University of Alberta

Cognitive Homology: Psychological Kinds as Biological Kinds in an Evolutionary Developmental Cognitive Science

by

Taylor Shaw Murphy

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Philosophy

Fall 2012
Edmonton, Alberta

Abstract

In this philosophy of science thesis I develop a conceptual framework for thinking about cognitive activities by drawing on the conceptual resources of evolutionary developmental biology, providing first steps toward an evolutionary developmental cognitive neuroscience. Focusing on ontological, epistemological, and explanatory dimensions, I develop a concept of *cognitive homology* grounded in sound theory and scientific practice. A *cognitive homologue* is the same cognitive character under every variety of form and function. Extending operational criteria from biology, I analyze evidential relationships between empirical data in cognitive science and cognitive homologies. The explanatory contribution is twofold. First, it succeeds in providing explanatory grounds for causal relationships among *representational formats*, and here I focus on pretense and the imagination in philosophy. Second, my account clarifies the relationship between evolution, development, and evolutionary history, pointing to promising contributions in the context of current philosophical theorizing on the evolution of the human mindreading (mentalization) system.

Acknowledgments

Special thanks to my thesis advisor and mentor Ingo Brigandt for providing invaluable support throughout the preparation of this thesis. I would also like to thank Alan Love for his insightful comments and discussion during my stay at the Minnesota Center for Philosophy of Science, where I was present while writing the first three chapters of the thesis. This work was funded by the Social Sciences and Humanities Research Council of Canada Joseph-Armand Bombardier Canada Graduate Scholarship – Master's (SSHRC CGS-M #766-2011-0108), and Social Sciences and Humanities Research Council of Canada Michael Smith Foreign Study Supplement (SSHRC CGS-MSFSS #771-2011-1145).

Examination committee: Ingo Brigandt (chair), Michael Dawson, Howard Nye, and Francis Jeffry Pelletier.

Table of Contents

List of Figures

Introduction

The core of this thesis aims to contribute to philosophy of cognitive science by applying the homology concept (from biology) to psychological traits. Philosophy of cognitive science is a field in philosophy of science that centers on theoretical and methodological issues that are raised in psychology and the cognitive sciences. The homology concept from biology is one which underlies the individuation of biological characters (e.g. limbs, organs, cells), but normally psychological traits are not viewed as biological kinds but rather functional roles (*viz., functionalism*). In this thesis I make a case for the viability and fruitfulness of thinking of psychological kinds as historical, biological kinds characterized by the homology concept.

The homology concept is one of the most fundamental concepts in biology (de Beer, 1971; Brigandt and Griffiths, 2007; Donoghue, 1992). The homology concept originated in the first half of the 19th century, and was first given a clear theoretical account by Richard Owen in 1843 as follows:

> **Homologue**: "the same organ in different animals under every variety of form and function" (1843, 374)

As is plain in this definition, a homologue is something whose identity conditions are independent of both its form and function. Although Owen originally conceived of a homologue's identity being given by its instantiating an ideal archetype, since Darwin homology has been viewed in historical (i.e., evolutionary, developmental) terms. That is to say, sameness in homology is usually "defined by the common phylogenetic origin of the associated structures" (Bergeron, 2010), such that two structures A and B are homologous if and only if they are derived from a single ancestral structure C in a common ancestor.

The idea that homology is relevant for understanding psychological kinds has a short but significant history in the philosophy of cognitive science, specifically in the study of the emotions. Paul Griffiths (1997) argued that any scientific theory of psychological kinds that identifies emotions with functional roles is a theory of psychological *analogies* (sets of analogues). Analogues are grouped according to functional considerations, such as wings in insects and birds (which do not derive from a common ancestral structure), because of their

function in flight. Griffiths argues that, for the purpose of a good psychological theorizing of the emotions, the *causal depth* of theories of psychological analogies are inadequate when compared to a theory of psychological homologies. Generally speaking, analogies are often comparatively shallow in their underlying causal commonalities, whereas homologies share 'deep' causal commonalities. For instance, when comparing an insect and bird wing (two analogues), although there are superficial similarities in form and movement dynamics of these wings, the "deeper you dig" the less similar they turn out to be in their morphological components, underlying ontogenies (developmental mechanisms), and the mechanism underlying the activity of flight. In contrast, homology implicates commonalities in underlying causal commonalities. If one takes two homologous characters (such as the bat wing and human arm) one often finds that they are highly convergent in underlying commonalities (bones, tissues, their organization, developmental mechanisms, and so on). These causal similarities also include computational similarities relevant for understanding psychological processes.

Since psychology and neuroscience are in the business of identifying mechanisms underlying behaviors, Griffiths argues, the adequacy of a kind of category (analogies or homologies) depends on its value in studying such underlying mechanisms. Included here is the condition that the causal and computational properties of a kind ought to be projectable (Goodman, 1983), supporting robust generalizations and ampliative reasoning about the properties of the members of the kind. Because analogy does not tend to group together deep causal and computational commonalities and homology does indeed tend to do so, emotions are therefore best viewed as homologues rather than as functional kinds like analogues. In summary:

Argument from causal depth

1. Homologues tend to be highly similar in underlying causal commonalities, and so tend to support robust causal generalizations.
2. Functional kinds (analogues) do not tend to be highly similar in underlying commonalities, and so do not tend to support robust causal generalizations.

3. A scientific theory of the emotions is best served by studying kinds that support robust causal generalizations.
4. So, the emotions should be viewed as homologues rather than functional kinds.

Although this argument potentially extends to all psychological kinds, most of the subsequent literature on particular applications of the homology concept in cognitive science has remained focused on the emotions (e.g., Clark, 2009, forthcoming; Charland, 2002; Griffiths, 2003). Why is it that the "bulk of the debate has centered around 'emotion' as an example of a psychological category ripe for reinterpretation within this new framework of classification" (Clark, 2009, 76)? There is much about emotions that make a homology construal particularly apt in this case, but why has it not been used to understand any other particular psychological kinds? One particularly salient reason for this has to do with the *kinds of comparisons* we are often interested in and how the homology concept is presently understood.

With emotions and the scientific study of emotions, much of the interest has been on using nonhuman animal models for understanding human emotional processes, and in these cases cross-species identification of homologies among emotions is particularly important. For instance, homology of fear processing in humans and rats legitimizes the study of rats as an animal model for human fear processing due to considerably conserved causal structures afforded by homology. In contrast, for many other psychological kinds there is more of an interest in understanding relationships between cognitive traits *within* one and the same individual, such as the relationships between language and music processing, working memory and episodic memory, etc.

The problem is that the notion of homology is well understood in the cross-species context but is comparatively poorly understood as a concept applied within one and the same individual. Complicating these matters is also how homology is only particularly well understood for morphological structures, which differ markedly from cognitive systems; in contrast, cognitive systems are *spatiotemporal* organizations of parts, and the identification of homology in spatiotemporal systems is poorly understood. It is the intersection of within-individual homologies and spatiotemporal systems in psychology that

makes the use of the homology concept difficult to employ in fruitfully understanding other psychological kinds.

This thesis aims to develop a concept of "cognitive homology" that is useful for understanding relationships between psychological processes within one and the same individual. The concept of homology that is germane for this is one found in the field of evolutionary developmental biology. Evolutionary developmental biology is an emerging field that explores the intersection of two fundamental processes in biology: the *development* of individual organisms (ontogeny), and the *evolution* of traits in the history of life (phylogeny) (Laubichler, 2008). The field of evolutionary developmental biology has its own methods, approaches and research questions but draws from diverse areas including development, evolution, paleontology, ecology, and molecular and systematic biology (Hall and Olson, 2003). The concept of homology from evolutionary developmental biology is one that focuses on the causal-developmental structure of an individual in such a way that allows for understanding 'deep' causal and computational relationships between psychological characters within one and the same individual. It is the notion of *serial homology*, which involves a repetition, duplication, or "redeployment" of an existing structure within an organism. A s*erially homologous cognitive character*, then, is the same cognitive character as it is redeployed within an individual under any variety of form and function – to borrow Owen's phrasing.

In elucidating a cognitive homology concept I aim to show how it can apply to psychological kinds by analogy from current understanding of the concept in biology, which historically speaking has primarily focused on cross-species homologies of morphological structures. In Chapter 1 I provide a discussion of the necessary background for understanding the notion of homology. Chapter 2 focuses on theoretical and methodological issues that arise from the notion of *cognitive homology* in particular. In Chapter 2 I draw on the way homology is understood in biology in order to bring it to bear on psychological kinds, tying the notion in with a number of scientific theories and research programs.

The focus of Chapters 3 and 4 are on applications of the cognitive homology concept to theories proposed by philosophers of cognitive science.

Chapter 3 engages with the relationship between imagination and belief. Shaun Nichols (2004) has argued that the reason why belief and imagination representations are processed in much the same way by cognitive mechanisms is that they are in a "single code," but leaves it unclear what this is or how it explains these similarities; indeed, "little if anything has been written about the criteria of sameness or difference for such codes" (Goldman, 2012, 73). I argue that cognitive homology has the requisite explanatory resources to explain these similarities and can ground an understanding of what it is to be a "code" in cognitive science. Chapter 4 engages with the architecture of the mentalization (mindreading, "theory of mind") system and how evolutionary developmental considerations bear on reasoning about the architecture of this system. I argue that the current focus on selection pressures as providing evidence to decide between "interpretive sensory access" (Carruthers, 2011) and "inner sense" (Goldman, 2006b) architectures depends on certain evolutionary developmental assumptions that cognitive homology can be used to clarify. In this chapter, I suggest that the core mechanism for mentalization is a homologue and that, under certain assumptions, the "default network" may be this homologue.

Chapter 1 – Background:
Homology from biology to psychology

Homology in biology

Richard Owen originally defined a homologue as "the same organ in different animals under every variety of form and function" (1843, 374). Formally speaking, homology is a binary relation, and identifies the *same* organ in different animals. Being a relation of identity (rather than similarity), it is transitive, and so a 'homology class' contains the set of traits that are all homologous. The mammalian forelimb is an example of a homologue, and this homology class contains all the mammalian forelimbs (Figure 1.1).

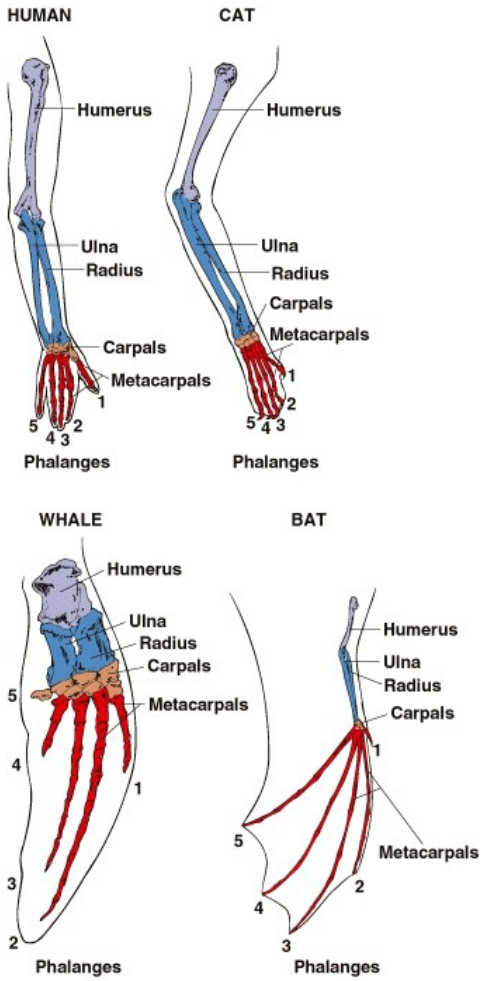

Figure 1.1
Mammalian forelimb as found in humans, cats, whales and bats.

The mammalian limb is depicted for humans, cats, whales and bats. Also included in this figure are the bones found within the limbs. Homology is the phenomenon of an organism being composed of homologues, and this example shows how homologues are *hierarchical* structures, such that homologues are decomposable into further homologues; the mammalian limb is composed of bones (homologues) and other structures, and these are in turn composed of cells, etc. However, it is not true that homologues always have exactly the same parts or emerge from identical developmental pathways or have identical genes. For example, the neural tube is homologous in vertebrates, but develops quite differently in fish and mice—a phenomenon known as developmental systems drift (True and Haag, 2001; Robinson, 2011). In general, homology at one level does not necessitate homology at another level; there may modifications or changes in developmental mechanisms, variation at the genetic level, as well as variation in the parts and configuration of parts that compose the homologue. Despite tending to be highly similar, homologues can be dissimilar in a wide variety of ways.

An important distinction here is between *characters* and *character states*, where character states are variable properties of evolutionary characters. For instance, the human eye (a character) may vary in eye color (its character state). The variation in size and shape of the bones of the mammalian forelimb are another example of variation in character states—as illustrated by the variation in size and shape of the humerus, metacarpals and other bones in Figure 1.1. "Character states" are the varieties of "form" that characters/homologues can take on.

Individuals are composed of homologues/characters that can vary in character state and this reflects how phenotypic evolution can be studied on a *character by character basis*, with homologues being the bearers of such morphological variation:

> The fact that phenotypic evolution can be studied on a character by character basis suggests that the body is composed of locally integrated units. These units can be considered as modular parts of the body which integrate functionally related characters into units of evolutionary transformations. (Wagner, 1996, 36)

The idea that homologues can be considered as *modular* parts of the body suggests that what makes two structures distinct homologues/characters in an organism is that they may vary in a more-or-less independent fashion from one another, permitting evolutionary change on a character by character basis:

> Homology is the phenomenon of an organism being composed of several homologues/characters, where one such character can phenotypically vary and evolve independently of the others (evolvability on a character by character basis). (Brigandt, 2007, 710)

In biology there are several different conceptions of homology, including transformational, taxic, and developmental accounts. The *transformational* approach owes to an interest in how characters evolve and change in their character states over the course of evolution, and has a home in evolutionary taxonomy (Donoghue, 1992). The *taxic* approach is interested in how a derived character (different from the more ancestral condition) originating in a species is inherited and shared by all descendant species and thus characteristic of the taxon containing these species (synapomorphy). Homologies are viewed as evidence for the monophyly of taxa (Ereshefsky, 2012). The taxic approach reflects an interest in marking the boundaries of taxa and so has a home in cladistics or phylogenetic systematics (Donoghue, 1992). Both the taxic and transformational approaches can be grouped together under the banner of *phylogenetic homology*, as the unifying feature for both is that the 'sameness' of homology is defined by the common phylogenetic origin of the associated structure (Wagner, 1992; Brigandt, 2002). In other words, what makes two structures the same (homologous) is that these structures are derived from a common ancestral structure. For example, eyes evolved independently in cephalopods (e.g., squids) and vertebrates (e.g., humans) and so there is no common ancestor of vertebrates and cephalopods with eyes. Consequently, the vertebrate and cephalopod eye are not homologous.

The *developmental* approach to homology in contrast serves to highlight ontogenetic mechanisms that retain and constrain evolutionary characters and has a home in evolutionary developmental biology. A developmental account of homology appeals to causal-developmental factors in determining homologies. For instance, two traits are homologous if they are caused by a special common

developmental module among other variable developmental processes (Wagner, 1996).[1] Developmental approaches to homology are in tension with phylogenetic accounts, which in contrast do not require a same developmental module while requiring a common ancestor, and this has led to skepticism over the coherence and usefulness of developmental accounts of homology (Cracraft, 2005). Serial homology (the focus of this thesis) involves the repetition or duplication of an existing trait within an individual, and so falls under a developmental approach to homology.

Spinal vertebrae are an illustrative example of serial homology in morphology, as in Figure 1.2 below. On the left is a human spine and on the right are idealized drawings that represent characteristic morphological components of ranges of spinal vertebrae. The vertebrae are the same structure owing to their shared causal-developmental properties, and the vertebrae exhibit variation in form along the vertebral column. Token vertebrae within an organism are serially homologous. Cervical, thoracic and lumbar vertebrae are serially homologous types of vertebrae.

Serial homology used to be seen as very important in biology, and identification of homologies between organisms and within organisms were part of the same project of identifying the parts of organisms (e.g. by Owen). However, with the advent of evolutionary theory, serial homology became much less prominent. When the task is to sort out evolutionary relationships between organisms in phylogenetic systematics, or to understand the adaptive modification of evolutionary characters, developmental approaches to homology are less relevant than phylogenetic approaches. Phylogenetic accounts of homology thus rightfully gained central importance, and consequently serial homology has received much less theoretical interest and is less well understood. But since the interest here is homology within an individual, serial homology is of central importance for thinking about homology and cognitive systems.

[1] One example of this are gene regulatory networks that function as character identity networks (ChINs) present in any instance of a character across species, such that other genes (not part of the character identity network) underlie the variation in the character state of that trait (Wagner, 2007).

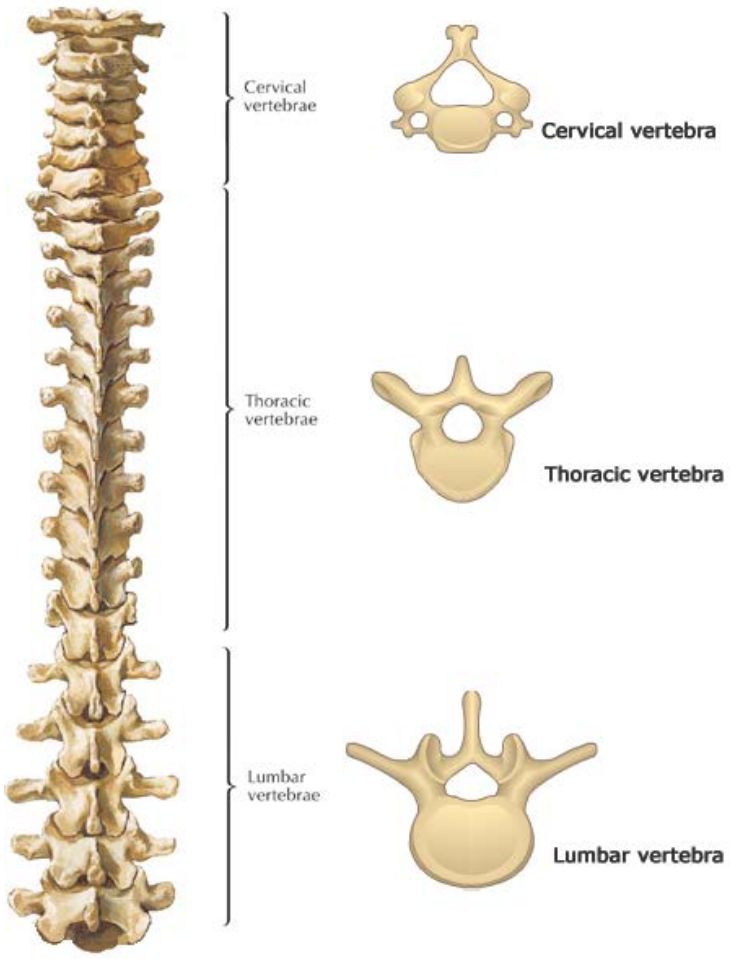

Figure 1.2

Spinal vertebrae as illustrations of serial homologues. On the left is a drawing of a human spine, and on the right an idealized drawing highlighting some common features of cervical, thoracic and lumbar vertebrae.

Operational criteria have been developed for identifying homologies, most prominently Adolf Remane's (1952) classic criteria for identifying morphological homologies (that is, for bones, tissues, cells, and other morphological features). These criteria represent the categories of evidence that can be used to support identification of homologies, and have been used to establish homologies independent of the approach used (i.e., for both phylogenetic and developmental homology). Remane's operational criteria are as follows:

Position: the relative position of a trait within a more general pattern of organization.

***Continuity**:* identifying a continuum of evolutionary intermediates, from a simpler and earlier character state to a more complex and derived state.

Special Quality: the complexity, distinctiveness, or specialization of that trait.

The first criterion is relative *position*, which refers to a correspondence in a more general pattern of morphological organization, including for instance adjacent tissues, organs, and bones. It is evidence that two bones in different organisms are homologues if they both occupy corresponding relative locations among other bones and structures. An example of this is the carpals in the forelimb of different mammals (Figure 1.1), which are such that they are situated between the ulna and radius on the proximal side, and metacarpals on the distal side. The criterion of *continuity* of evolutionary intermediates consists in identification of a series of character states between morphological structures in different species, thereby tracing a continuum of character states from its ancestral state to its derived state, for instance as established through a fossil record. Finally, the criterion of *special quality* a broad category which is motivated by the consideration that the more specialized and distinctive a character state is, the less likely it is to have evolved more than once (Matthen, 2007)—such shared features are unlikely to be due to convergent evolution (arising from the occupation of the same ecological niche). For instance, a correspondence in distinctive color, cell type, or internal organization of structures in two species provides evidence that they are homologous.

Homology and functional kinds

In biology, homology is usually contrasted with *analogy*, a notion which is tied to functional considerations. Two organs are analogous if they have a *similar* structure owing to common selection pressures. Because the analogy relation is one that marks out *similarity* rather than *sameness*, analogy is not a transitive relation. When two organisms inhabit similar ecological environments and are exposed to some of the same selection pressures, these structures tend to exhibit convergent evolution and response to these selection pressures. For example, insect and bat wings are analogous due to their similar function in supporting flight, as is the camera eye of vertebrates and cephalopods—these structures bear some similarities owing to their common functional demands.

Despite arguments from causal depth in favor of using homologues rather than analogues in scientific theorizing in psychology (discussed in more detail below), it would be a mistake to conclude that analogy is always unable to capture large amounts of causal depth or be scientifically useful. The case of the similar body morphologies of tunas and deepwater sharks shows how this is so (Donley et al., 2004). Despite 400 million years of independent evolution, deepwater sharks and tunas independently converged on highly similar body morphology related to locomotion, likely due to being under similar selective pressures for fast and continuous locomotion. These similarities do not just extend to the gross bodily shape of tunas and deepwater sharks but actually extend into fine-grained muscular organization and the mechanical dynamics of propulsion movements (Donley et al., 2004, 61), traits which are distinctively found in just these species.

In light of this, both analogy and homology should be recognized as valuable concepts in scientific reasoning, and that both *can* point to strikingly deep similarities. Generally speaking, in cases of *short* phylogenetic distances the homology concept is often more valuable than analogy in scientific reasoning, whereas relations of analogy become highly interesting in cases of large phylogenetic distances.

