by shared underlying neural mechanisms, the view does not require there to be anything magic about brains beyond their workings. In particular, we can envision prosthetics and artificial agents trafficking in representations that were flexible in all the right ways. As I suggested above, however, there are more constraints on actually building such systems than we might initially suppose. The current account moves such discussions away from vague notions of information processing or skillful coping to more specific models of plausible mechanisms that achieve a specific behavioral profile. Thus while recommending a critical attitude towards borderline cases of cognition, the present view rebuts charges of conservative prejudice by offering a principled account against which such claims can be empirically assessed. Though this yardstick for cognition may not be suitable for all purposes, it offers a precedent for others to consider in crafting alternative accounts.

Acknowledgments

I am grateful to Colin Allen, Jacob Beck, Jose Luis Bermúdez, Aaron Blaisdell, Robert Briscoe, Carl Craver, Fred Dretske, Jim Grau, Mitchell Herschbach, Celia Heyes, the attendees of the 2010 Winter Conference on Animal Behavior and Learning and a 2011 Pacific APA symposium, and several anonymous reviewers for discussion and feedback on earlier drafts of this paper. This work was supported in part by a grant from the Center for the Integrative Study of Animal and Behavior at Indiana University and by the Alexander von Humboldt Stiftung.

[1] In this paper, I will write in terms of "representations" and "representational abilities," but nothing turns on whether we construe these abilities as representational or in terms of the causal relationships that representations bear to their purported referents. Indeed, if naturalists such as Dretske have succeeded in reducing representation to types of causal **covariation**, then these are simply two different ways of describing the same thing. The important point is that cognitive processing allows agents to track specific types of environmental structure in the relevant ways. Hopefully, this account of representational flexibility articulates what is meant by those who suggest that "original" or "non-derived" representations are central to cognition (e.g. Adams & Aizawa, 2008). However, my own view of the explanatory work done by representational attributions in comparative psychology may surprise critics of representationalism (e.g. Ramsey, 2007), for I take the relationship to be more complex than one of mere "causal relay" (Buckner, forthcoming).

[3] The remainder of the paper might appear to defend the traditional distinction between cognition and association as strictly mutually exclusive. However, the current account holds rather only that at least some associative processing is non-cognitive, offering specific guidelines as to how to decide when a system built from associative principles would be sophisticated enough to count as cognitive. For more on this subtle issue, see Buckner (2011, p. 321-323).

[4] Some authors worry that representationalists take it to be analytic that cognition is representational, and so doing places undue constraints on empirical discovery (Ramsey, 2007). To clarify, the current account holds not that the kind of cognition discussed by comparative psychologists is necessarily (i.e. by definition) representational, but rather that it is contingently representational, that representationalism is the most plausible empirical hypothesis as to how humans and animals are able to pass the behavioral benchmarks that comparative psychologists associate with cognition.

[6] One may worry that the HPC approach's reliance on the identification of underlying mechanisms rules out non-representational dynamical models at the outset. While there is clearly some tension between dynamical and mechanistic approaches to explanation, recent work suggests that they can in some ways be complementary (Bechtel & Abrahamsen, 2010) and that dynamics themselves can constitute a mechanism (Zednik, 2011).

[7] This language comes from Morgan's Canon (Morgan 1903, p. 59). For a reasonable interpretation of the Canon, see Sober (2005).

[8] Notably, even cognitive networks that do not require online contribution from the MTLs for their typical operation may still require the MTLs for their initial organization. For example, though Wernicke's and Broca's areas (left-hemisphere regions implicated in linguistic comprehension and production in humans) have long been considered prime examples of functionally independent modules, children with early-onset MTL lesions on the left side of the brain develop right-lateralized or bi-lateral language-related brain activation, suggesting that even these regions depend for their organization at least in part on interstimulus relations discovered by the MTLs (Knecht, 2004). The independence of cognitive processing from MTL tissues after consolidation would also explain how patients with severe MTL lesions (e.g. H.M.) continue to display cognitive skills acquired before the onset of lesions (e.g. H.M. enjoyed doing crossword puzzles, chatting with experimenters, and watching television), but are impaired in the acquisition of new cognitive skills (Corkin, 2002).

References

Adams, F., and K. Aizawa. (2008). *The bounds of cognition*. Malden, MA: Blackwell.

Allen, C. 2006. Transitive inference in animals: Reasoning or conditioned associations? In Hurley and Nudds (2006), 175-185.

Anderson, M. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4), 245.

Arleo, A. & Rondi-Reig, L. (2007). Multi-modal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. *Journal of Integrative Neuroscience*, 6(3): 327-366.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscien*ce. Lawrence Erlbaum Associates.

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

Bird C., & Emery N. (2010). Rooks perceive support relations similar to six-month-old babies. *Proceedings Biological Science* 277:147–151.

Bond, A., Kamil, A., & Balda, R. (2007). Serial reversal learning and the evolution of behavioral flexibility in three species of North American corvids. *Journal of Comparative Psychology* 121(4):372-379.

Boyd, (1999). Kinds, complexity and multiple realization. *Philosophical Studies* 95, no. 1: 67-98.

Buckner, C. (2011). Two approaches to the distinction between cognition and 'mere association'. *International Journal of Comparative Psychology*, 24, 314-348.

Buckner, C. (Forthcoming). The semantic problem(s) with research on animal mindreading. *Mind & Language*.

Chemero, A, and M Silberstein. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science* 75, no. 1: 1-27.

Colombo, M., & Broadbent, N. (2000). Is the avian hippocampus a functional homologue of the mammalian hippocampus? *Biobehavioral Reviews* 24:465-484.

Cook, R. (2002). Same/different learning in pigeons. In: *The cognitive animal*, ed. M. Bekoff, C. Allen, & G. Burghardt, 229-238. Cambridge: MIT Press.

Corkin, S. (2002) What's new with amnesic patient H.M.? *Nature Reviews Neuroscience* 3(2):153-160.

Craver, C. (2006). When mechanistic models explain. *Synthese*, 153(3): 355-376.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology* 22:575-594.

Day, L. (2003). The importance of Hippocampus-Dependent Non-Spatial Tasks in Analyses of Homology and Homoplasy. *Brain, Behavior, and Evolution* 62:96-107.

Denniston, J., Savastano, H., Blaisdell, A., & Miller, R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In: *Handbook of contemporary learning theories*, ed. R. Mowrer & S. Klein, 65-117. Mahway: Earlbaum.

Dusek, J., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of the Sciences* 94:7109-7114.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain.* Oxford University Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3-71.

Fountain, S., & Doyle, K. (2011). Association and abstraction in sequential learning: 'What is learned?' revisited. *International Journal of Comparative Psychology*. 24(4), 437-459.

Frank, M., Jerry R., & O'Reilly, R. (2003). Transitivity, flexibility, conjunctive representations and the hippocampus: II. A computational analysis. *Hippocampus* 13:341-354.

Gillan, D.J., Premack, D., Woodruff, G. (1981). Reasoning in the chimpanzee. Analogical reasoning. *Animal Behavior Processes* 7: 1–17.

Giraldeau, L., Lefebvre, L., & Morand-Ferron, J. (2007). Can restrictive definitions lead to biases and tautologies? *Behavioral and brain Sciences* 30:411-412.
Giurfa, M. (2003). Cognitive neuroethology: Dissecting non-elemental learning in a honeybee brain. *Current Opinion in Neurobiology* 13:726-735.
Gluck, M. & Myers, K. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning.* Cambridge: MIT Press.
Gluck, M., Meeter, M., & Myers, C. (2003). Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends in Cognitive Science* 7(6):269-276.
Harlow, H. (1949). The formation of learning sets. *The Psychological Review* 56: 51-65.
Heyes, C., & Dickinson, A. (1990). The intentionality of animal action. *Mind & Language* 5:87-104.
Hihara, S., Obayashi, S., Tanaka, M., & Iriki, A. (2003). Rapid learning of sequential tool use by macaque monkeys. *Physiology and Behavior* 78: 427-434.
Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioural Biology* 12:421–444.
Hochner, B., Shomrat, T., Fiorito, G. (2006). The octopus: A model for a comparative analysis of the evolution of learning and memory mechanisms. *Biology Bulletin* 210:308-317.
Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review* 112, no. 1:75-116.
Hurley, S., and M. Nudds, eds. (2006). *Rational animals?* Oxford University Press.
Knecht, S. (2004) Does language lateralization depend on the hippocampus? *Brain* 127:1217-1218.
Kondo, N., Izawa, E., & Watanabe, S. (2012). Crows cross-modally recognize group members but not non-group members. *Proceedings of the Royal Society B* (online first).
Lefebvre L. (2010). Taxonomic counts of cognition in the wild. *Biology Letters.* doi:10.1098/rsbl.2010.0556 .
Lefebvre L., Bolhuis J. (2003). Positive and negative correlates of feeding innovations in birds: evidence for limited modularity. In: *Animal Innovation* (Reader SM, Laland KN, eds), pp 39–62. Oxford, UK: Oxford University Press.
Lefebvre, L., and Sol, D. (2008). Brains, lifestyles and cognition: Are there general trends? *Brain Behavior and Evolution* 72: 135–144.
Matsuzawa, T. (2009). Symbolic representation of number in chimpanzees. *Current Opinion in Neurobiology* 19(1): 92-98.
McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102:419-457.
McGonigle, B., and Chalmers, M. (2010). Insight and relational mechanisms. In Mills, D. S. ed., *The encyclopedia of applied animal behaviour and welfare*. CABI Publishing.
Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183-198.
Mizutani, U. (2001). *Introduction to the electron theory of metals*. Cambridge: Cambridge University Press.
Morgan, C. L. (1903). *An introduction to comparative psychology* (2nd Ed.). London: Walter Scott.
Nadel, L. and Hardt, O. (2004). The spatial brain. *Neuropsychology*, 18, 473-476.
Nicoll, R. A., & Schmitz, D. (2005). Synaptic plasticity at hippocampal mossy fibre synapses. *Nature Reviews Neuroscience* 6, no. 11 (November):863-76.
O'Keefe, J.., & Nadel, L. (1978). *The hippocampus as cognitive map*. Oxford: Clarendon Press.
Papineau, D., & Heyes, C. (2006). Rational or associative? Imitation in Japanese quail. In: Hurley & Nudds, 198-216.
Penn, D., & Povinelli, D. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology* 58:97-118.