In philosophy of mind and cognitive science, the homology concept has received relatively little attention. Those who have discussed homology typically contrast it with *functionalism*, according to which psychological kinds are functional (causal) roles instead (e.g., Matthen, 1998; Griffiths, 1997). In light of the homology concept's employment in philosophy of mind as well as cognitive science, it is important to distinguish between two different projects concerned with giving an account of what it is to be a psychological kind with respect to functionalism. In philosophy of mind and language there is an *analytic metaphysics* project that aims to theorize the (ordinary) concept of psychological kinds using intuitions and conceptual analysis. In contrast, there is a philosophy of cognitive science project that aims to theorize the notion of a psychological kind in a way that supports good scientific research in psychology. The difference is in what factors determine how psychological kinds are to be construed: whether this construal describes our ordinary way of thinking of

psychological kinds, or whether it is methodologically advantageous to construe psychological kinds as one kind over another. These two may come apart; our ordinary concept of a psychological state might not correspond to the best way to construe psychological kinds for the purpose of fruitful psychological research. My concern is not with the analytic metaphysics project of accounting for our ordinary concepts of mental states, but rather with the philosophy of cognitive science project of giving an account of psychological kinds that supports good psychological research.

For the analytic metaphysics project in philosophy of mind and language, functionalism is the is the widely accepted thesis that mental state concepts (such as being in pain, belief, desire, intending and so on) correspond to functional roles. For instance, David Lewis (1980) famously argued that pain should be identified with its functional role by considerations of what it would mean for a Martian to be in pain. We can imagine a Martian with a very different biology than ours being in a state of pain, and reflection on this suggests that it is in virtue of the Martian's mental state having the same functional role as pain, such as being caused to wince, to thereby desire that the pain stop, and various other relationships to actions and mental states. An account of desiring that *P* might be similarly viewed to consist in (inter alia) a disposition to intend to bring about *P* will obtain given certain conditions; in conjunction with a belief that for *P* to obtain, one must perform a particular action *Q*, one might thereby be disposed to infer that one ought to *Q* and to subsequently intend to *Q*. In short, functionalism about mental states specifies them according to characteristic inferences that these tend to support, their relations to other mental states, and dispositions for certain sorts of actions, and this is a claim about the (ordinary) concept of a mental state. In this context, Mohan Matthen (2000) argued that mental kinds may be instead best viewed as evolutionary characters (homologues), citing how this accords with other ordinary intuitions, such as the intelligibility of a person being in pain without this state satisfying the presumed functional roles characteristic of pain (a dysfunctional state of being in pain, e.g., a person in pain without being disposed to avoid the painful stimuli).

For the philosophy of cognitive science project central to this thesis, functionalism is the thesis that psychological kinds (such as pain, belief, episodic

memory, intending and so on) are functional roles. Because the aim of this project is not to identify ordinary concepts but to rather to support good scientific theorizing, the considerations that support functionalism over alternatives are methodologically driven and relate to the fecundity and explanatory adequacy of viewing psychological kinds as functional roles. Influential here has been David Marr's ([1982], 2010) distinction between three levels of description: computational, algorithmic and implementational, which go (roughly) as follows. At the top is the *computational (task) level*, which describes the *task* to be solved by the system—the causal/functional role of the cognitive system. Next is the *algorithmic* or *cognitive level*, which specifies a *functional architecture* (an organized system of functional components) that together suffice for accomplishing the task specified at the computational level. Functional architectures are realized in physical systems at the third, *implementational* level, which describes physical or biological systems that behave according to the functional architecture at the cognitive level.

The relation between the algorithmic and implementational levels in human cognition is captured by Block's (1995) idea that the mind is the "software" that runs on the brain. Like the computers that are common today, the software is organized quite independently of the hardware that it runs on, and this justifies treating the hardware and software separately. Moreover, it is thought that multiple realization of psychological states further supports the independence of the cognitive and implementational levels, as the same cognitive system is expected to be implemented in heterogeneous neural systems, and this presents a further reason for viewing psychological kinds as functional roles (Polger, 2012). In summary, psychological kinds are functional roles, and the loosely coupled cognitive and implementational levels requires psychological kinds to be theorized as functional systems rather than as biological (e.g., neural) systems.

Homology of "cognitive function"

The idea of "cognitive homology" may seem like a contradiction in terms at first, as homology is sameness independent of any particular functional role (the mammalian forelimb in Figure 1.1 varies in its function, from supporting flight to swimming). Consequently, entities at the cognitive (functional

architectural) level cannot participate in relations of homology. However, a straightforward and consistent account of cognitive homology can be modeled on the resolution of the analogous issue of "functional homology" or less misleadingly "homology of function" in biology. Alan Love (2007), building off of Arno Wouters (2003), has clarified how to make sense of homologies of biological function in a way consistent with biological practice. Wouters distinguished between four kinds of function in biological practice: dubbed activity-function, biological role-function, biological advantage-function and selected effect-function:

1. function as *activity* (function$_1$)—what an organism, part, organ, or substance by itself does or is capable of doing;
2. function as *biological role* (function$_2$)—the way in which an item or activity contributes to a complex activity or capacity of an organism;
3. function as *biological advantage* (function$_3$)—the advantages to an organism of a certain item or behavior being present or having a certain character;
4. function as *selected effect* (function$_4$)—the effects for which a certain trait was selected in the past which explain its current presence in the population. (Wouters, 2003, 635)

Love argues that only the first of the four—*activity-function*—can appropriately participate in relations of homology. This is because the other three senses of 'function' are *use-functions* (what it is *for*, rather than what it *does)*. they are what they are used for, rather than how they are used. they are all specified *relationally* by their causal or functional contribution to a larger system of organization, whereas activity-functions are independent of use-function. Since one and the same homologue may vary in its use-functions, it follows that selected effect, biological advantage, and biological role functions are not the right kinds of things to be homologues (indeed, correspondence in selected effect functions are a case of *analogy* rather than homology). In biology, talk of "functional homology" is more precisely homology of function, where "function" here means *activity*. This is the sense in which the notion of "cognitive homology" must be understood: cognitive homology is homology of cognitive function, where "function" here means cognitive *activity*.

To illustrate, consider the heart. The *use-function* of the heart is to pump blood in the circulatory system. Its contribution to the larger system of organization is that it provides the pressure differences in blood that are

necessary for appropriate blood flow. This can be contrasted with the heart's *activity-function* i.e., what the heart, by itself, does or is capable of doing, and this is that the heart beats. *Beating* is constituted by a spatiotemporal organization of further activity-functions which are all internal to the heart (the cardiac cycle, outlined in Figure 1.3): the relaxation of the atria and ventricles, followed by the firing of the sinoatrial node in the heart and subsequent contraction of the atria, the contraction of ventricles, and so forth throughout the cardiac cycle of the heart. It is important to stress that activity-functions are constituted by further activity-functions, rather than needing to have causal roles as parts, much as morphological structures have further homologues as parts.[2] Certainly, for any structure of activities such as the those in cardiac cycle, each activity has a role in that system, but it is the activities themselves and not their causal roles that are the parts of the heartbeat activity.

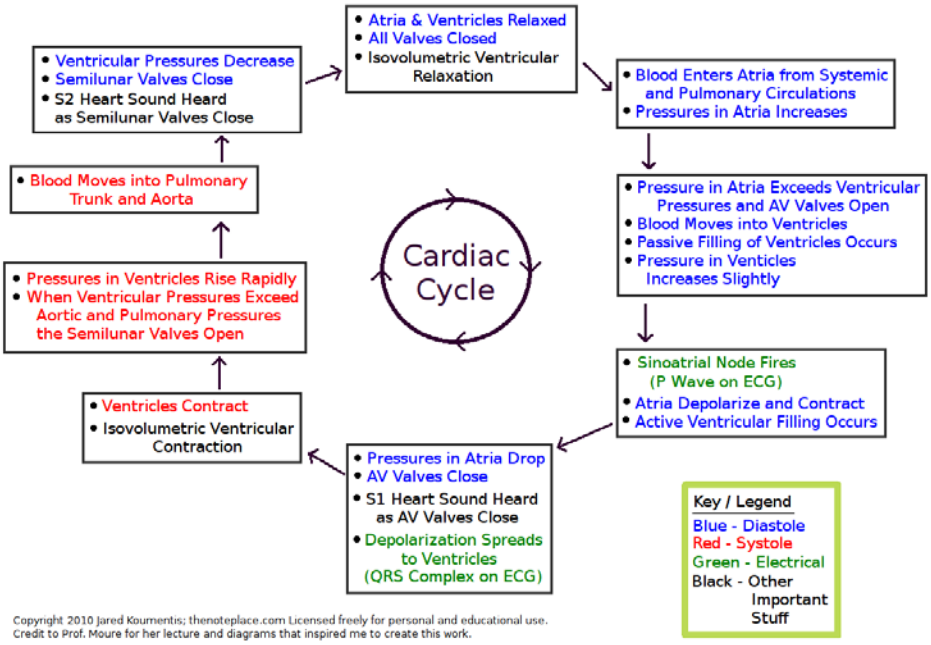

Figure 1.3
Architecture describing the heart's activity-function: the cardiac cycle.

[2] That is to say, activity-functions have activity-functions as parts, rather than being composed of functional roles as parts.

To provide an example from cognitive neuroscience, consider Broca's area. The activity-function corresponding to Broca's area refers to what this region by itself does or is capable of doing, and it is the internal (neural) activity found at this region independent of its functional role in any system. In terms of cytoarchitecture (which bears on the spatiotemporal structure of the activity-function), its relatively large pyramidal cells in layers III and V as well as its differentiated laminar pattern suggest that the structure of the activity is organized in terms of a nested hierarchy (Amunts et al., 2007; Keller et al., 2009). Indeed, Broca's area activity has a functional/causal role in a number of different tasks, including language comprehension (specifically syntactic processing), script cognition, and thinking about musical structures, which has led to the proposal that it tends to be deployed in representing abstract hierarchical structures regardless of modality (Fadiga et al., 2009; Farag et al., 2010). Such use-functional considerations suggest that its internal activity-function is organized so as to 'encode hierarchical dependencies'—and this is a description of the activity-function's architecture. Due to the above considerations, Broca's area activity has been suggested as a salient candidate for being a cognitive homologue (Bergeron, 2010). The value of homology thinking here is that it accounts for the potentially confusing notion of a function-independent function by distinguishing functional roles from activity-functions and identifying the latter with evolutionary characters.

Causal depth and mechanistic explanation

Having provided an indication of how to construe cognitive homology (which receives greater treatment in the next chapter), it is opportune to revisit the relationship between homology and explanation, in particular with respect to causal similarities and with mechanistic explanation. These two topics are of importance for later chapters, so in this section I discuss the argument from causal depth in more detail and contrast mechanisms with homologues.

Recall the argument from causal depth sketched in the introduction, which was modeled after Griffiths' argument for emotions as homologues. The argument from causal depth aims to establish that emotions are homologues rather than functional kinds due to how homologues stand to have a superior role in the development of scientific theory of the emotions, as follows:

Argument from causal depth

1. Homologues tend to be highly similar in underlying causal commonalities, and so tend to support robust causal generalizations.
2. Functional kinds (analogues) do not tend to be highly similar in underlying commonalities, and so do not tend to support robust causal generalizations.
3. A scientific theory of the emotions is best served by studying kinds that support robust causal generalizations.
4. So, the emotions should be viewed as homologues rather than functional kinds.

Premise 1 says that homologues have a tendency to share causal commonalities in such a way so as to support robust causal generalizations, so some account of this tendency and how it supports these causal generalizations is in order. The claim that homologues tend to share causal similarities is not one concerning any particular homologous characters sharing any particular causal similarities, but that homologous characters have a general tendency to share deep causal similarities. By appealing to the natural kind status of homologues as *homeostatic property clusters*, one can ground the claim that homologues tend to share deep causal commonalities.

The notion of a homeostatic property cluster (HPC) was introduced by Richard Boyd (1991) who argued that some natural kinds are HPCs that maintain homeostatic integrity in virtue of underlying causal processes. For HPCs, the different members of this natural kind tend to share various properties, where these properties need not be shared by every member of that kind. This clustering is not accidental, but due to some objective features of nature that is the causal basis of this clustering (the "homeostatic mechanism") The HPC conception of homologues as natural kinds is attractive due to its ability to accommodate how homologues retain their integrity in supporting scientific theorizing while not committing to any necessarily shared causal property of homologous characters (Assis and Brigandt, 2009; Rieppel and Kearney, 2007; Wilson et al., 2007; Wagner, 1996). Homeostatic mechanisms that support the shared causal properties of homologues come from the causal-developmental structure of the developing individual (e.g., developmental constraints).

Combined with the consideration that homologues are historical kinds, the HPC account of homologues accounts for and grounds this tendency. Homologues are historical kinds because they share common origins, either in phylogeny or due to the redeployment of developmental mechanisms in ontogeny. In either case, it is because of these common origins that homologous structures are initially highly similar (if not numerically identical) and absent further modification, they remain so. As homeostatic property clusters, homologues have homeostatic mechanisms that retain these property clusters. Importantly, as homologues are composed of further homologues as their parts, there is a correspondence in homeostatic mechanisms throughout homologues' parts as well. Homologues tend to behave similarly in causal interactions because for any given causal property, they will share this property unless there has been subsequent change in the homeostatic mechanisms. The extent to which various structures are likely to be conserved depend on the particular causal-developmental structures of the homologues and its parts—phenomena such as generative entrenchment are helpful for this (Wimsatt, 1986), which refers to how traits that appear earlier in development on which a larger number of subsequently developing entities depend on are harder to change without making the whole system nonfunctional.

Premise 2 of the argument from causal depth says that functional kinds such as analogues are not disposed to be highly convergent in causal commonalities (cf. Brigandt, 2009; Ereshefsky, 2012). It is well known that functional kinds such as analogues are extremely diverse in biology, as they should be given that they are defined in terms of relations to other functional kinds—relations that hold independently of whether objects' internal structures differ. For example bat and insect wings differ substantially in how they achieve flight, both in terms of their morphological structures as well as in the dynamics and mechanisms of flight. Analogues are grouped by functional considerations, such as wings being analogous due to their role in flight, rather than through any historical factors, and the multiple realizability of functional kinds implies that many different physical systems may realize the functional kind, so similarity need extend only to superficial features that are more or less necessary to realize the functional role. For instance, a camera eye requires a lens of some sort,

receptors that are sensitive to light, a focusing mechanism, and so on.[3] As functional kinds are not historical and are multiply realizable, there is nothing about them that supports a disposition for sharing deep causal commonalities.

The most prominent account of explanation in cognitive science is that it consists in mechanistic explanation (Craver, 2007). Given that the present account of cognitive homology is concerned with methodological considerations about the nature and role of psychological kinds in cognitive science, one might wonder about their relationship. At present I only want to draw out one brief point of difference between homologues and mechanisms in cognitive science: mechanisms are a broader class of things than homologues are, and (consequently) mechanisms, it seems, are not necessarily homologues. This is because there can be mechanisms for things that seem to not be evolutionary characters, such as sets of evolutionary characters (that do not compose another evolutionary character).

Explanation in terms of mechanisms consists in elaborating a mechanism that exhibits the "explanandum phenomenon" (Craver, 2007, 7). Mechanisms consist in activities and entities arranged so that they exhibit the behavior or properties to be explained. What makes something a part of a mechanism is that it is causally relevant to the mechanism exhibiting the explanandum phenomenon (as appropriately characterized, including by-products and other causal features; Craver, 2007, 153).

There is, in general, a mechanism for *T*, where *T* is any arbitrary task or cognitive function that is exhibited by a system. The mechanistic explanation of a task *T*, say, *T* = 'top-down face processing of fear-inducing stimuli under low-light conditions', consists in entities and activities arranged so that the causal role of the organization of parts in the mechanism is identical to the appropriately characterized causal role *T*. Sub-mechanisms, such as (say) the face processing mechanism, are constituted by whatever actual organization of activities and entities have the same causal role as the 'face processing' component of the

[3] Ereshefsky (2012) puts the relationship between analogues and homologue in the following way: an analogy class (group of analogous characters) consists of multiple homology classes (groups of homologous characters); the class of analogues 'wing' contains multiple classes of homologues (bird wing, bat wing, insect wing), and consequently the variation among the analogues is greater than the variation within any of the individual analogues.

mechanism for *T*. However, there is no *requirement* that this organization of parts comprising the mechanism for *T* (or any sub-mechanisms) is itself an activity-function homologue; an evolutionary character of its own, beyond whatever evolutionary characters are its parts. Satisfying a particular causal role does not a homologue make; the conditions for there being a mechanism for something requires only a causal role (for which there can be given a mechanistic explanation), and causal roles do not map one to one to activity-function homologues.

To be sure, this is compatible with there being mechanisms that *correspond* to homologues, such as the mechanism for the cardiac cycle. And mechanisms *can* (and often) have homologues as parts; the entities and activities in a mechanism can obviously be homologues (e.g., a neuron, an action potential, or a heart and a heartbeat). It is just that explanation in terms of mechanisms extend to phenomena that do not correspond to individual homologues, making any straightforward account of mechanisms in terms of homology difficult. I return to mechanistic explanation and an account of how mechanisms and homologues may be brought into closer alignment in the next chapter when considering whether "soft assembled" systems (like *T*, in the previous example) can be viewed as homologues.

Cognitive networks are a challenge for cognitive homology

Recent advances in cognitive neuroscience make an investigation into the use of homology particularly timely. In imaging neuroscience there has been a shift from brain activation to brain connectivity. Until recently, the vast majority of imaging neuroscience was focused on brain *activation*—isolating spatially defined zones that contribute to cognitive functions through subtraction from a baseline in an fMRI study—such as the so-called fusiform face area which processes face-configuration information. Now *connectivity*—the coordinated interaction of cognitive activities—is of much interest, and is fueled by a number of advancements in methodology and knowledge in imaging neuroscience, including large databases of imaging studies (Fox and Friston, forthcoming). Figure 1.4 shows this trend in general and for the default network in particular.

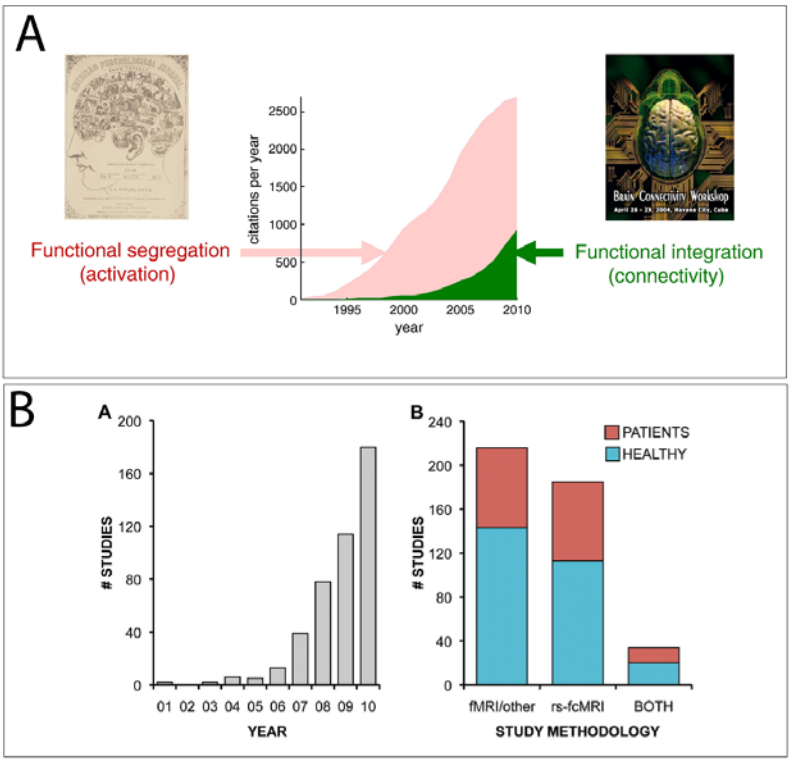

Figure 1.4

Top: Studies published on activation versus connectivity (Fox and Friston, forthcoming). Bottom: Studied published on the default network, on the left is total studies, on the right is broken down by participant type and methodology (fMRI, resting state functional connectivity MRI, or both) (Andrews-Hanna, 2012).

This shift to networks and related advances are important for the use of the homology concept in two ways. First, the identification of robust 'intrinsic' networks allows for analysis of cognitive activities as homologues on a network-level scale. Second, there is interest in identifying task-independent functional descriptions of cognitive activities, which are related to their redeployment in similar cognitive tasks. The task-independent function of networks and other cognitive activities describes their internal configuration that accounts for how and when it can be redeployed under a variety of functions, and homology thinking contributes to understanding what task-independent functions amount to. These developments in cognitive neuroscience have been used to call into question the assumed independence between cognitive architectures and their realization in systems of the brain (*viz.*, functionalism in cognitive science).

> Furthermore, the relevant question is not whether it is possible to study cognition "without making any statements about the neural structures involved" but whether this is the best strategy for making progress. … Given the tremendous recent advances in cognitive neuroscience, we are not convinced that it still makes sense, in understanding cognition, to talk about a "functional architecture" that is independent of the actual neural architecture. This may have been a convenient approximation when our knowledge of brain structure and function was still quite limited, but it is now too coarse to be useful. (Plaut and Patterson, 2010, 14; see also Coltheart, 2010; Patterson and Plaut, 2009)

> [We argue that] structure–function mappings can be defined and will lead to new cognitive ontologies that are grounded on the functional architectures that support them. However, to access the mappings and ontologies may require us to disassemble current views of cognition and use a more physiologically and anatomically informed approach. ….. In other words, we can apply current techniques to look not just for connections between brain regions but for connections between regions and cognitive processes in (abstract) [task-independent] cognitive spaces. (Fox and Friston, forthcoming)

Serial homology of cognitive activity-functions presents a promising avenue when considering how to conceptualize the bridge between these task-independent functions in abstract cognitive spaces with a physiologically and anatomically informed approach.

Chapter 2 – Cognitive homology

A cognitive homologue is the same cognitive activity under every variety of form and function, where these activities consist in spatiotemporal organizations of further activities and share a common origin of some sort. This chapter develops the notion of cognitive homology in detail by considering operational criteria (kinds of evidence for establishing cognitive homologies) and linking it with other theories in cognitive science. These theories are that of grounded cognition (Barsalou, 2008) and neural reuse (Anderson, 2010, 2007a).

I begin in Section 1 by reviewing some recent literature on homology in cognition and highlight how the subject matter of this thesis can be distinguished from this very recent work, which is by focusing on *serial* homology (which relates different parts of an individual) and cognitive *activities* from cognitive neuroscience. Section 2 revisits Remane's operational criteria of homology, which is formulated for phylogenetic approaches to homology of morphological structures, and raises analogous criteria that are formulated for serially homologous cognitive activity in particular. For each operational criterion, I provide illustrations to show how they are relevant to establishing cognitive homologies. Section 3 focuses on the hierarchy and connectivity in the context of cognitive networks so as to consider cognitive homology in light of the shift in imaging neuroscience to networks. Here I introduce several concepts from cognitive neuroscience, in particular intrinsic and resting state networks and ontology (*hubs* and *small-worlds*) which are inspired by graph theoretic approaches to cognitive networks (hubs and small-worlds). Section 4 ends the chapter by returning to a general discussion of cognitive homology and mechanistic explanation in cognitive science, where I suggest how they might be integrated by way of viewing "soft assembled" systems as being part of their own homeostatic mechanisms.

Recent work on cognitive homology

There has been a recent flurry of interest in applications of the homology concept to cognitive science and psychology, both in philosophy as well as in cognitive science venues. In the following I *briefly* survey this literature. I cannot do justice to all of these papers or discuss them in much detail in any reasonable

amount of space, but the aim here is to give the reader a general picture of the kind of interests people have in homology and cognition as well as to situate this thesis within the context of recent work.

A handful of articles discuss applications of homology and serial homology in psychology in general (Ereshefsky, 2007; Clark, 2009, 2010, forthcoming; D.S. Moore, 2012c). However, these papers tend to restrict their discussion and evidence of homology to overt behaviors and do not refer to cognitive activities *per se*.[4] In Clark's (2010, 2009) discussion of the possible relationship of serial homology between basic emotions and higher cognitive emotions, the focus is on neuroanatomy, omitting a discussion of cognitive activity during its appeal to non-behavioural empirical evidence.

A number of discussions explore conceptual issues that are raised from applying the homology concept in cognitive science. Ploeger and Galis (2011) explore the topics of modularity, developmental constraints and evolvability, arguing that they may be fruitful conceptual tools for cognitive neuroscience research. Balari and Lorenzo (2008, 2009) apply Pere Alberch's notion of morphospaces to cognitive functions in order to describe the concept of a "computational phenotype." However, by defining computational functions (functional roles) as phenotypes these authors do not sufficiently distinguish between homologues and the function of homologues. Garcia (2010) provides an account of "functional homology" in evolutionary cognitive science, and evaluates Remane's operational criteria in order to argue that these criteria are germane to cognitive science. However, Garcia seems to argue that there is no principled distinction between activity-functions and causal-role functions, since every activity-function within a system also has a causal-role function within a system. This would not be a valid inference; an activity-function may vary in its causal-role in different systems, and this serves to distinguish them.[5]

Some scientific papers have emerged focusing on the evolution of human language and cognition in evolutionary developmental terms (both traits are seen to be intimately related). Lanyon (2010) for instance stresses the emergence

[4] In a related vein, Katz (2011) discusses evolvability of behaviors by focusing on correspondence in underlying neural circuitry, highlighting the fact that parallel behavioural evolution may be very common due to the conserved structures in the brain.
[5] This response was suggested to me by Alan Love (personal communication).

of the granular layer in the prefrontal cortex, arguing that this novelty supports a saltationist view of human evolution, although again the focus is on morphology rather than cognitive activity *per se*. In a related vein, Scharff and Petri (2011) evaluate the role of the FoxP2 transcription factor and regulatory molecular network in the evolution of speech and language—i.e., *deep homology*, which is homology on a genetic level among quite unrelated species (Shubin et al., 2009). In a special issue on "cognitive homology" (Platt and Spelke, 2009), a number of papers focus on comparative and developmental evidence for explaining the evolution of various cognitive capacities, such as number cognition, mindreading, and higher order human cognition more generally. Platt and Spelke (2009) define cognitive homologues as "those psychological and neurobiological traits that evolved in the common ancestor of related phyletic groups that emerge from shared developmental pathways and serve closely related functions." This definition of homology reflects a phylogenetic approach, which it seems tends to be in play whenever cognitive activities are the subject matter. Bergeron uses Broca's area as an example of cognitive homology among different tasks, but still defines the sameness of homology "by the common phylogenetic origin of the associated structures" (2010).

In summary of the above, what is often lacking is a discussion of serial homology in general, and in particular serial homology of cognitive activities. Those that do raise the notion of serial homology leave out cognitive neuroscience and cognitive activity. These approaches are weakened by not engaging with the substantial literature on cognitive activity from cognitive neuroscience. When they *do* engage with the brain sciences explicitly, they tend to discuss homologies in brain anatomy, and activity *per se* goes unmentioned. And for those that do engage with cognitive neuroscience and cognitive activity, a phylogenetic approach to homology is in play. As previously discussed, it is at the intersection of serial homology and cognitive activity that it is particularly unclear how to understand "cognitive homology" and this is the central subject matter for this thesis.

Finally, just prior to completing this thesis, most of a special issue of *Developmental Psychobiology* has been published online on cognitive homology as the result of a large workshop in summer 2011, but these have not been included

in this section's discussion (D.S. Moore, 2012, C. Moore 2012; Michel, 2012; Lickliter and Bahrick, 2012; Hall, 2012; Blumberg, 2012; Anderson and Penner-Wilger, 2012; Clark, 2012).

Operational criteria for cognitive homology

This section discusses Remane's traditional operational criteria for establishing homologies, as they may be modified for use in identification of serial homology of cognitive activity.[6] Recall that the operational criteria are (1) *continuity*: the presence of an evolutionary continuum of properties from a more primitive character state to a more complex and derived state, (2) *special quality*: the complexity, distinctiveness or specialization of that trait, and (3) *position*: the relative spatial position within a larger system of morphological organization. Each will be discussed in turn. Importantly, satisfaction of any operational criteria is neither a sufficient nor a necessary condition for traits to be homologous, but rather provides *evidence* for homologies.

Continuity in Remane's formulation placed a focus on continuity throughout *evolutionary* lineages, and consists in the identification of a continuum of evolutionary intermediates from a simpler and more primitive state to a more complex and derived state of that trait. Continuity in the form of a morphological structure over time (such as the shape of a bone in a fossil record) supports the contention that the ancestral and derived traits are homologous. There is continuity in two senses: continuity within an evolutionary lineage over time, and continuity in form for a trait among extant species or individuals. Since serial homology involves the repetition or duplication of a structure within an organism, this indicates that the relevant time scale for continuity may not so much be phylogeny but ontogeny—that is, throughout the organisms' lifespan. This analogous criteria can be defined as a correspondence in the causal properties of cognitive activities as they appear within the organisms' lifespan. This is continuity in two senses: continuity in a cognitive activity's form throughout maturation (i.e., at different developmental time periods), and continuity in the varieties of form of tokened cognitive activities.[7]

---

[6] I present these in a different order from when they were introduced in Chapter 1 because it is easier to provide empirical illustrations in this order.

[7] I return to a discussion of cognitive homology in terms of types/tokens in the next section. It works the same way as with types and tokens of morphological structures in

When an activity undergoes duplication or repetition within one and the same organism, these activities are likely to be highly similar. Identifying a continuity from the earlier state of the activity to the mature state presents one kind of evidence for homologies between activities. Given two serially homologous activities, these activities are likely to share many properties that diverge as they take on mature character states specialized for their particular functional demands. Token activities of the same type of cognitive activity (i.e. serial homology of token activities in an organism) should also exhibit continuity, which will be manifest in a correspondence in causal features for these tokened activities throughout development.

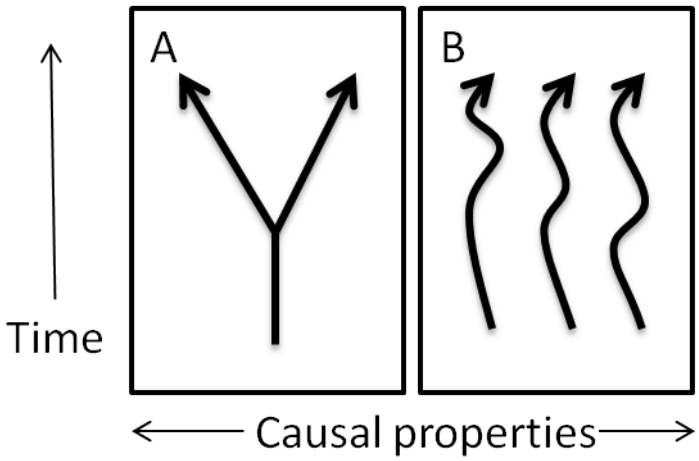

Figure 2.1
Two salient scenarios for locating continuity. (A) the duplication of an already existing cognitive activity, which diverge in causal properties; (B) correspondence in causal properties throughout the maturation of the cognitive activities.

Often cases of continuity will exemplify one of the following patterns: (a) a cognitive activity A develops and is duplicated such that there are two highly similar activities B and C, which subsequently diverge from each other in certain ways over the course of the organisms' lifespan, or (b) there is continuity in the variation of causal properties for tokened cognitive activities. These are illustrated in Figure 2.1 (A and B). Figure 2.1 (A) shows divergence after duplication. The time dimension in (A) refers to an organism's lifespan, so that

biology. The spinal vertebra is a homologue and the cervical vertebra type is serially homologous to lumbar vertebra type, and this relation of homology *also* holds for token vertebrae in individual organisms (such as my spinal vertebrae).

there is first one character state and subsequently two character states that come to differ throughout maturation. Figure 2.1 (B) shows continuity in variation for tokened cognitive activities, and here the time dimension refers to the duration of tokened activity-functions at some particular developmental time period.

Recall also that homologues may come to differ in their developmental mechanisms/pathways. We saw in Chapter 1 that this was true of morphological structures such as for the neural tube, and the structure may be homologous while the ontogeny is strikingly dissimilar. In the context of serial homology of activities, this means that similarity is not guaranteed at the outset. Duplicated activities may be initially highly similar, but over the course of evolution may vary in the dynamics of their maturation, including their initial developmental mechanisms.

The development of motor imagery illustrates continuity. Unlike other imagery modalities, motor imagery has clear behavioral components that can be easily studied in children. In a series of studies, Caeyenberghs et al. (2009a, 2009b) compared motor imagery and motor execution in children and adults. One notable property of motor imagery is that it follows Fitt's Law, which states that the speed of moving an object between two points varies characteristically as a function of the target size and distance (a speed-accuracy tradeoff). If the task is to move a block back and forth between two regions on a desk, variation in distance or target region sizes should similarly affect both the estimated time to complete the task and the actual time it takes to complete the movement task. For children under approximately age 7, only motor execution tasks follow Fitt's Law, which indicates that children are not using motor imagery before this age.

At 7-8 years old, motor imagery task performance begins to conform to Fitt's Law and also start displaying other typical dynamics of motor imagery, such as a correlation between actual and estimated action speeds. To be sure, such changes are associated with concurrent differentiation among key posterior brain regions that typically underlie motor imagery and execution performance in adults. Crucially in this illustrative setting, it is at this time that children start displaying overt, only partially-suppressed task related movements during motor imagery tasks, such as body rocking and limb and head movements. Over time, however, these overt movements are gradually suppressed, and motor

imagery is no longer present with any significant overt behavior. This continuity indicates a "developmental trajectory that is entwined with the development of movement skill in children" (Caeyenberghs et al., 2009a). Interestingly, there is further convergence of motor imagery and execution *performance* during maturation through adulthood (e.g., less difference in time for imagining and executing a given action).

The development of motor imagery illustrates how continuity provides evidence for serially homologous cognitive activities. Some suitably defined activity related to overt motor execution is redeployed, and *overt* action is gradually suppressed until in adults where overt action is almost entirely suppressed in imagery. Initially there is high correspondence in overt behavior for motor imagery and execution, but overt behavior is not present in adult motor imagery (as in Figure 2.1a). The significance of partially (and increasingly) suppressed overt movements is that it suggests that the repeated or duplicated activity used in imagery tasks is one that generates overt movement, and so the mature state of motor imagery is a serial homologue for of this motor execution process. The significance of the tighter coupling of performance in motor imagery and execution is that it displays correlations in causal properties of token activities later in development (as in Figure 2.1b). These in sum illustrate how 'continuity' can be used to identify serially homologous cognitive activities.

Consider now Remane's criterion of *special quality*, which refers to the "complexity, distinctiveness or specialization" of a character – that is, a correspondence in any especially distinctive property of the character. Understood as the "complexity, distinctiveness or specialization" of a character, this criterion's formulation will not need to be modified to accommodate serial homology or activity-functions. It is also a category of evidence that one expects to be highly heterogeneous since it may extend to nearly any property of an activity.[8]

The criterion of special quality is motivated by the consideration that the more distinctive a trait is, the less likely it will have evolved more than once and so shared special features are unlikely to be due to homoplasy (or convergent evolution). Special quality is additionally bolstered when there is no adaptive

[8] A list of all particularly relevant special qualities to look for in cognitive systems (given current technology) would be very interesting, but I do not try to present such a list here.

reason for this special feature being present. For example, a distinctive cellular makeup of an organ could be a special quality that is evidence of homology. In the case of serial homology of cognitive activities, special quality is motivated by the consideration that the more distinctive an activity is, the less likely it will have evolved/developed separately (rather than being a duplication using the same developmental pathway), and so shared special features are unlikely to be due to convergent evolution of activities which are not serially homologous. In other words, the special quality thought to be present as a result of duplication of an activity, rather than the de novo origin of that activity. For example, a distinctive internal configuration of the activity-function (which is closely related to cytoarchitecture for local activities) is one kind of 'special quality' evidence of homology.

Correspondence in a distinctive internal configuration for an activity is evidence of serial homology. Neurophysiological measures such as event related potentials (ERP) from electroencephalography (EEG) provide another candidate for a special quality when they are distinctive. The following example of mismatch negativity illustrates both of these special quality considerations. Mismatch negativity (MMN) is a commonly observed event related potential (ERP, as measured by EEG recordings on the surface of the head), which is associated with a deviant or "oddball" stimulus, such as an auditory pattern of tones ABABAB**B** (Figure 2.1a). One interesting quality of MMN is that it is found for stimuli among many sensory modalities, including auditory, visual, tactile, and associative (crossmodal) systems. This is notable because EEG recordings can be used to approximate localizations of the source of an ERP in the brain, and MMN appears to be related to the activity localized in these different sensory systems.[9]

[9] Interestingly, MMN is often modulated in the same way across modalities in clinical settings, such as in schizophrenia or prematurely born infants. This would be another example of continuity of the second type in Figure 2.1b, where there is a correspondence in the causal properties of suspected homologues over time in development.

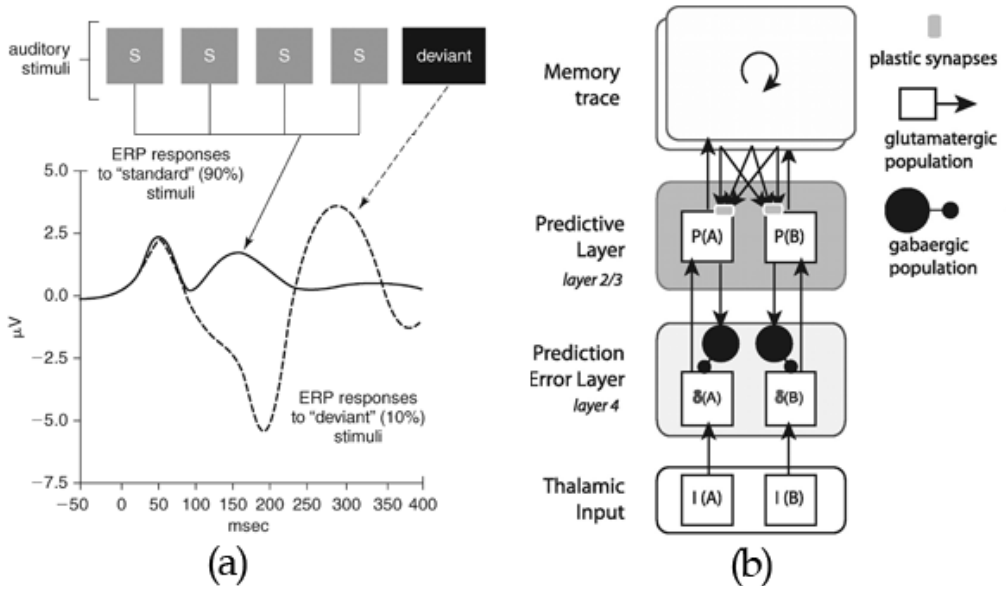

Figure 2.2

(a) Mismatch negativity ERP response. The dotted line corresponds to EEG recordings for a deviant stimuli. The negative ERP response at 200ms is the mismatch negativity. (Adaped from Light et al., 2010.) (b) Neuronal model of predictive coding for mismatch negativity. (Adapted from Wacongne et al., 2012.)

Inspired by these and other considerations, Wacongne et al. (2012) developed a neuronal model for predictive coding that provides a mechanistic explanation of MMN (Figure 2.1b). In this model, which is developed for auditory stimuli, each column represents a cortical column with thalamic input responding maximally to one of two frequencies of auditory input. The "predictive layer" population of neurons corresponds to layers 2/3 of the cortex and prediction of particular sounds are coded in population firing rates. These predictions are compared with the incoming inputs from the thalamus through a number of synaptic mechanisms in the "prediction error" population at layer 4 of the cortex, resulting in this population firing whenever thalamic input is not cancelled by predictive signals. This activity in layer 4 would result in the appearance of MMN (Wacongne et al., 2012).

Throughout the cerebral cortex, including in other sensory modality systems, there exists a closely similar neuronal architecture of cortical layers. Wacogne et al. (2012) argue these relevantly similar neuronal architectures in sensory and multimodal systems supports the idea that a similar architecture of

predictive coding may apply to these other sensory systems where mismatch negativity is also found.

Mismatch negativity and the associated neuronal architecture are both examples of special qualities. MMN is distinctive due to the relatively robust temporal evolution of this event related potential (a decline followed by a sharp ERP negativity at approximately 200ms). Its appearance across sensory modalities provides suggestive evidence that the cognitive activities it reflects are serially homologous. Assuming that other sensory modalities do indeed have relevantly similar neuronal architectures that underlie the appearance of MMN (i.e., assuming they have similar mechanisms), this also serves as 'special quality' evidence that the predictive coding activities among modalities are serially homologous.[10]

The third operational criterion is *relative position*. The criterion of relative position was originally formulated in order to identify homologies between morphological structures across individuals. As originally formulated, relative position is a structure's position relative to other bodily parts/homologues of the organism, and a correspondence in relative position across individuals is evidence that the associated structure is homologous. This owes to the consideration that it is unlikely for a character to be lost and subsequently replaced by another in the same relative position. Although activities are performed by morphological structures, they are independent from any particular morphological structure.[11] How can relative position extended to apply to activity-functions, given that they are not morphological structures? Since activity-functions are *spatiotemporal* organizations, this indicates that an analogous criterion of relative *spatiotemporal* position may apply (Love, 2007).

Spatiotemporal position or "organization" refers to how activities are arranged in order to perform functional roles (Love, 2007; Clark, 2009). More precisely, correspondence in spatiotemporal position refers to when an activity is

[10] To be sure, this is a *good candidate* for special quality. One would have to first exclude adaptive reasons for this similarity, and any firm conclusion on this matter depend on showing, for instance, that there are other possible mechanisms that may be better, so that the occurrence of this shared similarity is unlikely to be due to convergent evolution.
[11] At least conceptually. However, at present I know of no easy or non-controversial cases that clearly show the same cognitive activity in two distinct cytoarchitectures or distinct brain morphological areas (e.g., visual cortex and hippocampus). But see Chapter 3 for a further discussion of this and some more detailed examples.

in the same relative spatiotemporal position among other activities/homologues, independent of its current functional role in a system. Correspondence in organization independent of task, in other words, is the analogous criterion for spatial position. The criterion of spatiotemporal position is motivated by the consideration that it is unlikely for an activity to be lost and subsequently replaced by another in the same relative spatiotemporal position.

However, finding a useful account of relative spatiotemporal position proves difficult. Shared causal role is evidence of homology, but what makes serial homology so interesting is precisely when there is variation in functional role. The problem is that difference in functional role always shows up as differences in spatiotemporal positions as given by functional connectivity in imaging studies. As Anderson and Penner-Wilger observe:

> one of the oft-cited criteria for homologous structures is that they occupy the same position with the same connections in two different species … [but] the different uses of a given circuit are differentiated precisely by their different (functional) connections to other neural structures (Anderson and Penner-Wilger, 2012).

For instance, functional connectivity between the insula and the rest of the brain can differ extensively depending on what the task is – the opposite of a correspondence in spatiotemporal position (Jabbi et al., 2008; Friston, 2009). So either spatiotemporal position is not a very useful criterion (given that the interesting cases are those that vary in functional role), or current imaging techniques are not always able to yield the relevant correspondences.

My suggestion is that spatiotemporal position is relevant even in cases where there is variation in functional connectivity, and this is because imaging technology does not presently provide a high enough resolution of the structure of cognitive activities to locate correspondences. That is to say, more detailed models are needed than what can be arrived at solely by looking at functional connectivity in imaging data. In particular, it is by specifying the mechanisms that underlie connectivity in imaging studies that one can discern relevant spatiotemporal correspondences.

Here is one example of such a mechanism. Zanto et al. (2010, 2011) studied top-down (attentional) modulation of visual feature processing of motion and color stimuli. Using fMRI they found activity in region of the right

inferior frontal junction (IFJ) occurs in motion and color stimuli conditions. Interestingly, the activity for motion was dorsal to that of color in neighboring regions of IFJ (corresponding with motion and color being processed in dorsal and ventral streams respectively). Zanto et al further suggest that the mechanism for the influence on these areas of the right IFJ on visual processing centers is long-distance alpha band (8-12 Hz) phase coherence between IFJ and visual centers. So, if they are correct, the two activities in dorsal and ventral IFJ are bridged by the phase coherence mechanism (likely via thalamic input that is modulated by TPJ; Foxe and Snyder, 2011). This kind of fine-grained spatiotemporal structure that can provide evidence of homology using spatiotemporal position, since the two activities in IFJ occupy the same spatiotemporal position with respect to activities in the phase coherence mechanisms.