Petrov, A., Jilk, D., and O'Reilly, R. (2010). The Leabra architecture: Specialization without modularity. *Behavioral and Brain Sciences* 33: 286-287.
Piccinini, G. & C. Craver. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183(3):283-311.
Poldrack, R., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia* 41(3): 245–251.
Proops, L., McComb, K., & Reby, D. (2008). Cross-modal individual recognition in domestic horses. Proceedings of the National Academy of Sciences 106(3): 947-951.
Radick, G. (2007). *The simian tongue: The long debate about animal language.* Chicago and London: University of Chicago Press.
Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.
Reiner, A., Perkel, D., Bruce, L., Butler, A., Csillag, A., Kuenzel, W., Medina, L., et al. (2004). Revised nomenclature for avian telencephalon and some related brainstem nuclei. *The Journal of Comparative Neurology* 473:377-414.
Rolls, E T. (2000). Memory systems in the brain. *Annual Review of Psychology* 51, no. 1:599-630.
Roth, T., Brodin, A., Smulders, T., LaDage, L., & Pravosudov, V. (2010). Is bigger always better? A critical appraisal of the use of volumetric analysis in the study of the hippocampus. *Philosophical Transactions of the Royal Society of London B* 365(1542):915-931.
Seyfarth, R.M. & Cheney, D.L. (2009). Seeing who we hear and hearing who we see. Commentary. *Proceedings of the National Academy of Science* 106, 669-670.
Slotnick, B., Hanford, L., & Hodos, W. (2000). Can rats acquire an olfactory learning set? Journal of Experimental Psychology: *Animal Behavior Processes* 26(4), 399-415.
Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology Animal Behavior Processes* 36, no. 1:54-65.
Sporns, O. (2010). *Networks of the Brain.* Cambridge: MIT Press.
Squire, L.R., & Schacter, D.L. (ed) (2002). *Neuropsychology of memory* (3rd edn). New York: Guilford Press.
Sweatt, J. D. (2004). Hippocampal function in cognition. *Psychopharmacology* 174:99-110.
Taylor, A., Hunt, G, Holzhaider, J., & Gray, R. (2007). Spontaneous metatool use by New Caledonian crows. *Current Biology* 17: 1504–1507.
Toates, F. (1998). The interaction of cognitive and stimulus-response processes in the control of behaviour. *Neuroscience and Biobehavioral Reviews* 22:59-83.
Tricario, E., Borrelli, L., Gherardi, F., and Fiorito, G. (2011). I know my neighbor: Individual recognition in octopus vulgaris. *PLoS ONE* 6(4): e18710.
Van Orden G., Pennington, B., & Stone, G. (2001). What do double dissociations really prove? *Cognitive Science* 25:111–72.
Von Fersen L. et al. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology Animal Behavior Processes* 17(3):334-341.
Wasserman, E., & Zentall, T. (2006). Comparative cognition: A natural science approach to the study of animal intelligence. In: *Comparative cognition* ed. S. Wasserman & T. Zentall, 619-36. Oxford: Oxford University Press.
Watanabe, S. (2006). The Neural basis of cognitive flexibility in birds. In: *Comparative cognition* ed. S. Wasserman & T. Zentall, 619-36. Oxford: Oxford University Press.
Zednik, C. (2011). The Nature of Dynamical Explanation. *Philosophy of Science*, 78(2), 238-263.
Zhao L, Beverlin B, Neto T, & Nykamp D. (2011). Synchronization from second order network connectivity statistics. *Frontiers in Computational Neuroscience* 5:1-16.