The upshot is that relative position needs to be construed as relative position within a set of *nearby* activities and not functional connectivity of any sort (Figure 2.3). In morphology, structural position is more precisely relative position with respect to *proximal* rather than distal morphological structures/homologues. A distal relative spatial position such as being located somewhere below the neck does not provide the relevant spatial relationship for establishing homologues, whereas proximal (adjacent) structures are relevant. Functional connectivity establishes some sort of spatiotemporal relationship, so that one activity predicts another at a later time, and this is analogous to a spatial position such as "being below the neck." Connectivity inferred from imaging data tends to provide *distal* spatiotemporal positional information, but this is not the relevant kind of position. Rather, what is relevant are the proximal (adjacent) activities that are not separated by a number of other intermediate activities. In addition to not providing the right kind of evidence, not distinguishing between distal and proximal functional connectivity subsumes functional role under spatiotemporal positional criteria, but it is precisely from this that we want to keep separate from a cognitive homologue.

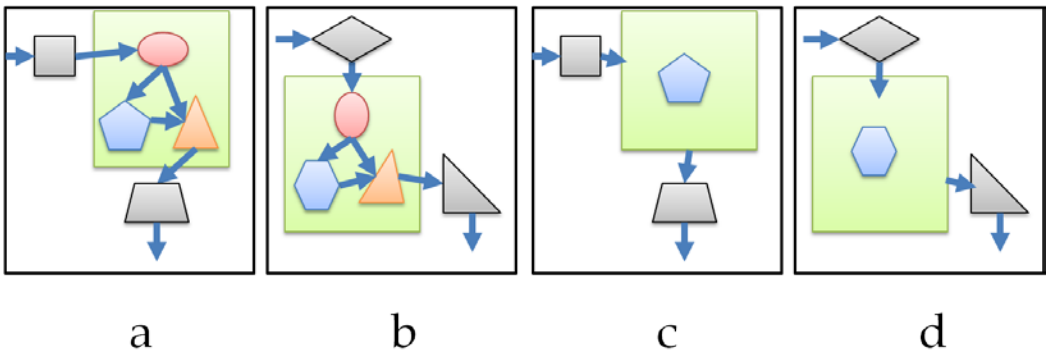

Figure 2.3

Spatiotemporal position, proximal and distal. **(a & b)** proximal correspondences in spatiotemporal position. The activities (polygons) inside the shaded box show a correspondence in spatiotemporal position. For example, the pentagon/hexagon has the same spatiotemporal position with respect to the circle and triangle. So even though it may appear different, the correspondence in spatiotemporal position is evidence for homology. **(c & d)** distal spatio-temporal position, representing what might be revealed by fMRI functional connectivity. The pentagon/hexagon shows different functional connectivity with the rest of the brain.[12]

In summary of this section on operational criteria, the following is the reformulated version of Remane's operational criteria that I began with:

Continuity: identifying a correspondence in causal properties of an activity as the activity develops from an immature state to a more mature state of that trait or the manifestations of the trait at some time.[13]

Special Quality: the complexity, distinctiveness, or specialization of that trait.

Position: the proximal spatiotemporal position of an activity within a more general pattern of organization.

Theories of grounded cognition and neural reuse

Having discussed operational criterion for (serial) cognitive homology, I now turn to a discussion of two theories that bear close relationships to cognitive homology: that of grounded cognition (Barsalou, 2008) and neural reuse (Anderson, 2010). Connecting homology with these theories allows for better

[12] A detailed example is given below, in the discussion of cognitive networks immediately following the next section.

[13] To be sure, the disjunction here is found in the original criteria but is buried within the distinction between "primitive" and "derived" states, which allows for continuity within a phylogenetic branch of one species and among extant species to be made in the same terms.

understanding the role cognitive homology can play in supporting scientific theorizing.

Grounded cognition is a theory according to which conceptual representations are "grounded" in modal systems for perception and action, as opposed to being grounded in a distinct, amodal conceptual system (reviewed in Barsalou, 2008). This theory aims to reject the view that thinking draws on conceptual representations independently from modal systems involved in perception and action, in favor of a view that embeds the former within the latter. In his review, Barsalou distinguishes between a number of grounded cognition theories. Cognitive linguistic theories argue that abstract concepts are grounded metaphorically in embodied and situated knowledge (Lakoff and Johnson, 1999), such as happiness and sadness being grounded in spatial systems (happiness is up, sadness is down). Social simulation theories argue that representations of others' minds involves simulation of one's own mental states (Goldman, 2006b). Cognitive simulation theories argue that conceptual processing (such as in abstract reasoning) is grounded in perceptual and motor systems (e.g., Glenberg et al., 2008).

Consider cognitive simulation theories of conceptual processing. Evidence for grounded cognition here include behavioral evidence such as property verification; for instance, the size of an object in perception predicts how long it takes to verify whether a property is true of that object, e.g., <horse, tail> or <squirrel, tail> (Solomon and Barsalou, 2004). Category-specific deficits in object recognition from lesions involve selective loss of the ability to name certain kinds of objects, such as tools or colors, and these are associated with lesions to the dominant modality for interacting with the modality (e.g., lesions to color processing visual areas for loss of color knowledge).

Neuroimaging evidence also shows that when conceptual knowledge is represented, this involves activation of areas that represent their properties during perception and action (Martin, 2007; Simmons et al., 2007). Activity in the so-called fusiform face area (FFA) consistently occurs when seeing or occurrently thinking about faces, and lesions to the FFA produce prosopagnosia (an inability identify faces), deficits in face imagery, inabilities to describe facial features in detail, and deficits in drawing faces (Martin, 2007). Conceptual processing of

actions and tools involves reuse of premotor processes (Pulvermüller, 2005). Similarly, observation, imagination and experience of disgust all converge on anterior insula activation, despite being part of different global patterns of activity (Jabbi et al., 2008). There is a lot of evidence along these lines.

How do grounded cognition theories of conceptual processing relate to cognitive homology? The central claim for these theories is that conceptual processing in thought reuses the same systems that conceptual processing in perception and action use. In terms of homology, a reasonable interpretation is that some core component activities central to conceptual processing are serially homologous to activities in perception and action execution. For instance, fusiform face area (FFA) activity is serially homologous in activities of perceiving and thinking about faces (though these two whole systems of perceiving and thinking about faces need not be homologous).

However, a construal in terms of serial homology is not necessarily straightforward. For serial homology requires numerically distinct entities (which can be serially homologous). A morphological structure may be used for different functions, but this does not yield two numerically distinct morphological structures. For activities, the same point holds: adopting a different functional role does not make the activity two numerically distinct kinds. There are two ways to respond to this potential issue by focusing on how the concept of homology applies to types and tokens. In biology, both types and tokens are homologized: cervical and lumbar vertebrae are serially homologous vertebral structures, but so are token spinal vertebrae within an individual organism.

One potential solution focuses on serial homology of types by arguing that these cases do indeed differ on the type level. In order for serial homology to exist at the type level, there have to be numerically distinct types (that are homologous). This can be understood in terms of character states: if a character has two types of character state these can serve to demarcate serially homologous characters at the type level. That is to say, the difference between the type and token situation is that in the former, the tokens can be grouped into distinct types (where each type includes many tokens). So the FFA-activity that occurs in thinking about faces and the FFA-activity that occurs in perception of faces

would be distinguished by their character state: the FFA-activity tokens in thinking about faces are one type, whereas the FFA-activity tokens in face perception are of another type. It is not altogether implausible that this is the case. For instance, consider how the vividness of mental imagery is predicted by the strength of the fMRI BOLD response in perceptual systems (Olivetti Belardinelli et al., 2009). One could use such differences in the character state of the activities that the differential BOLD response reflects to individuate serially homologous FFA activity-functions on the type level.

A second response focuses on token activities, claiming relations of homology obtain between tokens of the same type. Token structures are homologized in biology (e.g., token structures in different species, or within an individual organism), so perhaps relations of serial homology obtain between token cognitive activities. Anderson and Penner-Wilger suggest something along these lines:

> [There] need not be a copy of the neural structure that is adapted to new uses; instead the very same structure comes to participate in functional complexes. Thus, neither the physical structure nor the developmental pathway is duplicated, and while the function is in some sense duplicated, it is a temporal rather than a physical or spatial duplication. (Anderson and Penner-Wilger, 2012)

Under this construal, one focuses on token activities that are distinguished temporally (hence they are not numerically identical tokens). In other words, repeated instances of an activity can be serially homologous because they are distinct tokens of the same homology class. Given that both types and tokens are homologized for structures in biology, I see no reason to view these two positions as mutually exclusive. What this highlights is the need to be clear about whether types or tokens are being homologized: the claim that activity A and activity B are serial homologues may be ambiguous between whether only the tokens are serially homologous or whether there is serial homology at the type level as well.

Additionally, one must be careful with the level of generality with which one describes traits due to the hierarchical nature of homologues. It is not just that homologues contain further homologues as parts, but descriptions of homologues such as "red blood cell" are ambiguous in the homologue they pick out in the first instance. Recall that homologues are homeostatic property

clusters that can vary largely independently of each other. For "red blood cell," there are at least two relevant HPCs/homologues: an HPC common to all blood cells (the blood cell HPC), and a second HPC unique to red blood cells in particular (not found in white blood cells). A token white blood cell and token red blood cell are serially homologous *blood cells* (owing to duplicated blood cell HPC), but a token white blood cell and token red blood cell are *not* serially homologous tokens of *red* blood cells (white blood cells are not duplications of the red blood cell HPC). For blood cells in general (the blood cell HPC), the blood cell's type (red/white) are character *states* of the blood cell, meaning that red/white blood cells are serially homologous types of blood cell and tokens in either state are serially homologous blood cells.

These same considerations apply to cognitive activities as HPCs/homologues, *mutatis mutandis*. For "motor imagery," (suppose) there are at least two relevant HPCs/homologues: an HPC common to all imagery (the imagery HPC), and a second HPC unique to motor imagery in particular (not found in visual imagery). A token visual imagery activity and token motor imagery activity are serially homologous *imagery activities* (owing to the imagery HPC), but a token visual imagery activity and token motor imagery activity are *not* serially homologous tokens of *motor* imagery (the motor imagery HPC, not the same in visual imagery). For imagery activity in general (the imagery HPC), the imagery activity's type (motor/visual) are character *states* of the imagery activity, meaning that motor/visual imagery activities are serially homologous types of imagery activities and tokens in either state are serially homologous imagery activities.

Michael L. Anderson's theory of neural reuse (or "massive redeployment hypothesis") is a theory of the functional structure of the brain, according to which it is extremely common for neural circuits to be exapted (recycled, redeployed) during evolution and during development (Anderson, 2007a, 2010). Accordingly, higher cognitive activities (e.g., abstract reasoning, language understanding) have to find their neural niche in existing systems. For instance, it seems that egocentric (body-oriented) attention processes are redeployed in representing magnitude forming a number-line in egocentric space, with small magnitudes on the left, and large magnitudes on the right (Hubbard et al., 2005).

Finger representational processes (i.e., in finger gnosis and learning to count) are reused for mathematical calculation (Andres et al., 2008; Penner-Wilger and Anderson, 2008). This theory shares many similarities with Barsalou's theory of grounded cognition, particularly in terms of the evidence there is for neural reuse (so I will not repeat any of it here). There are two especially notable aspects of this theory. First, it has been interpreted in terms of homology (Bergeron, 2010; Moore and Moore, 2010; Anderson and Penner-Wilger, 2012). In particular, Anderson views his theory of neural reuse as a theory of the organization of token cognitive activities. The "redeployment" in neural reuse is not neural reuse where deployed circuits have different character states, but temporally distinct tokens of a numerically identical type.

Secondly, it is an evolutionary developmental theory of cognition that makes specific claims about the intersection between the evolution and development of cognitive activities. The core component of neural reuse is that evolutionary and developmental considerations lead to the expectation that the brain will reuse existing components for new tasks rather than developing new circuits *de novo*. The reason that massive redeployment of existing neural circuits is expected because it is costly to sustain the existence of new circuits when redeployment of existing systems is sufficient for accomplishing a task, so redundant structures are highly unlikely to be selected for. Increases in the number of circuits and systems suggests more potential for neural reuse, as functional reorganizations of more existing activities will likely be sufficient for accomplishing a wider variety of tasks.

At least three predictions can be made based on this core idea. First, brain regions are expected to actually participate in diverse task categories. Second, there should be a correlation between the phylogenetic age of a brain area (the cognitive activity, more precisely) and the frequency with which it is redeployed, as older areas have been present for redeployment longer and are more likely to have been integrated in later developing / evolving systems. Third, there should be a correlation between the phylogenetic age of a cognitive function, and the degree of localization (average spatial distance between) the cognitive activities for the given task. Several publications by Anderson test and support these three predictions (Anderson, 2007b, 2007a, 2008).

Aside from making these evolutionary developmental predictions, this theory suggests that there ought to be some mapping between cognitive activities and domain-independent functional roles. The objective of such a task-independent structure-function mapping is that it provides an architectural template which can be used to explain why certain tasks involve redeployment of certain cognitive activities and not others. As previously discussed, Broca's area has been suggested to be a domain-independent processor of hierarchical structures. Anderson and Penner-Wilger (2012) survey the way in which finger representation processes are reused among multiple task domains in order to determine "what the shared circuit is *doing* during all these various tasks":

> For reuse to have occurred, the service offered by the shared circuit must be something that the different uses could benefit from incorporating. Applying this perspective to the uses found in the database search, we identified some common requirements across uses, including ordered storage of discrete representations and mapping between representational forms. Although neural activations are generally assigned functional processes specific to the domain under investigation, cross-domain structure–function mapping requires a domain-independent vocabulary. Thus, using vocabulary drawn from computation, our proposal for the structure–function pairing that could meet the functional requirements imposed by the multiple uses is an array of pointers. An array is an ordered group, and a pointer is a data structure that designates a memory location and can indicate different data types. (Anderson and Penner-Wilger, 2012)

In sum, the suggestion is that the domain-independent function for this circuit is an 'array of pointers' (which indexes where information/things are stored rather than containing the information itself).

My suggestion is that homology provides an account of what these domain-independent functions are. For these domain-independent functions are not specified by their use in any particular task (though they can be gleaned from the pattern of reuse among multiple task domains). Instead, these functions refer to what the activity by itself does or is capable of doing, which is to say they are activity-functions. The computational description of an 'array of pointers' refers to the structure of the activity-function performed in the finger representation part of the somatosensory cortex. This reflects a very important point about activity-functions: the value of homology thinking here is that it accounts for the potentially confusing notion of a function-independent function by distinguishing functional roles from activity-functions.

Homology in cognitive networks

One of the more interesting developments in cognitive neuroscience relevant to cognitive homology is a shift to networks and connectivity. Central to this shift has been a combination of developments: meta-analytic modeling using large imaging databases, the use of graph theory to describe cognitive networks, and intrinsic and resting state networks (Fox and Friston, 2012; Bullmore and Sporns, 2009). The networks and concepts involved have not been explored in relation to homology, as to the best of my knowledge discussions of homology and cognition have remained focused on highly localized activities in particular brain regions.

It is not uncommon for any functionally connected set of brain activity to be called a network when it is involved in a task, such that there are an extremely large amount of "networks" that correspond to the diversity of tasks that can be performed by a cognitive system. Under this way of speaking about networks, they are similar to mechanisms in that they are more precisely *networks-for*, they are task-defined ensembles of activation (Seeley et al., 2007).

More restrictive accounts of networks have been proposed, though there is no clear consensus about the nature of these entities and their precise relationships with one another. These are resting state functional connectivity networks, meta analytic connectivity modeling networks, and intrinsic connectivity networks (Deco et al., 2011; Laird et al., 2011; Rubinov and Sporns, 2010). Resting state functional connectivity networks are functional networks detectable at "rest"—i.e., when there are no explicit task instructions. Meta-analytic functional networks are consistently coupled activation patterns among a variety of task domains. These two types of networks show strikingly similar structures and do so across imaging techniques, which has led to the proposal that these reflect the existence of stable organizations of cognitive activities or *intrinsic networks,* which some have argued are the "fundamental, organizational elements of human brain architecture" (Laird et al., 2011, 4022).

Graph theory has also been used as a mathematical framework for studying networks, providing ways of quantifying hierarchy and substructure of network graphs, traffic flow, and other properties of networks (Power et al., 2011). Networks are formalized as mathematical objects consisting of 'nodes' and

'ties' between these nodes, where nodes are anatomically defined regions and the strength of ties between nodes is a measure of causal influence or statistical dependence of activity between nodes (effective or functional connectivity).

The graph theoretic concepts of *small worlds* and scale-free network properties such as highly integrated *hubs* are particularly germane to theorizing the architectures of cognitive activities (Sporns et al., 2004; Rubinov and Sporns, 2010). A *small world* is a tightly coupled neighborhood of nodes, such that most ties are among neighboring nodes and only a few ties exist to distant nodes. Scale-free networks are networks such that the degree of connections for nodes are distributed according to a power law (over orders of magnitude) rather than having a typical number of connections (within an order of magnitude). One feature of such scale free networks are *highly integrated hubs*, which are nodes with very high connectivity—potentially orders of magnitude above average.

In the following I show how the cognitive homology concept applies to intrinsic networks and these graph theoretic structures by focusing on how they apply to the most well understood studied network: the default (mode) network. The motivation for connecting these concepts is because intrinsic networks can be viewed as homologues and these graph theoretic structures can play a useful role in understanding the ways that homologues can vary in character states.

Consider Figure 2.4, which shows correlated activity for three cognitive networks (left) and graph theoretic measure of connectivity (right). As can be seen by inspection, the default network, dorsal attention network, and task control network all engage frontal, parietal and temporal regions, and these regions are largely non-overlapping.

Figure 2.4b shows a graph theoretic representation of the functional connectivity between nodes (regions) of the three networks among others (see Power et al., 2011). The bottom arrows point to the default network (red), showing high functional connectivity between nodes and low connectivity to other network nodes (the other two arrows are to auditory and visual systems). This local integration and global isolation of networks illustrate *small worlds*, defined as groups with high inter-connectivity among neighboring nodes but low connectivity to distant nodes.

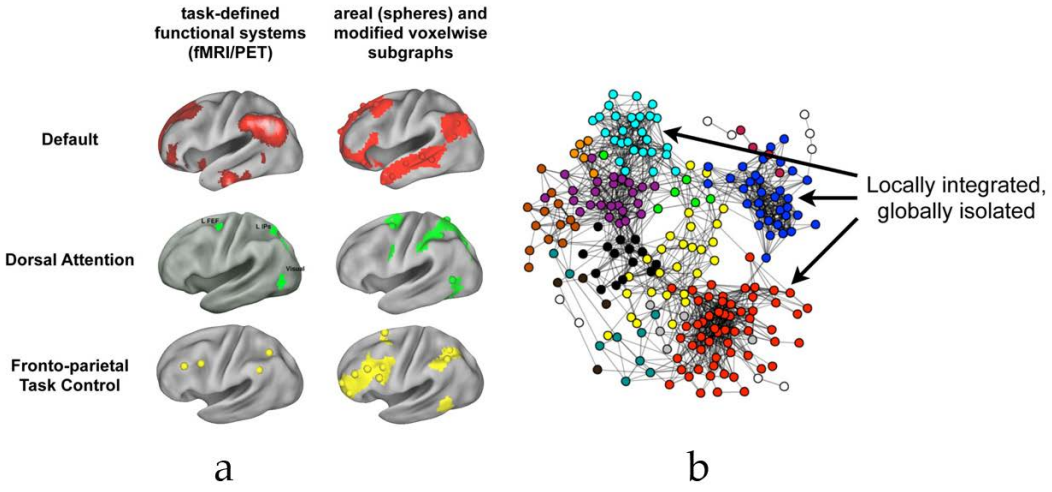

Figure 2.4

(a) three cognitive network localizations using two methodologies (task defined and resting state graphs with voxel nodes): the default mode network, dorsal attention network, and fronto-parietal task control network. (b) small world architectures illustrated. Note that distance between nodes corresponds to integration between nodes rather than spatial distance. Adapted from Power et al. (2011)

Highly connected *hubs* are nodes with high degrees of connectivity. The posterior cingulate cortex (PCC) and precuneus contains such a cortical hub and this "hub" is well established as a core part of the default network (Leech et al., 2012). Leech et al. (2012) presented imaging data that suggests that "the activity of functionally distinct distributed brain networks is echoed in spatially overlapping but distinct parts of the PCC" (Leech et al., 2012, 220). For instance, ventral regions were particularly implicated in default network processes, and two dorsal regions showed functional connectivity with the dorsal attention and task control networks (Figure 2.4a). Leech et al. (2012) argue on these grounds that the PCC acts as a cortical hub that integrates activities originating from multiple distinct brain networks. A graph theoretic representation of this connectivity would show a remarkably high number of ties to nodes in other brain areas, when the whole region is used as a node.

When subareas of the PCC treated taken as nodes (rather than as one large node), it is intuitively clear how PCC also likely forms a "small world" network in terms of its internal structure. As with early visual processing areas, nodes *within* the PCC form a tightly functionally coupled neighborhood, but connectivity to other systems is maintained by select regions, such that only

particular subregions of the PCC exhibit connections to distant nodes.[14] To the best of my knowledge, there have been no investigations of the nature of the connectivity among regions of this posterior hub of the default network. However, many studies have investigated the functional connectivity of particular regions of the PCC/precuneus hub (Margulies et al., 2009; Figure 2.4). What Figure 2.5 shows is how parts of the PCC/precuneus hub are selectively connected to other systems, so that within the hub, only particular subregions high connectivity to other networks.

[14] However, it is important to also note that from the perspective of graph theory, hubs and small worlds are not necessarily localized within a particular anatomical zones; they are measures of functional connectivity, independent of spatial location. Likewise, ties between nodes are not measures of physical distance between nodes but rather functional integration (Tomasi and Volkow, 2011). In other words, although local anatomical regions may often form small worlds and highly connected hubs, functional small worlds and functional hubs are not limited to such regions. This is all to say, there are two interpretations of "hubs" – anatomical hubs, which for developmental reasons are likely functional hubs, and functional hubs, which are not necessarily spatially localized.

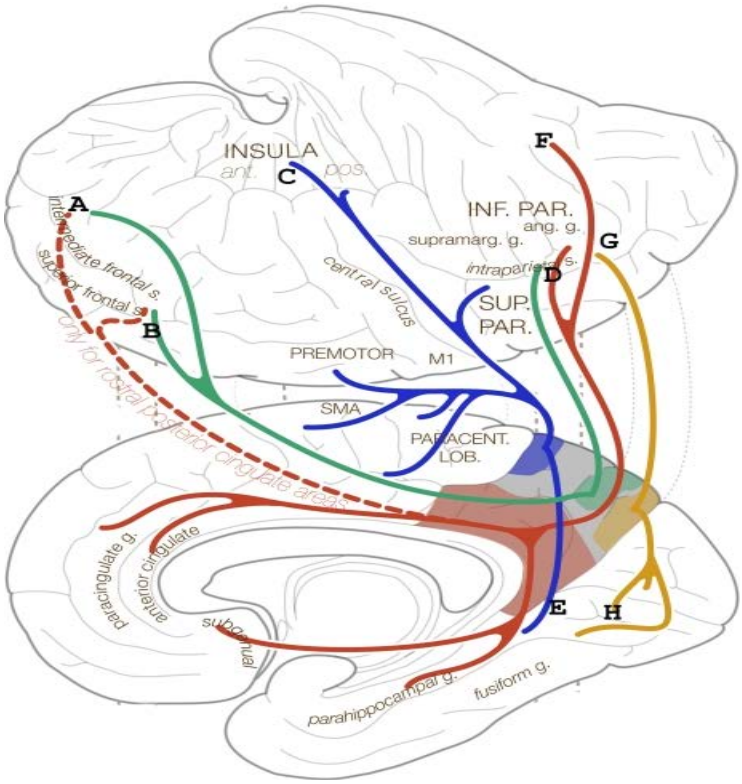

Figure 2.5

Functional connectivity maps of regions within the posterior cingulate/precuneus to the rest of the brain. *Within* this region there is likely a small world of high interconnectivity where only some nodes within this region are connected to other systems. Red: Posterior cingulate, which has connectivity with limbic structures and the frontal lobe as depicted. Blue: sensorimotor precunal region, connected to the insula, somatosensory cortex and premotor areas. Green: cognitive/associative precunal region. Yellow: visual posterior precunal region. (Adapted from Margulies et al., 2009.)

Having provided some examples of these graph theoretic structures, let us now consider the default network in more detail. The default network is named after its discovery in the observation that subjects at rest—"rest" being where participants are left in an fMRI machine without any explicit task instructions—appear to show a consistent, robust pattern of cognitive activity (Raichle et al., 2001; reviewed in Buckner et al., 2008; Buckner, 2012). Since its discovery the default network has been implicated in many other non-rest tasks such as mentalization (it is not "task negative"; Spreng, 2012). Anatomically, the default network is summarized in the figure below (Figure 2.6).

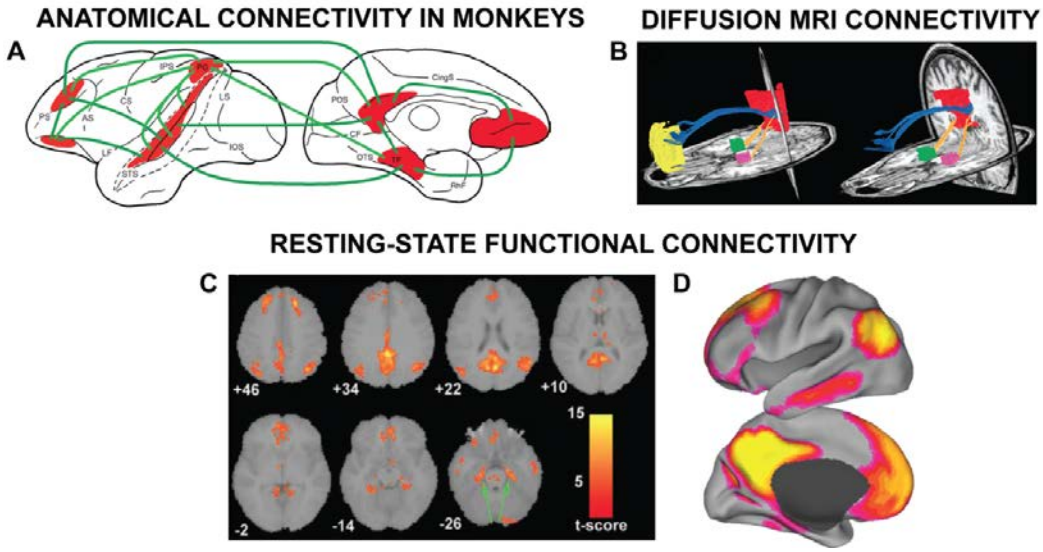

Figure 2.6

(A) major white matter connections in monkeys (macaques) through anatomical tracing studies in macaques, (B) diffusion tractography for a human participant highlighting white matter tracts, (C) functional connectivity map of default mode network using independent component analysis, (D) resting state connectivity using seed-based blood-oxygen-level-dependent (BOLD) response. (Adapted from Andrews-Hanna, 2012)

As can be seen by inspection of Figure 2.6 (c and d), the default network comprises coordinated activity throughout a number of regions: briefly, it includes the medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), superior and inferior frontal giri (SFG & IFG), the posterior inferior parietal lobule (IPL) and medial and lateral temporal lobes (MTL) (Spreng, 2012; Andrews-Hanna, 2012).

The default network is organized into two subsystems and has two major hubs (Figure 2.6a). One hub is the aforementioned posterior cingulate cortex (PCC), which seems to function as an integrative mechanism between networks. A second hub is found in the anterior medial prefrontal cortex. These regions are hubs because they are highly connected nodes within the default network (as well as with other systems of the brain). The two subsystems (sets of nodes within with high functional integration) are categorized into a medial temporal lobe subsystem and a dorsal medial prefrontal cortex subsystem.

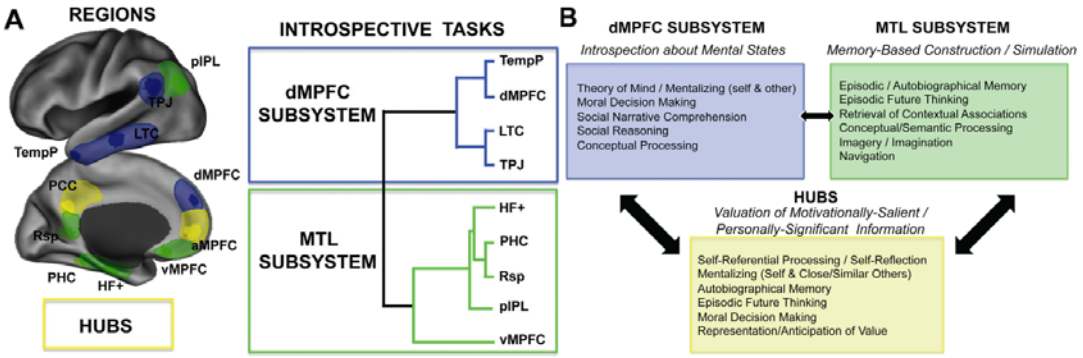

Figure 2.7

(a) Two subsystems of the default mode network, comprising of the dorsal medial prefrontal cortex system and a medial temporal lobe subsystem, as identified by a graph theoretic analysis; (b) Summary of tasks that these subsystems and hubs play an important functional role in. See (Andrews-Hanna, 2012) and text for a detailed discussion.

In her review of the adaptive role of the default network in internal mentalization, Andrews-Hanna describes how each subsystem is preferentially associated with a certain range of tasks (Figure 2.7b). The dMPFC subsystem (blue) is involved in mentalization, moral decision making, social narrative comprehension and social reasoning, as well as being preferentially involved in introspective tasks concerned with mental states. The tasks that the MTL subsystem (green) is especially involved in episodic autobiographical memory and future thinking, contextual associations, imagery and spatial navigation, as well as being preferentially involved in memory-based scene construction.

With the default network's parts and some subsystems briefly introduced, let's return to a discussion of the default network and cognitive homology. If any network deserves to be called a cognitive homologue, it is the default network. Homologues are homeostatic property clusters that can vary largely independently of each other in their character state (both across individuals and within individuals). If the default network is a cognitive homologue, it should be describable in these terms and operational criteria can guide one towards evidence supporting this view.

Interestingly, the default network appears to be evolutionarily ancient. Lu et al. (2012) report that they found a network "homologous" to the default network in rats, which extends previous research that has shown the default network in nonhuman primates:

> [We] have identified multiple networks in the anesthetized rat brain that are homologous to those reported in humans. … Our data suggest that, despite the distinct evolutionary paths between rodent and primate brain, a well-organized, intrinsically coherent DMN appears to be a fundamental feature in the mammalian brain whose primary functions might be to integrate multimodal sensory and affective information to guide behavior in anticipation of changing environmental contingencies. (Lu et al., 2012)

The evidence that these are homologous networks come from high correspondence in the structures involved, anatomical connectivity among structures, and functional connectivity among the network—special qualities distinctive of the default network. The network's functional connectivity is heritable as well. In fact, its functional connectivity is influenced by genetic factors that cannot be attributed to anatomic variation of the components in the network (Glahn et al., 2010).

The default network's characteristic changes in disease support the view that it is a homologue. It appears to be tightly coupled with Alzheimer's disease, such that its morphological structures (in the subsystems/hubs) are affected similarly by amyloid plaques (Buckner et al., 2008). These structural changes reflect changes in functional connectivity in aging. Indeed, a large amount of research is directed at analyzing the default network's role in mental disease, as in many cases of mental disease such as schizophrenia there are characteristic changes over time that depart from normal functional connectivity within this network and among other networks (Broyd et al., 2009). Between individuals there is a correspondence in the way that the network varies in its character state (whereas other characters, such as motor systems, are not so affected).

This suggests that the default network is a homologue, at least in humans (i.e., across individuals), but let us turn to the topic of serial homology in particular, and consider the revised operational criteria developed earlier in this chapter. In order to do this, at least two putative serially homologous default network activity-functions must be proposed. A reasonable suggestion is that two serially homologous activities correspond to the behavior of the two subsystems of the default network: (1) the default network in a character state described by dMPFC-subsystem and hub activity ("DN-dMPFC"), and (2) the default network in a character state described by MTL-subsystem and hub activity ("DN-MTL"). Such putative serial homologues seem to be associated

with ways in which humans reason about the mental states of others: for tasks involving close/similar others (e.g., family), activity corresponds closely with the MTL+hub activity similar to that in scene construction and imagery, whereas for distant others activity corresponds more closely with dMPFC+hub activity similar to that in social reasoning and narrative comprehension (Andrews-Hanna, 2012). Although I consider the operational criteria below, the current state of research makes it is difficult to provide more than suggestive evidence or to say more than what kind of evidence would be relevant.

Consider continuity first. Continuity was defined as identifying a correspondences in causal properties of an activity as the activity develops from an immature state to a more mature state of that trait. I highlighted two important kinds of cases of continuity: (1) activity that shows duplication and divergence in character state over time, and (2) correspondence in the ways that character states are modulated over time. Evidence of the first kind could be provided by identifying two tasks that in adulthood produce activity corresponding to DN-dMPFC and DN-MTL and comparing these to DN activity in children. Evidence in the first form of continuity would be present if DN activity that normally activates one particular subsystem does so less selectively. Evidence of the second kind can be found in mental disorders that alter the integrity of the default network. In Alzheimer's disease, connectivity is altered among a number of components in both subsystems, tokens of DN activity should be altered in similar ways (in either character state), e.g., in a correlated decrease in node connectivity strength within these system.

Now consider position: the proximal spatiotemporal position of an activity within a more general pattern of organization. Without detailed models (of which there are none at present), it is difficult to find any examples of a correspondence in spatiotemporal positions. There is some suggestive evidence however, and this concerns the interaction between the default network and other cognitive networks. Gao and Lin (2012) provide some evidence that a frontoparietal task control network regulates the default network and dorsal attention networks (Figure 2.3a). Combined with results that indicate that this is so for tasks that show activity corresponding to the putative serial homologues

(e.g., a social reasoning and episodic thinking task), this would provide positional evidence of serial homology.

Finally, consider special quality: the complexity, distinctiveness or specialization of that trait, which refers to correspondence in properties that are unlikely if the traits are not homologous. One potential special quality could come from epiphenomenal byproducts resulting from the recruitment of a cognitive network, by which I mean cognitive activity that owes to the recruitment of a network but does not contribute to accomplishing the task at hand (Klein, 2010). The default network is often anticorrelated with the dorsal attention network, so that increased activity in one network is correlated with decreased activity throughout the other. Task-induced deactivations of the default network show deactivation of regions in both subnetworks (at least in meta-analyses) (Andrews-Hanna, 2012). The networks are not just anti-correlated but also cooperate in certain tasks, such as in purposeful visual imagery (which would exemplify the DN-MTL character state). Special quality would be present here if, in a task involving the dorsal network and DN-MTL, there was less task induced deactivation in DN-dMPFC as well.

Networks provide unique challenges in identifying special qualities. The criterion of special quality is motivated by the consideration that distinctive properties are unlikely to be due to homoplasy or convergent evolution (by functional demands in the environment). For serial homology of cognitive activity, whether some shared property should be considered a special quality indicative of homology turns on whether it might instead by attributed to common functional demands. The structure of cognitive activities at the network scale are highly reorganizable, and this presents a difficulty for evaluating special qualities at this scale (cf. Anderson et al., 2012).

Cognitive homologues as mechanisms

In closing for this chapter I would like to return to mechanistic explanation and offer some speculative suggestions for integrating it with the notion of cognitive homology. First consider functionalism as a theory in cognitive science, which says that psychological kinds are at the cognitive level, a level which is best viewed in functional terms and as being very loosely coupled with the underlying neural architecture. The multiple realization of mental states

and other psychological kinds is supposed to support the view that these are loosely coupled levels, and this in turn preserves the autonomy of psychology as a science without sacrificing the causal efficacy of mental states or physicalism (Polger, 2012a, 2012b). The realization relation is central, as it says how psychological traits are made up by brains but are not identical to states of the brain; the brain-psychology relationship is many-to-one.

It has been suggested that homology also exhibits some sort of multiple realization (Ereshefsky, 2012). Ereshefsky points out to a case of "multiple realizability" of grasshopper mating calls—a type of activity—as it is realized by non-homologous morphological parts in different species of grasshoppers (different organs are used to produce the mating call). Ereshefsky calls these cases of "hierarchical disconnect" and they are a part of the broader phenomenon of homologues being able to vary in their parts and developmental mechanisms; there is hierarchical disconnect in both spatial and temporal scales. One wonders whether the hierarchical disconnect in homology can play the same role as multiple realization does in keeping the brain-psychology relation many-to-one.

The answer is that hierarchical disconnect does and doesn't: on the one hand, hierarchical disconnect does say that the brain-psychology relationship is many-to-one, but this does not imply the multiple realizability of mental states. Let me explain. It *does* support the many-to-one relationship in so far as serially homologous cognitive activities can vary in their architectures or their component parts. However, multiple realization is often taken to require some substantive difference in kind at the level of the realizers; it is more demanding than hierarchical disconnect. Polger (2012) uses the analogy of the multiple realization of corks: having two corks in different colors does not count as a proper instance of multiple realization, even though the properties of these corks vary in one way or another. What is needed is a *relevant* difference, one that marks a difference in type at the level of the realizers, so as to prevent type identities from reducing psychological kinds to neural systems. If two *types* of things realize being a cork, *then* there is multiple realization of corks.

When this is applied to cognitive homology and hierarchical disconnect, it is not immediately clear that one have or even can have multiple realizability

in the requisite sense. Imagine a cognitive activity, such as the default network, and suppose that it there are serially homologous states of this activity corresponding to the two subsystems discussed above. Here we have some fairly interesting differences in organization, component activities and the like—in short, there is hierarchical disconnect. But do these differences constitute differences in kind? It seems the answer is no, for (non-)homology is precisely the relation that distinguishes between types of cognitive activities. Accordingly, if psychological kinds are types of cognitive activities, psychological kinds are thereby type-identical to types of cognitive activities, and the differences between the properties of these systems at various levels are thereby, of conceptual necessity, *not* the kind of relevant difference that supports genuine multiple realization of psychological kinds. In summary, whereas cognitive homology allows for hierarchical disconnect and so makes the brain-psychology relation many-to-one, it does not do so in a way that allows for multiple realization—the cognitive activity-psychology relation (between types of cognitive activity and homologues) is one-to-one.[15]

There are a number of advantages of homology thinking over functionalism (in addition to considerations of causal depth). First, homologous traits in different organisms are phylogenetically related when homologous in the phylogenetic sense, so explanations of evolutionary transformations are possible, in contrast to explanations of evolutionary transformations of functional roles (they do not evolve).[16] Secondly, developmental commonalities among homologues suggest shared developmental constraints, which define the potential for future morphological change. Finally, even though there may be

[15] This does not conflict with how for serial homology, there are two types/states for that homologue. For serial homology involves two (sub-)types of one and the same activity, but these pick out two different levels of generality (sub-types and types). The level of generality is the issue here: e.g., red and white blood cells are two (sub-)types of one and the same blood cell, but they are also one type simpliciter: namely, the blood cell. Red and white blood cells, taken as sub-types, map one-to-one with sub-types of the blood cell, and red and white blood cells, taken as blood cells, map one to one with the type of blood cell.

[16] That is to say, functional roles do not evolve except in so far as there is a change in functional role due to the evolution of its realizing parts. The function itself, however, is not the kind of thing that evolves; when a function changes in function, it is just a different function, rather than the same function taking on a different functional role. This is the point that only *activity*-functions are, strictly speaking, evolutionary characters.

hierarchical disconnect, the homology perspective plays close attention to the compositional and causal relations among entities on several levels.

Recently, Craver and Piccinini (2011) argued that functional analyses are best construed as *mechanism sketches*, which refer to mechanisms but omit many details of these mechanisms. This is in line with the way functional architectures are increasingly used in cognitive science and the overall shift in cognitive science away from the classical picture of loosely coupled computational and implementational levels. This allows for mechanistic models to include entities, activities *and functional roles* as parts, so long as the latter are understood as sketches of mechanisms.

How might cognitive homology and mechanisms be integrated? In Chapter 1 I pointed out that mechanisms are a broader class of things than activity-homologues, as mechanistic explanations can be given for things that do not correspond to particular cognitive homologues. But with some tweaking, mechanisms and homology can be brought into close alignment. In particular, it has to do with how mechanisms are characterized and the nature of cognitive systems as being highly dynamic. Mechanisms are activities and entities arranged so that they *do something*, they are "acting entities." This is ambiguous between two senses of use: how they are used—that is, how they operate when they are used—and what its operation is used for accomplishing.[17] Mechanistic explanation proceeds by way of characterizing the explanandum phenomenon *completely*, in such a way that includes its "precipitating, inhibiting, modulating and nonstandard conditions, and of its by-products" (Craver, 2007, 128). This indicates that properly characterized mechanisms are not characterizations of what they are used for, but *how* they are used. So mechanisms are about activities (or sets of interacting activities), even though it seems possible that they not every group of activities corresponds to a cognitive homologue.

I want to tentatively suggest that mechanisms, properly characterized and in the context of cognitive systems, correspond to cognitive homologues. That is to say, mechanisms are amenable to homology thinking, so that one could

[17] Example: *how* hammers are used = hammering, throwing, and other activities; *what* hammers are used *for* = driving a nail through a structure, aligning a board, cracking open an object, and other things it can be used for. When a person uses a hammer to drive in a nail by hammering it in, how it is used = the hammering activity, what it is used for = driving the nail in.

gain certain explanatory and methodological purchase when thinking about mechanisms as with cognitive activities. So far my discussions have been restricted to fairly localized cognitive activity and a few relatively robust cognitive networks, but what about all of the other cognitive activities? The organization of cognitive activities for a given task exemplifies what some have called "soft assembly." Softly assembled cognitive systems are such that these cognitive activities are task-facilitative groupings that are temporarily constrained to act as a single coherent unit (Kelso, 1995; Kello et al., 2007). My suggestion is that homology thinking can be used in conceptualizing softly assembled cognitive systems, and in doing so allows for bringing mechanisms and evolutionary characters in closer correspondence. A number of potential problems are immediately apparent. First, how can homology thinking even apply to 'task-facilitative' groupings (as I have called them) in the first place, given that their reliance on functional demands? Secondly, in virtue of what can softly assembled systems be said to be historical, homeostatic property clusters, as expected by this account of cognitive homology? Finally, what is the value in homology thinking here? I address these three issues in turn.

It will help to have an example of such an activity. An example borrowed from cross-cultural cognitive neuroscience is particularly germane as it illustrates a number of relevant issues (Kitayama and Uskul, 2011). Murata et al. (2012) showed that although Asians and European Americans can suppress overt emotional reactions to stimuli, Asians but not European Americans can suppress processing of emotional reactions to disturbing stimuli, as revealed by a parieto-frontal inhibitory activity in Asians, but a different activity in European Americans. It turns out that, when the task is to inhibit emotional processing of disturbing stimuli, the softly assembled systems differ according to one's cultural practices. These differences owe to differences in cultural attitudes/practices regarding emotional expression and regulation. The softly assembled systems here refer to the coordinated interactions of parts of motor, visual, attentional and affective systems—the mechanism for the cognitive activity.

Let us consider the potential issues for homology thinking here. One might wonder how task-facilitative groupings can be viewed as homologues, given that they are assembled in accordance with one's current task demands.

The issue is that these softly assembled systems rely largely on functional demands, and so these seem more like analogies or some other functional kind rather than homologues. However, it is important to clearly distinguish between what kind of thing a psychological trait is taken to be, and its causal-developmental mechanisms. It is one thing to say that a trait is a homologue rather than a functional kind, and quite another thing to say that its developmental mechanisms do or do not heavily on environmental input. After all, many activities and behaviors depend heavily on environmental conditions and are learned such as a language, in humans, or directly suppressing emotional processing of disturbing stimuli, in Asian cultures, yet these can be considered as evolutionary characters all the same.[18] So a reliance on task demands and environmental input does not preclude homology thinking.

Since softly assembled cognitive activities can be viewed as homologues, it should be possible to describe them in terms of what homologues are – that is, in terms of being historical (evolutionarily or developmentally) homeostatic property clusters. Consider the system that is assembled for emotional suppression Asian participants versus that of European American participants. Let us suppose that these are not homologous cognitive activities. If so, then these activities have their evolutionary/developmental origins in particular cultural practices (plus soft assembly mechanisms, which may not differ between cultures). The softly assembled systems have different developmental mechanisms that are redeployed in this system. What about being homeostatic property clusters? Within a culture, the assembled system for the emotional suppression task seems to be a stable type, as its homeostatic integrity is maintained by common soft assembly mechanisms and cultural practices. Such cultural practices are just a special case of how cognitive activities become more tightly coupled systems as they are repeatedly used, and soft assembly mechanisms are sensitive to this fact. This further indicates that the homeostatic integrity of a softly assembled cognitive activity is substantially dependent on the causal-developmental history of component activities.

[18] Indeed, even the development of anatomical structures depend on environmental input and past activities. For example, starting on the third day of incubation, the chick embryo assumes a pattern of increasing and decreasing motor activity, activity which is necessary for the subsequent formation of cartilage, bones, joints, and other structures (Müller, 2003).

For the above reasons, and in accordance with the cross cultural neuroscience example, serial homology of tokened cognitive activities within an individual can be expected to occur to a greater degree than homology of softly assembled cognitive activities between individuals (people, cultures, etc.). The cognitive homologue has a higher propensity for redeployment among tasks. And this relationship holds in so far as these groups or individuals exhibit high similarity in the causal-developmental properties that are homeostatic mechanisms for such softly assembled systems.[19]

Finally, what is the value in homology thinking here? First, by construing softly assembled systems as cognitive homologues the psychological kinds amenable to homology thinking include all psychological kinds, rather than being limited to a number of intrinsic networks and local cognitive activities.[20] Secondly, and consequently, it allows for a construal of mechanisms in terms of homology. Previously I suggested that not any group of cognitive activities will not correspond to a cognitive homologue, which indicates that mechanisms outstrip cognitive homologues in their scope. This is so even when mechanisms are characterized by *how* they are used, including by-products, side effects and other properties of the explanandum phenomenon, for there will only be a correspondence between mechanisms and cognitive activities/homologues for intrinsic networks and local (small-scale) cognitive activities. But when the scope of cognitive homology includes softly assembled systems, there will be a correspondence in mechanisms and cognitive homologues that is not limited to intrinsic networks and local cognitive activities. Much like the mechanism for an action potential corresponds to the action potential's activity-function (they have corresponding structures), the mechanism for softly assembled systems such as the emotion suppression task corresponds to an activity-function homologue. [21]

---

[19] To be sure, this is distinct from the claim that for a given activity, it will show less *variation in form* within an individual than between individuals (even if this is also reasonable to suppose).

[20] One might think that this does not show why we need homology thinking as opposed to just thinking about activity-functions. However, this is two ways of saying the same thing, provided one individuates activity-functions as biological characters rather than as functional roles.

[21] A reviewer has emphasized that I have not presented any detailed examples for what one gains by doing so, and this section is underdeveloped in this important respect. To the question, "why does anyone care that it's a homologue?" I have not presented any defense, except in so far as one might think that a homology approach to mechanisms

Chapter 3: Imagination and representational codes

Using the cognitive homology concept, this chapter aims to provide some theoretical contributions for thinking about the imagination and its relationship with belief. Homology is a biological relationship of sameness; it says what makes two characters one and the same thing. Homologues are also natural kinds, and in being the historical kinds they are one can generally expect for homologous traits to share 'deep' causal commonalities. In the context of cognition, identification of homologues consists in identifying the same cognitive activity among its varieties of forms and functions, and cognitive homologues can be expected to share deep causal and computational commonalities—that is to say, homologues are disposed to behave in much the same way in cognitive systems. These considerations form the backbone of the upshots for understanding the imagination.

This chapter consists in two main parts, the first on a specific application of the homology concept in relation to the imagination and the second is a more general proposal concerning the nature of representational codes or "formats." For the first part, I argue that cognitive homology allows us to understand why corresponding imagination and belief representations (or non-imaginative representational states more generally) are disposed to behave in much the same way among potential cognitive systems despite having different functional roles in these systems. For this, I focus on Shaun Nichols and Stephen Stich's (2000, 2003, 2006) influential work on the functional architecture of the imagination and Nichols' (2004, 2006, 2008) "single code" explanation of this similarity. I argue that as it stands this explanation is inadequate, but cognitive homology can accomplish the explanatory goal of common coding explanations.

For the second part, I argue that cognitive homology can be used to give an account of what a representational "code" or "format" is. Although the notion of a code is familiar in cognitive science, there are no theories about what it is to be a code, "and little if anything has been written about the criteria of sameness or difference for such codes" (Goldman, 2012, 73). I argue that representational

---

and to soft-assemblies seems to be topics which may be promising to explore in more detail. I hope to develop this line further in future work.

codes or formats are *homology classes* (relatively large sets of cognitive homologues).

Before beginning, a brief discussion of the imagination is in order. Accounts of what the imagination is vary extensively in the literature, and are usually characterized by a series of examples. Contemporary accounts of the imagination make claims about its overall computational architecture, such as metarepresentational and offline simulation theories (reviewed in Gendler, 2011). For example, metarepresentational architectures involve the representation of cognitive states as such, which as units are operated on in cognition, whereas off-line simulation theories need not involve manipulation of representations of cognitive states but rather manipulation of one's own (first-order) cognitive states and some way of distinguishing offline and online states. However no particular account need be assumed in the following, as the empirical fact of certain causal similarities between imagination and beliefs or perception is independent of which architecture of the imagination is correct.

The existing literature in philosophy distinguishes between aspects of the imagination in a number of ways (Figure 3.1). Steven Yablo (1993) divides the imagination according to what kind of thing is represented—the kind of object one's representational state refers to; these three types of imagination are *propositional*, *objectual*, and *active* imagination, depending on whether one represents a state of affairs, an object, or an action to oneself, respectively. One can also distinguish between the way that imagination is connected to other cognitive systems, as these may vary among modes of imagination. For example, Kenneth Walton (1990) distinguishes between *spontaneous* and *deliberate* acts of imagination as well as *occurrent* and *nonoccurrent* activities of imagination (which do or do not occupy the subject's attention, respectively). Many accounts distinguish between a *belief-like* imagination and "perception-" or *sensation-like* imagination. One may also add, that there is *action-like* imagination. These kinds of imagination are defined by their close relationship with occurrent belief, sensory states and action execution. Belief-like imagination has been of special interest to philosophers, and has been refined in theories of "cognitive" (McGinn, 2004, 2009), "suppositional" (Goldman, 2006b), or "propositional" imagination (Nichols and Stich, 2006). Overall, these accounts cross-classify each other with

some independence. In particular, although representational objects are associated with a resemblance to other types of cognitive states, these two properties can vary independently. An action could be represented in a belief-like state, for instance when supposing that a person has performed an action rather than imagining it by using vivid motor imagery.

Individuation basis	Example	Label
Representational object	State of affairs	Propositional imagination
	Object	Objectual imagination
	Action	Active imagination
Involvement of cognitive systems	Action control systems	Deliberate vs. spontaneous imagination
	Attentional awareness systems	Occurrent vs. nonoccurrent imagination
Analogy with other type of cognitive state that shares distinctive features	Occurrent belief	Belief-like imagination (cognitive, propositional, suppositional)
	Sensory states	Sensation-like imagination
	Action execution	Action-like imagination

Figure 3.1

Table of different classificatory systems for the imagination, as based on the representational object, involvement of other cognitive systems, or by analogy with another type of cognitive state that shares distinctive properties with the corresponding imaginative state. The third type, highlighted in red, is the most germane to this chapter.

Whatever the imagination is, and whatever its distinguishable components are, it seems that there always is some corresponding non-imaginative counterpart (such as an executed action, sensory state, or belief), but the latter are not easy to describe at once. I know of no easy word or phrase that characterizes all of these non-imaginative counterparts at once. In the following, I will simply refer to the non-imaginative counterpart to an imagination representation as "nonimagination" *simpliciter*. However, I also refer to this contrast of nonimagination and imagination as one between "belief" and "pretense" representations as these are the terms Nichols uses. Although there may be

differences in what these are (or how Nichols and I construe the terms), these do not affect the explanatory project in this chapter and so can be ignored for our purposes: "belief" and "nonimagination" are used interchangeably, as are "pretense" and "imagination."

Common coding explanations of computational similarity

This first section argues for the relevance of cognitive homology in explaining the similarities between imagination and corresponding nonimagination representations, in particular a tendency for common causal/computational behavior in cognitive systems. I present this in the context of Nichols and Stich's (2000, 2003, 2006) influential work on the architecture of the imagination and Nichols' (2004, 2006, 2008) favored explanation of this similarity.

This first section sets up the explanatory project of explaining a disposition for corresponding imagination and nonimagination representations to be processed in much the same way by cognitive systems. This disposition is inferred from the similarities between corresponding imagination and non-imagination representations. The similarities have to do with the way in which corresponding, "isomorphic," or "parallel" representations seem to be disposed to be processed in much the same way by cognitive mechanisms.[22] Nichols' explanation is that this is because corresponding representational states are in a "single code," or a *common* code, to put the thesis in more general terms.

Before getting into the purported explanation of the relevant similarities between imagination and nonimagination, a survey of some of them is in order. First, both imagination and nonimagination representations exhibit *inferential orderliness*. Inferential orderliness refers to the *pattern* of inferences that tend to be followed by corresponding tokened imagination and nonimagination representations. If one supposes that there is a full glass of water on a table and that someone knocks the glass over, one readily infers that the water spills out of the glass. Likewise, if one actually believed that a glass of water was tipped over,

[22] What makes two representational states, such as an occurrent belief and occurrent pretense, isomorphic? It is hard to say exactly, but it has a lot to do with representing the same thing in the same way: for the most part, believing that p and supposing that p are corresponding states, as are being in a sensory state p and vividly imagining p using the same sensory modalities. But supposing that p (=one sees an explosion) and being in a sensory state p (=seeing an explosion) are not corresponding states.

one would infer that the water has spilled. This characteristic pattern in inferences is followed by both imagination and nonimagination representations.

Both imagination and belief also engage emotion or affective systems in similar ways, exhibiting a sort of *affective orderliness*. One tends to have the same characteristic pattern of affective responses to imagined scenarios as one would have from corresponding beliefs or sensory states. Fiction is a clear case in which this is so: one does not believe that the events in a work of fiction have actually occurred, but one can feel saddened, elated or fearful all the same when engrossed in a work of fiction. In this case, it seems that the more vividly one imagines a sensory-perceptual situation, the more affective systems seem to be activated.

A third example: actions and imagined actions also share a number of striking similarities. For instance, imagining performing an action such as writing, walking or tapping one's finger. Generally speaking, imagined actions tend to preserve the same spatiotemporal characteristics and obey the same laws of movement control, such as Fitts's law[23] (Papaxanthis et al., 2002). The list goes on, but this gives some idea of the similarities under question. Overall, there are deep causal and computational commonalities in the way corresponding imagination and nonimagination representations are processed by cognitive mechanisms, including affective, inferential, motor control, and other mechanisms.

In general, it seems that belief and imagination representations are processed in much the same way by cognitive mechanisms. How can these similarities be explained? One answer is that they are in a *single* or *common representational code*.[24] By having a common representational code, one hopes to explain why belief and pretense are so similar and in contrast to desires, whose dissimilarities in these respects are consistent with them *not* having a shared representational code. An early common coding explanation is found in Alan

[23] Recall that Fitt's law is the speed-accuracy tradeoff, such that the smaller an object is, the smaller the movement destination is, and the larger the distance, the slower the action.

[24] Leslie argued that there was just a single, primary representational code. Nichols (discussed below) also originally argued for a "single code" explanation, but later modified his view to allow for the possibility of more than one code (Nichols, 2008). The important thing here is that the code for corresponding imagination and pretense representations is *in common*.

Leslie's (1987) metarepresentational account of pretense. Briefly, "primary representations" are the representations used in beliefs and desires. In order to imagine something, the primary representations are decoupled from this primary representational code (for belief), and then "marked" as pretense. So <*I BELIEVE* 'there is a cup on the table'> feeds into a decoupling mechanism and then is tagged as a pretense representation, resulting in the cognitive attitude <*I PRETEND* 'there is a cup on the table'>. Yet the proposition 'there is a cup on the table' is in one and the same representational code throughout, an identical state found in both belief or pretense representations. I take it that what this means is that in a suitably constructed functional architecture, one would expect that the decoupled representations are one and the same thing; they correspond to one and the same "box" in that architecture.

Here is one sketch on how this could go. This is not intended to be a correct proposal (nor Leslie's architecture), but rather to show how it could be that being in a single code might explain similarities in processing under an decoupling-and-tagging architecture.

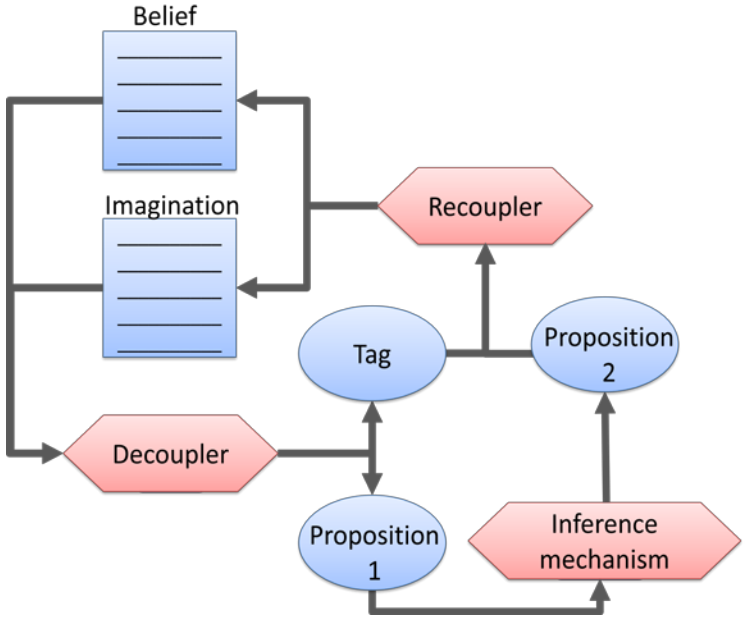

Figure 3.2
Example of an architecture for a tagging architecture in inference formation. Untagged propositions are the one and the same, computationally speaking, regardless of whether they are decoupled from belief and imagination representations, and this implies similar treatment by the inference mechanism.

What the architecture in Figure 3.2 depicts is a decoupling-and-tagging architecture for inference formations. In the figure, the code for imagination and belief boxes are drawn, and they house belief and imagination representations respectively. In order to enter into inference formation mechanisms, the 'decoupler' separates the 'tag' from the 'proposition', and the decoupled proposition is what is sent to inference formation mechanisms (or affective systems, etc.), yielding a further uncoupled proposition that is subsequently re-tagged appropriately to form a new cognitive attitude token. A tagging account says that cognitive attitudes are composed of cognitive tags and propositions in a primary representational code, and cognitive mechanisms standardly operate on the decoupled proposition. Pending the addition of further computational steps, mechanisms will not differentially respond to the tag of the input proposition, and this implicates a disposition to be processed similarity. Because the decoupled representation in corresponding tokened belief and imagination representations are in one and the same representational code, the decoupled proposition for corresponding imagination and belief representations will be one and the same, and this explains why they tend to be processed similarly by cognitive mechanisms that take both as input.

Inspired in many ways by Leslie (1987), Nichols (and Stich) also puts forward a common coding explanation. This is done under the background of a particular cognitive architecture for pretense and imagination and is with respect to this architecture for pretense that Nichols aims to explain how representations are processed in much the same way by cognitive mechanisms by the appeal to a common code. Nichols & Stich's architecture is for pretend *play*, but "pretense representations" are imagination representations all the same. Nichols and Stich argue for the existence of three distinct cognitive attitudes (DCAs, for short). These DCAs are pretense representations, belief representations and desires.[25] The architecture is displayed in Figure 3.3 below.

[25] It has been recently suggested that a "single attitude" account is correct, where both imagination representations and beliefs are reduced to beliefs only (Langland-Hassan, forthcoming). These debates are orthogonal to the issue of understanding the relevant computational relationships between corresponding imagination and nonimagination representations, since the general explanatory strategy is at issue here and not how many attitudes there are. If reduced to one, there are still corresponding states, but in this case they are both belief states—still, an explanation needs to be given why they tend to be processed similarly.

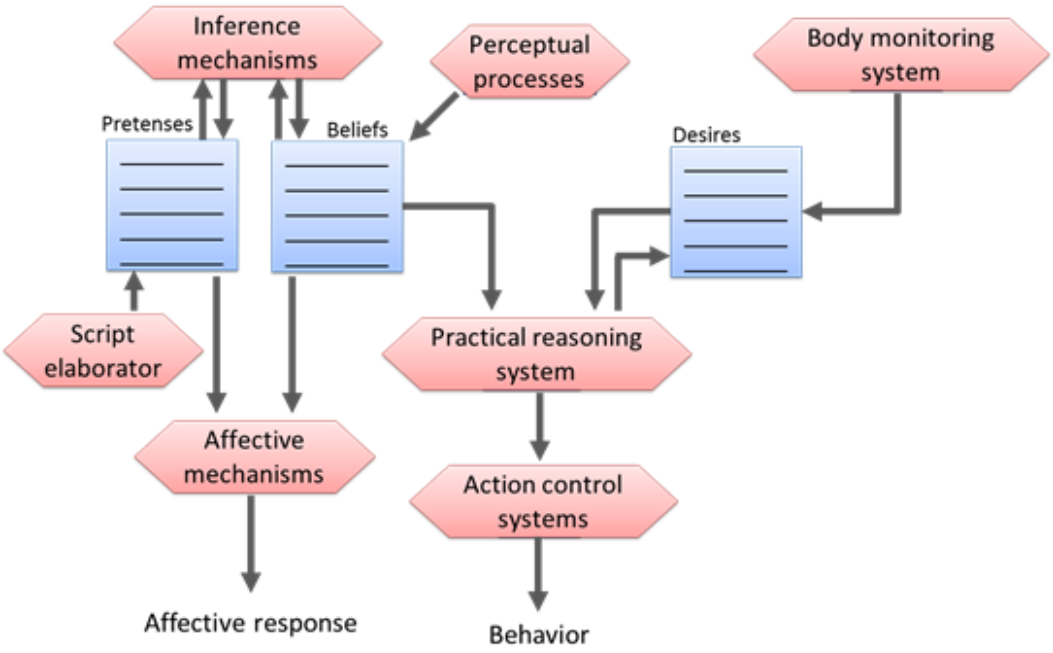

Figure 3.3
Functional architecture for pretend play. (Adapted from Nichols, 2004.)

This architecture presents a number of cognitive mechanisms (hexagons) and cognitive attitudes (squares), which are connected by arrows indicating causal connections between these functional components. The units in the architecture are largely individuated by their causal role within the overall cognitive system, where this causal role consists in how it figures in the overall architecture. One thing that distinguishes beliefs from pretenses is that beliefs, but not pretenses, enter into practical reasoning systems. Pretense representations are not directly influenced by perceptual processes, but instead require perceptual processes to give rise to beliefs as an intermediate computational step. At the same time, both pretenses and beliefs interact with inference formation mechanisms and affective mechanisms

Nichols argues that common coding explains the similarities in processing between imagination and belief representations. What in particular does common coding say about the cognitive attitudes in this architecture?

> The key point is just that, if imagination representations and beliefs are in the same code, then mechanisms that take input from the 'imagination box' and from the 'belief box' will treat parallel representations much the same way. For instance, if a mechanism takes imagination representations as input, the single code hypothesis maintains that if that mechanism is activated by the occurrent belief that *p*, it will also be *activated* by the occurrent imagination representation that *p*. More generally, for any mechanism that takes input from both the pretense box and the belief box, the pretense representation *p* will be processed much the same way as the belief representation *p*. (Nichols, 2008, 525, cf. Nichols 2004, 130, Nichols 2006, 461)

Under this construal, common coding grounds the counterfactual claim about the behavior of these cognitive attitudes: if a given mechanism were to take both kinds of representations as input, that mechanism "will process the pretense representation much the same way it would process an isomorphic belief" (Nichols, 2006, 462). Importantly, this *is* a counterfactual claim according to Nichols:

> This is a general hypothesis about psychological mechanisms, not specific to the inference mechanisms. If, for instance, emotion systems receive input from the imagination, the single code hypothesis predicts that these systems should produce affective output similar to the output that would be produced by isomorphic beliefs. (Nichols, 2006, 462)

So in addition to explaining how belief and pretense representations are *in fact* processed in much the same way by cognitive mechanisms that they interact with, Nichols argues that common coding also entails that belief and imagination representations are *disposed* to be processed in much the same way, by any given cognitive mechanism, selected at random, *should* this mechanism take both as input. Illustrative of this is Nichols' claim that, if emotion systems take input from both, common coding predicts that corresponding beliefs and imagination representations will be processed in much the same way. In this way, common coding is a *trans-architectural* hypothesis; it makes predictions about similarities in processing among many architectures such as non-standard ones.

One class of non-standard cognitive architectures are dysfunctional ones, such as would occur in a psychological disorder that disrupts the functional organization of a healthy cognitive system. In Nichols & Stich's architecture,

belief and desire representations feed directly into practical reasoning systems, but pretense representations do not. However, it is possible for pretense representations to be taken as input by the practical reasoning system, and this would look like Figure 3.4.

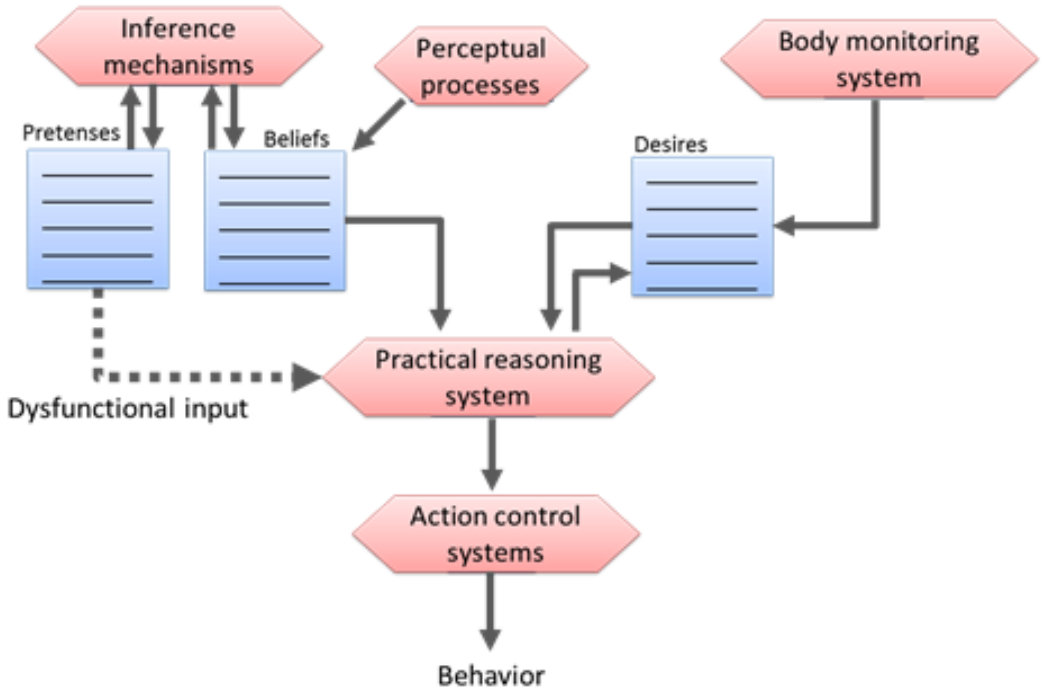

Figure 3.4
Dysfunctional cognitive architecture: pretense representations feed into the practical reasoning system (dotted line).

Common coding says that, if pretense representations were to feed into practical reasoning systems directly, they stand to be processed in much the same way as the corresponding belief representations are. Because pretense representations are in the same code as beliefs, they will be treated similar to beliefs by the practical reasoning system. (In contrast, as they are by hypothesis not in the same code as desires, they will not necessarily be treated in the same way as a desire with the same propositional content.) In a dysfunctional system this may be the case, and would describe a dysfunctional architecture according to which practical reasoning systems do not distinguish between what is imagined and what is believed. However, other features of the causal role of pretenses still distinguish them from beliefs; beliefs but not pretenses take input from perceptual processes.

Limits of common coding explanations

Alvin Goldman (2006a, 2006b, 282) has raised concerns regarding the adequacy of Nichols' common coding explanation. In order to be explanatory, after all, the *explanans* (common coding) should entail or make probable the *explanandum* (similar processing by mechanisms). Accordingly, if two tokens of the same representation in the same code occur in distinct cognitive attitudes, it ought to be guaranteed or quite probable that they will be processed in much the same way. However, since cognitive attitudes such as imagination and belief are functional roles—individuated largely by their dispositions to interact with other functional roles—one would expect them to be, in general, processed *differently*:

> It could happen, of course, that some mechanism would process a pretense representation p and a belief representation p equivalently. But why is this implied, predicted, or made probable, for a random mechanism? ... On the contrary, one would think that the distinctive functional role associated with each box or attitude type would also be relevant. And it would tilt in the general direction of difference of treatment. So why does the sameness of code imply, or make probable, sameness of treatment? (Goldman, 2006b, 282)

In sum, how can one understand a disposition to be processed in much the same way, when it is the very difference in treatment by cognitive mechanisms that is used to distinguish the attitudes to begin with? This appears to be a general problem: dispositions for largely distinct causal interactions individuate functional roles, so any appeal to their functional role cannot also ground the contrary disposition for largely similar causal interactions. According to this worry, there seems to be no principled way of holding that imagination and belief are different functional roles (= are largely causally dissimilar) while also holding that they are disposed to be treated in the same way by mechanisms – this being the claim that they are largely causally similar.

It seems that in reply one could try to distinguish between two types of causal similarity/dissimilarity. On the one hand, there are dispositions for interactions between cognitive attitudes and mechanisms *simpliciter*; for instance, beliefs but not imagination representations are disposed to enter into practical reasoning systems. And on the other hand, there is a disposition for a correspondence in the *structure* of input-output relations to cognitive mechanisms, independent of their dispositions to interact with these systems in the first place. A correspondence in structure of input-output refers to a common

pattern in the way that mechanisms process representational states. These patterns are what are highlighted in inferential, affective and other forms of orderliness shared among belief and imagination: the same affective responses to isomorphic inputs, the same inferences from isomorphic suppositions and beliefs, and so on.

As far as I can tell, Nichols' account would fare well by appealing to separate types of dispositions, viz. dispositions for causal interaction between functional roles simpliciter, and the pattern of input-output relations in the interaction with these mechanisms. After all, Nichols does seem to distinguish between them in framing the explanatory hypothesis:

> For the single code hypothesis is a proposal about the relation between certain kinds of representations and certain kinds of cognitive mechanisms. As a result, the hypothesis can only be framed against a background of cognitive architecture. Once one has posited a background of cognitive components, one can then explain some processing differences between imagination and belief by noting that some of these cognitive components take input from beliefs but not from imagination. The single code hypothesis maintains that one can also explain many of the similarities between imagination and belief with the proposal that some of the cognitive mechanisms do process input from both imagination and belief, and further, that pretense representations are in the same code as belief representations. (Nichols, 2004, 131)

What this means is that coding specifically refers to the underlying causal commonalities between corresponding imagination and nonimagination representational states, independently of the distinct functional roles of these cognitive attitudes in the given architecture. Hence the reason for saying that the hypothesis can only be framed against a background of (a given) cognitive architecture. Given a cognitive architecture that does not say why such representations are disposed to be processed in much the same way, common coding explains why, at some other, deeper level, these attitudes tend to share extensive computational similarities.

This reply seems promising, except of course that it is still at best unclear how to understand these two dispositions. What, in other words, could provide the conceptual resources for making a distinction between dispositions for interaction in general (i.e., the functional role of cognitive attitudes) and the way in which a psychological mechanism, should it take corresponding representational states as input, is disposed to treat them similarly? There are

two potential ways of going about doing this: either you specify a further functional architecture or you don't. If you don't, then there has to be a way of guaranteeing the relevant computational and causal similarities independently of describing the underlying architecture.

Take the first option first. Perhaps one could just go about describing an underlying functional architecture. So there would be two architectures: the *macro*-architecture which is the architecture for pretense (e.g., Figure 3.3, Nichols and Stich's architecture), and a *micro*-architecture that accounts for the relevant causal similarities (e.g., Figure 3.2, the tagging architecture). Multiple architectures or levels of computational analysis are required if one wants to refer to the units in the original architecture and to explain the further fact of similar treatment these are disposed to receive by cognitive mechanisms. This strategy involves elaborating a further architectural description (similar in spirit to the tagging architecture) that describes the relevant underlying sameness of code. One would describe, in detail, an underlying micro-architecture, that implies or makes it probable that mechanisms will process corresponding representational states (on the macro-architectural level) in the same way.

For example, perhaps these representations are coded as tagged perceptual symbols, such that this architecture would place both imagination and nonimagination representations in perceptual systems associated with a cognitive tag. This corresponding microarchitecture among cognitive attitudes would predict similarities in processing by any cognitive mechanisms, should they take these cognitive attitudes as inputs. It would predict similarity because, pending the subsequent addition of further computational steps in the system, mechanisms will not be able to differentially respond to decoupled imagination and nonimagination representations. In sum, a description of the code provides the second requisite architecture needed to ground the trans-macroarchitectural counterfactual about how cognitive mechanisms would treat corresponding representations, should they take both cognitive attitudes as input in another architecture.

Unfortunately, even if we had such an architecture, it would still be lacking as an *explanation*. What would a specification of an underlying architecture explain? Such an account endeavors to explain the relevant

similarities in the way that mechanisms handle imagination and nonimagination representations by describing the corresponding underlying computational architectures in detail. But this seems to restate what is being explained with a description of the shared properties to be explained; the similar underlying computational structures are appealed to in order to explain how, on the level of this underlying computational structure, they are so similar, and this is portrayed alongside another macro-architecture that distinguishes the functional role of these commonly coded cognitive attitudes. If one wants to *explain* why there are such computational similarities (i.e., to do more than model or gesture towards the underlying computational architecture to explain this underlying architecture), one should to appeal some *further* fact that says why they should be so similar.

In many ways, the problem with trying to explain common coding by appealing to the shared computational properties resembles Molière's famous quip in *Le Malade Imaginaire*, where a doctor might try to explain why opium puts people to sleep by appealing to the fact that it has a "dormitive virtue," where having a dormitive virtue just is to be disposed to cause people to sleep. One needs to appeal to a further fact, independent of having the property of putting people to sleep, in order to explain why opium puts people to sleep. The relationship between having a dormitive virtue and putting people to sleep is of conceptual necessitation, not explanatory connection. When codes are identified with underlying computational structures, appealing to sameness in code just is an appeal to the shared causal commonalities one endeavors to explain, and what one instead needs is a further fact that establishes a genuine explanatory connection between codes and underlying computational structures. Otherwise, as it stands, sharing underlying computational structures is a conceptual necessitation of common codes, rather than a genuine explanatory relationship.[26]

[26] The account advanced here does share a similarity in form with this sort of "dormitive virtue" explanation. One interpretation of the dormitive virtue account is that it uses the evidence of the effects of opium versus a placebo to conclude that opium has a dormitive virtue, which entails or weakly explains this effect. The account of cognitive homology also uses some evidence of homology to conclude that there is often homology among these corresponding tokened representational states, and this entails or weakly explains a disposition for similar processing. A detailed mechanism would make it less superficial, but the homology account advanced here is one of *many* homologues: a *pattern* of homologies for corresponding representational states, each with their own distinct mechanisms. So in contrast to the dormitive virtue account, where one reasonably

This brings us to the second option: have some way of guaranteeing the similar underlying computational properties without restating them. In other words, have an account of what makes something the same code that that also implies or makes it probable that such representations will be processed in much the same way by cognitive mechanisms, such that this account appeals to some further fact beyond there being a correspondence in computational properties at an underlying level. One possibility was presented in earlier work by Nichols and Stich (2000, 2003), and this hypothesis says that being in the same code consists of (1) sharing the same logical form (roughly, the structure of logical relations in the proposition for a tokened cognitive attitude) and (2) having their representational properties determined in the same way:

> We have suggested that [the] inference mechanisms treat the [imagination] representations in roughly the same way that the mechanisms treat real beliefs, but we have said little about the representational properties and the logical form of pretense representations. One possibility that we find attractive is that [imagination representations] have the *same logical form* as [belief representations], and that their representational properties are *determined in the same way*. When both of these are the case, we will say that the representations are *in the same code*. Since mental processing mechanisms like the inference mechanism are usually assumed to be sensitive to the logical form of representations, the inference mechanism will handle [imagination] representations and belief representations in much the same way. (Nichols and Stich, 2000, 125–6)

Under this account of common coding, a correspondence in the way of determining representational properties and the logical form of a cognitive attitude suffices to make two representational states in the one and the same code. Moreover, the sharing of these properties should imply or make probable the relevant computational similarities. However, this account remains underdeveloped and I will suggest that it does not seem very promising. First of all, the notion of a "way of determining representational properties" is left unexplicated in this proposal, and Nichols and Stich admit that "logical form" is also a notion with which there is considerable uncertainty.

---

expects a mechanism if the explanation is going to be any good, the cognitive homology account is one that ranges over many mechanisms and so no single mechanism is expected to furnish the explanation. *If* there is any mechanism that is common, it has to do with those mechanisms governing softly assembled systems and redeployment, discussed in Chapter 3 (but there are still going to often be different mechanisms for different corresponding representational states).

More importantly, these do not seem to be relevant considerations for what makes something a representational code to begin with. Although the consideration that mechanisms are sensitive to logical form might go some of the way in explaining their similarities in processing, it does not seem relevant for distinguishing between codes in general. Desires, on Nichols' account, are in a different code. So the corresponding desire that *p* and belief that *p* either always differ in their logical form, or they have their representational properties "determined in a different way." Given that they can have the same logical form, the way of determining representational properties has to be decisive in individuating codes, but it is unclear what this would amount to. Indeed one might be inclined to think that their representational properties *themselves* might have to differ, rather than the *way* in which their representational properties are determined. One wonders about how this account could extend to other codes in the literature. Dehaene (1992) proposed a "triple-code" model for number cognition, positing an Arabic, verbal and analogical code used to represent numbers, and this has been extended with the addition of a fourth, finger-representational, code (Penner-Wilger and Anderson, 2008; Di Luca and Pesenti, 2011). So there should be a distinction in the way that representational properties are determined among Arabic and verbal codes for representations of mathematical entities, but without an account of what makes a representational property determined in a way it is unclear how it marks out the relevant differences for individuating codes. The point here is that it is unclear how logical form or the "way of determining representational properties" could be a satisfactory account of codes, given that logical form is not relevant for individuating codes and the notion of a way of determining representational properties is unexplained.

It is worth noting that in subsequent work, Nichols (2004, 2006, 2008) drops this account of codes and discusses it in general terms only, saying for instance that "it's far from clear what the code is for belief representations, so it's not possible to be specific about the details or the nature of the putatively shared code" (Nichols, 2008, 525). The important thing, he says, is that common coding

is that which explains similarities in processing,[27] and is widely used for this explanatory purpose, citing a range of authors who use common coding to explain the same phenomenon (*ibid*). This would be even worse of an account of codes (should it have been presented as such) since it is not a theory of codes but what *counts* as a theory of common coding. Hence, it describes the *explanatory goal* of common coding theories: to explain a disposition for exhibiting a range of (underlying) computational commonalities in the face of variation in (overall) functional role. The claim of common coding amounts to saying there is something, namely a common code, that explains why corresponding representational states are disposed to be processed in much the same way by cognitive mechanisms.

Cognitive homology grounds common coding explanations

Although I will propose that cognitive homology can underpin a general theory of representational "codes" or computational "formats" in this way, let us see how it applies here. A cognitive homology account follows the strategy of providing a way of guaranteeing the requisite similar underlying computational properties without referring to a description of them. Does cognitive homology have the resources to explain the way in which corresponding imagination and nonimagination representations share a range of fine grained causal similarities (i.e., commonalities in the structure of inferences and particular effects on affective systems, temporal properties for motor systems, etc.) in a way that is clearly distinct from functional role of the given cognitive attitude (i.e., its dispositions to interact with certain cognitive mechanisms *simpliciter*)? Conceptually speaking, the answer is yes, because cognitive homology and homology more generally is clearly distinguished from functional role, and homologues tend to share extensive underlying causal and computational commonalities. It provides a 'truthmaker' for the disposition/tendency to be processed in much the same way, in general, because the natural kind status of

[27] "For any mechanism that takes input from both the pretense box and the belief box, the pretense representation *p* will be processed much the same way as the belief representation *p*. I will count any theory that makes this claim as a 'single code' theory" (Nichols, 2004, 461).

cognitive homologues means that causal powers are projectable between serially homologous activities.[28]

The empirical claim of cognitive homology in this case is that, for corresponding tokened imagination and nonimagination representational states, there are relevant cognitive homologies. The evidence from conceptual processing in grounded cognition (Barsalou, 2008) and examples of neural reuse (Martin, 2007; Anderson, 2010) in Chapter 2 supports the claim of a widespread pattern of homologous activities among corresponding imagination and nonimagination activities. It is in virtue of these homologous activities that one can explain the way in which these representations are disposed behave similarly when interacting with cognitive mechanisms.

In addition to being adequate, a homology explanation seems to have some advantages over accounts that proceed from functional considerations. It is not just that there is currently no theory of what the belief code is or an understanding its nature from a functionalist perspective, but it is unclear whether one could even go about deriving the right kind of underlying causal commonalities in a "top-down" manner from principles akin to logical form or the way of determining representational properties. In contrast, the strategy for identifying cognitive homology proceeds from the "bottom-up" through a host of available empirical considerations (outlined in Chapter 2), and is informed by theories about cognitive evolution and development (e.g., theories of neural reuse). When one has established that the relevant relations of homology obtain, one gets a disposition for similar processing dynamics relatively cheap. The related issue of the scope of such an explanation highlights this advantage. Perhaps one wants to also explain, say, why adopting a particular body position improves the speed at which one can imagine rotating a body part in this same position in mental rotation task, and perhaps one thinks that this should be explained by being in a common code. A homology explanation can naturally extend to similarities over a very large range of properties and in this way be more detailed, extending to virtually any property of the homologue.

[28] Of course, since homologues can vary in their parts, causal properties, etc., which *particular* similarities will hold depends on the details of the case. But we are not trying to say which particular similarities will hold for particular activities, but rather that there tends to be a range extensive underlying computational similarities.

Representational codes as homology classes

Having outlined how cognitive homology could go about explaining what a common coding theory sets out to do in the context the imagination, I now turn to the more general question of whether cognitive homology can provide a base for developing a more general theory of "codes" or representational "formats." As coding has been discussed in a wide variety of domains in psychology, a 'cognitive homology' account of representational formats applies to this wider literature.[29] According to this hypothesis, psychological codes or formats are *cognitive homology classes* (sets of cognitive homologues; homologous cognitive activities). I argue in this section that this satisfies a range of desiderata for a coding theory. In particular, it is *consistent* with the way codes/formats are treated in psychology, and has a positive contribution to the understanding nature of these codes (more than merely being consistent with use). In closing, I show how it can provide an account of what it at stake in extant debates over representational codes/formats by raising one such debate.

In order to even qualify as a legitimate proposal, a theory of codes should provide an account of the identity of these codes (what kinds of things they are as well as what distinguishes between codes), and be able to serve the explanatory goal to which common coding has been put to use (e.g., in explaining the disposition for being processed in much the same way by cognitive mechanisms). At this point it should be clear that cognitive homology is up to those tasks. If representational codes or formats are homology classes, then representational formats are homeostatic property clusters with historical origins, and what distinguishes between codes is them being able to vary largely independent from each other. The question is whether this describes representational formats and whether it can also account for other desirable features of representational formats.

Goldman and de Vignemont (2009) sketch some properties of codes, identifying three: (1) codes or formats have something to do with the typical

[29] It need not apply to *every* way in which the notion of a 'code' is used in psychology. If there is more than one concept associated with "code" or "format" locutions, my account only is meant to apply to one. However, I do think the desiderata correspond to the most prominent way in which codes are discussed in psychology and philosophy.

contents of a tokened representation, (2) the neural systems underlying tokened representations, yet (3) they do not necessarily exclusively code for particular types of referents. One can code representations in systems associated with different sensory modalities: there is presumably a visual code, an auditory code, a verbal code, a finger-representational code, and so on. These codes are associated with the typical contents they bear: one tends to represent sounds in an auditory code, visual forms in a visual code, actions in a motor code, fingers in a finger-representational code and so on, and these in turn have something to do with the neural systems they are found in. For instance, visual coding has something to do with the neural systems that process visual information such as the primary and secondary visual cortex, rather than the primary or secondary auditory cortices. Likewise, being in a finger-representational code has something to do with the activity of the relevant region of the somatosensory cortex. Yet one can represent one and the same thing in a variety of modal formats. One can imagine moving one's arm up and down, and this may take the form of (merely) supposing that one's arm is moving up and down, visual imagery of seeing the motion of one's arm (a sensation-like form of imagination), or motor imagery (with a quasi-proprioceptive component in experience). One can code representations of numbers in any of the four formats: the Arabic code, analogical code, verbal code, and finger-representational code. In each case, one is still representing the same thing—the motion of one's arm, or a number—but coding these in different representational systems/formats.

Here is an empirical example of an alternative format being used to represent spatial relationships during a mental imagery task: the case of "blind visual imagery" (Zeman et al., 2010). Zeman et al. describe a case where "MX" suffered a subtle ischemic event during coronary angioplasty. MX subsequently reported being unable to form any visual images whatsoever and this claim was corroborated by a number of other sources of evidence (such as standardized imagery tests), yet his performance on visuospatial tasks that typically require visual imagery was normal for his age, and visual perception itself was unaffected. Zeman et al. suggest that MX had adopted a verbal strategy rather than using visuospatial imagery for the task, which was consistent with a number of behavioral tests (such as articulatory suppression—repeating words

out loud—severely disrupting performance) and fMRI results that indicated that verbal but not visual neural systems occurred in visual imagery tasks. By coding the representation of objects' geometric properties *verbally* rather than *visually* MX was able to match them with rotated objects in the visual imagery task.

Cognitive homology can accommodate these features. First, it is *consistent* with the way in which codes have something to do with their "typical" contents and underlying neural systems by bringing evolutionary and developmental considerations to the table. For instance, neural reuse theories hold that more recent cognitive activities develop later than evolutionary older cognitive activities, and the former redeploy the neural systems of the latter in development. The reason why representations of numbers are, for instance, coded in a finger-representational code rather than it being the other way around (fingers representations coded in a number format), is due to evolutionary and developmental considerations about the later origin of numerical representation than finger representation Because number representations redeploy the evolutionarily and developmentally prior cognitive resources finger gnosis, it is the case that the code is a finger-representational code that is redeployed for representation of numbers in number cognition (Di Luca and Pesenti, 2011).

Consistency with the way codes are usually described is what one wants in providing an account of what codes are. However, thinking of codes as homologues suggests that this way of referring to codes may be modified. Although evolutionary and developmental considerations provide a way of giving codes namesakes (as visual, auditory, and such), this way of describing formats may need to eventually replaced by a description of codes in a domain-independent vocabulary. Instead of conceptualizing number cognition to be coded in a finger representational format, both number and finger cognition are coded in the format of an "array of pointers" (Penner-Wilger and Anderson, 2008; Anderson and Penner-Wilger, 2012). The positive contribution of homology thinking here is that codes are best conceived of in terms of what they, by themselves, do or are capable of doing, rather than being what they were originally used for (the selected effect function, or the functional role earlier in development).

Secondly, cognitive homology provides a construal of how codes have something to do with the underlying neural systems. In particular, the relationship is that codes map on to the *activities* of these neural systems. For each distinct code, there will be one cognitive homolog (or one class of homologous activities). If there is more than one code in one and the same neural substrate, this indicates that there will be two (non-homologous) activity-functions performed by this neural substrate. Construing codes as cognitive homologs indicates that the relationship between codes and cognitive activity homologues is one-to-one, but the relationship between codes and neural substrates are many to one. This is a positive contribution of homology thinking in this context, as it clarifies the manner in which codes relate to underlying neural substrates.

Finally, an account of representational codes in terms of cognitive homology accommodates the idea that codes are independent of what they code for. Being in one representational state as opposed to another presumably depends on functional considerations; when fingers and numbers are coded in a finger-representational code, this has to do with the way these states behave in the overall cognitive system. A representational state can include more than just reference to a finger or a number; the state may be, say, 'the number of fingers on this hand is four'. In this case, numbers and fingers may be coded in a finger-representational code whereas the other parts are coded in different systems (the 'on' relationship in some other system, etc.). What makes the two commonly coded representations of fingers and numbers distinct is the role these activities are playing in the overall cognitive process of deciding what number of fingers are on one's hand. Homology thinking is quite consistent with the idea that codes and what they code for are independent features.

Codes interact with each other and can be composed of further codes (they exhibit spatiotemporal hierarchy). For an example of how such codes may interact, consider the thesis proposed in *Simulating Minds*, where Goldman argues for "a proprietary code, the *introspective code*, [which] is used to represent types of mental categories and to classify mental-state tokens in terms of those categories [such as representing a visual mental state *as visual*]" (2006b, 260). Goldman's introspective format represents mental states in terms of the latter's

modality, but the mental states that are represented by the introspective code are themselves in their own representational formats: an auditorily coded mental state is in its own auditory code, and this representation is coded *as* auditory in an introspective coding format. In terms of spatiotemporal hierarchy, Goldman suggests that levels of processing within each modality has its own code; "some modalities—certainly vision—have multiple levels of processing, each with its own code or format" (2012). Exemplary of this is Goodale and Milner's (1992) two visual streams hypothesis, according to which there is a dorsal visual stream involved in action guidance and coordination, and a ventral stream involved in conscious perception and perception for action. This two streams hypothesis suggests that objects' orientations, position, shape and other visuospatial properties are coded in two distinct visual formats, and within these systems there are further codes.[30]

Here is how cognitive homology is consistent with these two features. First, obviously cognitive activities can interact with each other, and so even though giving an account of something like the introspective code is far from easy, it is no conceptual difficulty for it to be one that interacts with modality specific representational systems. For the second consideration, according to which codes exhibit spatiotemporal hierarchy, consider the whole of the ventral stream as an activity in normal visual perception. This system starts in the primary visual cortex and proceeds through the ventral stream, through secondary visual systems, color processing systems through more anterior areas such as the so-called fusiform face area and parahippocampal place area. Within this spatiotemporally extended activity of the ventral stream, it can be further decomposed into such component activities that are their own representational formats (e.g., fusiform face area activity), which correspond to different spatiotemporal stages of the ventral stream's activity.

Representational codes are often treated as internal organizations of an activity. In cognitive neuroscience, representations are often said to be coded *in*

[30] Goldman seems to be thinking that the dorsal and ventral stream are part of one and the same visual format. It is unclear whether this is true, at least on a homology construal. One would have to look at the evolution and development of the two visual streams, and similarly for the analogous two streams in somatosensory and auditory systems, in order to determine how these streams relate to each other and to corresponding streams in other systems.

specific anatomical structures. Consider the work of Jabbi et al. (2008) on disgust processing. Jabbi et al. had people process disgust by experiencing disgust (via an odor), imagining disgust (by reading a story) and by seeing someone react to something disgusting (in a video), and they used conjunction analysis to find common regions of activity. Although depending on which modality was tested (imagining, perceiving or feeling disgust) there was an overall difference in effective connectivity with the rest of the brain, Jabbi et al. found that part of the anterior insula (the anterior insula and adjacent frontal operculum; IFO) is redeployed in all three modalities. They conclude that this "suggests that the IFO is a convergence zone where bodily feeling states relevant for the emotion of disgust are coded according to a common code, regardless of stimulus modality." A theory of representational codes should allow that a common underlying neuronal system is evidence for common coding, and cognitive homology does (as a special quality, to be sure). However, this is not all, for although codes have something to do with the structures that code for them, it is more precisely the internal spatiotemporal properties of the system that constitutes a representational code. When disgust is coded according to a common code in the IFO regardless of the overall cognitive context, as Jabbi et al. suggested, this code refers to what is in common in each case, which is to say, it refers to the particular intrinsic or internal characteristic of this region's activity rather than its role in a larger system of organization (its functional connectivity with the rest of the brain).

Thinking of codes as cognitive homologs also points to a more nuanced understanding of such codes, in particular variation in codes. As cognitive homologs can vary in their internal spatiotemporal structure and spatial position, one would want to resist identifying codes as particular spatiotemporal structures such as the IFO. In Chapter 2 a neuronal architecture was presented for predictive coding and mismatch negativity (MMN). Although the neuronal architecture reviewed was described for a particular sensory system (i.e., neuronal structure for *auditory* predictive coding), Wacongne et al. (2012) further argued that the closely similar neuronal architectures of cortical layers in a number of other modalities (i.e., visual, olfactive, somatosensory, association/crossmodal) may also perform predictive coding, which would

explain a wide range of similar computational and other causal properties found among these distinct modalities (i.e., "much beyond the specific domain of the MMN for which it was presently tested" (3676)). What is identical between such cases is not the neuronal architecture itself (as this may vary somewhat among sensory systems), but the activity of "predictive coding" which can exhibit variation in its neuronal architecture. The positive contribution of cognitive homology thinking here is that it holds that cognitive activities are *constituted* by particular spatiotemporal structures of activities without being *identical* to them.

Desiderata	Example for desiderata	Consistency	Illustration	Positive contribution
Relate to typical contents	Finger-representational format	Evolutionary and developmental considerations underlie typicality / namesake	Number representations coded in a finger-representational format, not vice versa	Codes may be better characterized in terms of their internal structure independent of original functional role.
Coupled with neural substrates	Finger-representational format in somatosensory cortex	Codes map on to activity-functions	Same activity for both number & finger components of representational state	Codes map 1-1 to activity-functions, but many-to-one for neural substrates
Independent of representational content	'The number of fingers on my hand is four'	Representational content depends on use-function, not activity-function	Same format for representational content of numbers and fingers, different functional role	*
Causal relata	Introspective code	Activities are causal relata	Introspective code takes activity from other formats as input	*
Spatiotemporal hierarchy	Ventral visual stream	Spatiotemporal parts of activities correspond to further representational formats	Fusiform face area activity as one component activity of ventral visual stream	*
Internal structure	Disgust in IFO & Predictive coding	Internal activity rather than functional role is what constitutes the code	Structure of IFO activity & Structure of predictive coding	Cognitive homologues are *constituted* by internal organization, but not *identical* to them

Figure 3.5

Overview of desiderata, an example that was discussed, the consistency claim and an illustration as it applies to the example, and the positive contributions from homology thinking (if applicable).

In summary of the above, cognitive homology seems quite adequate as an account of computational formats. It provides an account of the identity conditions for codes, and relations of cognitive homology support the kinds of inferences common coding is invoked to support. Moreover, it can accommodate

the other desiderata surveyed and in many cases contribute to a better understanding of these desiderata (Figure 3.5).

Cognitive homology can explain how codes are namesakes and what they have to do with 'typical' contents and underlying neural systems through evolutionary and developmental considerations of priority, though suggests that these may be the wrong way to characterize codes. It explains the relationship with neural substrates: the mapping is many-to-one, but one-to-one with activity-functions of various neural substrates. It allows for the codes to be understood so as to be independent of representational content, as the latter relate to functional roles. It also allows for causal interaction between codes and for such formats to exhibit spatiotemporal hierarchy. Finally, codes are constituted by their internal spatiotemporal makeup, but one can also accommodate variation in the internal spatiotemporal properties of a given code as codes are constituted by, but not identical to, their internal structures. Overall, I consider this to be a fairly strong case for thinking of codes/formats as homology classes.

One worry one might have about the above is that this all seems to require only an appeal to activity-functions, whereas the *historical* character might seem as if it could be dropped out entirely and replaced with some other, ahistorical essentialist or non-essentialist account of identity of cognitive activities. It is true that the conceptual contribution of homology thinking here does not seem necessary to gain purchase on these desiderata, at least not directly. In the previous section on common coding explanations I endeavored to show how the historical character of cognitive homologues is important for supporting an important explanatory goal: common coding explanations. In this section I focus on other desiderata, so as to provide further reasons for a cognitive homology interpretation of representational formats. The developmental-historical identity conditions for an activity-function are the relevant starting point for developing detailed common coding explanations and support that role, whereas these other desiderata furnish a more detailed account of why codes stand to be viewed as cognitive homologues.

Let's consider an objection. One might be tempted to object to an account of codes that is not in computational terms by appealing to an intuition about

that one would call two non-homologous, but highly similar computational structures the same code. Sameness of *code* refers to sameness of something *computational*, it might be thought, and this is demonstrated where computational similarity and cognitive homology come apart. However, the desiderata for a theory of representational formats does not include these kinds of a priori considerations about what we would call the same thing, any more than a theory that organs are homologues turns on whether one would call two non-homologous organs "wings." In both cases, functional/computational commonalities are what our judgments concern, but the existence of wings as analogues does not count against the fact that these traits *can* also be viewed as homologues (though not a single class of homologous parts), nor does it say that they *should not* be viewed as homologues. The goal is to provide some account of what makes two tokened representations in one and the same code or format—an account that is consistent with and supports good scientific theorizing (and not intuitions), and which can accomplish explanatory goals of attributing sameness of codes, and my argument is that cognitive homology can do this quite well. Individuating codes by their computational structures does not provide an explanation of why they are similar or share these functional/computational structures, as per the explanatory role of identifying sameness of codes, because codes would then be individuated in virtue of the features they are individuated to explain—no better than the case of the "dormitive virtue" mentioned earlier. What this objection has right is that relationships between the computational properties of representational states are often the motivating factor for positing codes, but this alone is does not carry weight independent of its relationship to the explanatory goal of positing a common code.

It is worth highlighting that homology thinking supports precisely the *opposite* intuition from the above; it *is* possible for two codes to differ remarkably in their causal/computational properties while being one and the same code. This just reflects the fact that homologues can vary in their properties, including their developmental mechanisms, components and internal organizations. There is also some precedent for there being representations housed in one and the same code that differ in very significant respects. Goldman suggested that the two

visual streams in the two-streams hypothesis are in one and the same "visual format." If they are the same code, dorsal and visual stream are two character states of one cognitive activity homologue. Assuming dorsal and ventral visual stream processes are homologous activities, here we have a case where there will be quite remarkable computational differences between corresponding representations in the same code. Although a representation of a spatial property such as the position of the object will be in the visual format, a representation of this property in the dorsal and ventral stream will differ in a range of significant ways. In the dorsal visual stream, spatial locations are coded egocentrically (the frame of reference being one's body), whereas ventral stream seems to code spatial locations allocentrically (i.e., the spatial information is among or within objects). Again, it remains to be seen whether the dorsal and ventral stream truly are in a common visual format (whether they are serial homologues), but the point is that they could be, and if so, they could indeed be strikingly different, *as expected* from homology thinking.

The fruitfulness of my account of formats in terms of homology is demonstrated in its applicability for evaluating theories that make claims about how concepts and other representational states are coded. In debates on grounded cognition, the main theoretical issue is whether representations are (sometimes) coded in an "amodal" language format, or whether they are always coded in perceptual and motor systems (Simmons et al., 2007; Barsalou, 2008). Codes also appear in accounts of grounded cognition; Goldman (2012) argued for a picture of embodied cognition according to which much higher cognition takes place in a number of "bodily formats," where bodily formats are those formats that typically process bodily information in perception and action. A cognitive homology account of coding formats helps say what sorts of issues are at stake in these theories, and points to a way forward in determining whether these accounts are correct.

Consider an exchange regarding grounded cognition. Simmons et al. (2007) used fMRI to analyze the relationship between activities involved in perceiving and knowing about color. They compared two tasks: color verification (e.g., whether <'taxi' - 'yellow'> was a congruent pairing) and color perception (i.e., whether color wedges on a color wheel are in order). Their results suggest

that posterior color-selective regions in the occipital cortex support passive color sensation (and representations of "lower level qualia of stimuli"), whereas anterior color-selective regions on the ventral temporal lobe are engaged for active processing of color information, so as to encode higher level color perceptual representations. These are partially overlapping regions. This is used to support the thesis of grounded cognition, that color representations in either case are coded in modal systems for perceiving color.

However, without an account of what codes are it is unclear how or to what extent an overlapping neural substrate implicates a common representational format. Edouard Machery points out that, actually, this result (and other similar cases) can be interpreted as evidence *against* grounded cognition, as the partial overlap suggests that anterior regions are in a different, amodal format:

> [C]ognitive neuroscientists have repeatedly found that the brain areas activated in the tasks meant to tap into higher cognition are near, and thus not identical to, the brain areas involved in perceptual or motor processing (a point acknowledged by Simmons et al., 2007) … Furthermore, the brain areas that are activated in the tasks meant to tap into the processes underlying higher cognition are not only different from the brains areas activated in perceptual processing, they are also anterior to them. A plausible interpretation is that the brain areas activated in the tasks tapping into higher cognition are amodal representations, which are distinct from the perceptual representations activated in the tasks tapping into perceptual processes. (Machery, 2010)

What Machery argues is that, although cortical processing regions involved in higher cognition tend to be *near* to brain areas in perceptual or motor processing, and are anterior to them, these may be instead viewed as amodal representations.

As it stands in these debates, it is not immediately obvious what would ultimately provide the answer. That is, because being a representational format is an unanalyzed notion, it is unclear what the relationship is between various kinds of evidence and the thesis that all representations are coded in modal systems. What is at stake in this particular exchange concerns whether or not these anterior regions are in a same or different code from their posterior counterparts, and my account says that the answer to this question turns on (is equivalent to) the question of whether these are homologous cognitive activities.

That is, it depends on whether or not these activities involve the repetition or duplication of the same cognitive activity.

The contribution from cognitive homology here is that it can provide a basis for bridging empirical data and theses about representational formats, and does so by appealing to the modified operational criteria presented in Chapter 2. One would ask whether these activities exhibit continuity: are anterior-posterior counterparts poorly differentiated in children for tasks that normally selectively activate one over the other? And is there continuity between the range of activities that span these anterior-posterior regions? For spatiotemporal position, is it true that both anterior and posterior activities have high correspondence in their proximal connections to other cognitive activities over a range of tasks? And for special quality, are there shared features such as a distinctive internal organization, type of neural substrate, and the like, that suggest they are homologous activities?[31] By thinking of representational formats as homology classes/homologous activities, one understands what is at stake in determining the sameness of representational formats (relations of serial homology), and this points to the aforementioned operational criteria as specifying the relevant evidence for deciding on the issue.

[31] A particularly difficult issue here is that it seems that these operational criteria point in different directions. Among these more anterior temporal regions, there seems to be a distinctive cytoarchitecture compared to that found in posterior/dorsal regions (Martin, 2007), but these anterior activities also organized in part by the input they receive from their corresponding posterior region and as such exhibit continuity.

Chapter 4: Evolution of the mentalization system

This final chapter focuses on the relationship between our evolutionary history and the architecture of cognitive systems, particularly as it applies to mentalization. The terms 'mindreading' and 'mentalization' (used interchangeably) refer to our capacity to think about other minds, which involves attributing and reasoning about mental states. The particular mental states in question are propositional cognitive attitudes (beliefs, desires, fears, and the like), rather than perceptual states.

This chapter has three sections. First, I clarify what is at stake in competing accounts of the mentalization or "mindreading system," arguing that the issue is providing an account of the core or primary system underlying our mentalization capacities. Second, I sketch two competing accounts of the mentalization system, according to which it is either interpretation-like or is rather like an inner-sense. I show how optimality considerations can be used to provide some evidence for these two accounts, and say how cognitive homology can address considerations that are not included in this (viz., evolvability). Third, I provide an illustration of how to address this component of the evolutionary process and how it may be used to understand the architecture of the mentalization system.

The core mentalization system and its structure

Over the last few decades, debates about the mindreading system have been partitioned into two general camps of theories, dubbed *theory-theory* and *simulation-theory*. Proponents of theory-theory argue that mentalization proceeds by way of inference that employs a folk-psychological theory. Proponents of simulation-based accounts argue that mentalization proceeds by way of mental simulation that uses one's own psychological states, rather than through any sort of folk psychological theory. In contrast to such early debates, most authors nowadays believe a hybrid theory is correct; the mentalization system has both simulative and folk-psychological elements in it. So in the following I will not discuss simulation-theory and theory-theory *per se*, but rather present in general terms two accounts that are consistent with hybrid theories as found in contemporary discussions of the mentalization architecture (Carruthers, 2009,

2011; Goldman, 2006b). These two accounts make claims about the architecture of the mentalization system by saying it resembles an inner-sense (on analogy with our access to perceptual states) or whether it is an interpretive system (which does not resemble our access to perceptual states).

Before discussing particular accounts of the mentalization system, I first would like to clarify what is at stake in such an account. This can be made in terms of the range of phenomena that count as proper parts of the mentalization system. In theorizing the architecture of the mentalization system, the aim is to give an account of the structure and components of the mindreading system, but what range of phenomena need to be explained by this mentalization system? A functionalist approach would say that mindreading occurs whenever one's thoughts are about mental states (tokened representations of mental states), and would count any instance of mental state attribution as one example of mindreading that needs to be explainable in terms of the architecture of the mentalization system. In contrast, a non-functionalist would use other criteria to determine what counts as a proper part of the mindreading system, which does not necessarily extend to every instance of there being tokened representations of mental states. I argue that the approach that is standardly taken is not functionalist and that it should not in any case be viewed as such.

That the architecture of the mentalization system is not standardly taken to be one that accounts for all cases of mental state attribution is reviewed in detail by Theodore Bach (2011). Bach points out how many authors over the last decade have explicitly distanced themselves from their theory being one that aims to explain all instances of mental state attribution. Instead, what is at issue is giving an account of a core mentalization mechanism. For instance, in a representative example Alvin Goldman says "I have no hesitation in agreeing that inductively based attributions sometimes occur, with no simulative ingredient at all … [there is an] obvious need to refrain from claiming that simulation is involved in all cases of third-person attribution" (2006b, 89). Goldman is committed to the core mentalization system being simulative, rather than all instances of mental state attribution. Hence, there are cases of mental state attribution that are not performed by the mentalization system. What criteria are used to identify the core mentalization system? There need be no

precise answer, but the thought is that the core mentalization system will be found to be in play for most ecologically valid contexts (the bulk of everyday mentalization), and it is a system that has particular developmental mechanisms that have been selected for supporting reasoning about mental states – it is tied to the developmental literature on mindreading, such that the mentalization system should develop consistently and fairly early in childhood.

It is also undesirable to count all cases of mental state attribution as part of one's range of evidence to be accounted for by one's mentalization system architecture. Mirror neurons are thought to be a component of the mentalization system, as they are neural populations that code for observed and executed actions (Gallese and Goldman, 1998). But are they part of the mentalization system, or just something that the mentalization system interacts with? A functionalist conception of mentalization would appeal to whether mirror neurons are necessary for mentalization by seeing whether they are always used in mentalization. For instance, Shannon Spaulding points out that mirror neurons are not necessary for mindreading because "subjects with damage to mirror neuron areas for fear retain the ability to attribute fear based on body language, semantic knowledge, and situations that typically evoke fear" (Spaulding forthcoming). As the language makes clear, the mindreading system in question is supposed to be responsible for any instance of the "ability to attribute fear," and if mirror neurons do not always underlie this ability, they are not necessary for mindreading. Consequently, although they may play a role in mentalization sometimes, they are on comparable ground with semantic knowledge.

However, we might do well to wonder what such an architecture would look like, if any mental state attribution counts as an instance of the mindreading system. If semantic knowledge, body language interpretation and general associative knowledge are all sufficient for mentalization, what are the necessary components of the mentalization system? Presumably, representations of mental states can be coded in a wide variety of systems, including semantic knowledge, procedural memory, visuospatial associative knowledge, and so on. If all of these cases tell us about the mindreading system, it is hard to see how the mindreading system will look much different from a theory of how we attribute

and reason about things more generally, with mental states being just a special case of the referents of our thoughts. At least, it is difficult to see how such an account would not end up looking like a disjunction of systems and activities that are individually sufficient for mental state attribution. This is undesirable because the increased scope of one's mentalization architecture comes at the cost of being able to less easily account for any distinctive mentalization system associated with ecological, evolutionary and developmental considerations.

In summary, what is at stake in competing accounts of the mindreading system (in cognitive science) is not a general architecture that accounts for all cases of mental state attribution but rather a specific system with an evolutionary history, as described by developmental psychologists, and which represents the bulk of our everyday mentalization activities. This is desirable anyways, since defining the mindreading system in terms of all cases of mental state attribution would end up describing a general system for reasoning about nearly anything, with mental states as a special case of referent for one's thoughts.

In the following, I refer to the mentalization system as the "mentalization mechanism." I do this because referring to it as a mechanism accurately captures the explanatory project as being one of *mechanistic explanation*. The goal is to identify the *core* mentalization mechanism/system. In accordance with mechanistic explanation, the mindreading system explains the capacities of a system as a whole (in supporting cognition about mental states) in terms of its components and their organization.

Evolutionary and developmental perspectives on mentalization

Evolutionary considerations are often used to support one theory of the mentalization system over another. Two accounts of the structural features of this mentalization system have been prominent in recent debates. *Interpretive sensory access* (ISA) accounts hold that our access to our own cognitive attitudes is indirect, and the mechanism operates by providing a sort of "best guess" based on the available sensory evidence (some of which might be simulated sensory states). ISA accounts model our self-knowledge process on folk psychological theories, where sensory information is taken as input and a body of knowledge is deployed in attributing a mental state. In contrast, according to *inner-sense*

theories, the access that humans have to their own cognitive attitudes is both direct and introspective. It is direct and introspective in so far as it is *not* mediated by an inferential processes such as through a folk psychological theory. Instead of being modeled on the operation of a folk psychological theory, our access to our own beliefs, desires and other cognitive attitudes is modeled after our access to our own perceptual states (such as whether one is seeing an image as an image of a duck or a rabbit). Inner-sense accounts argue that our mentalization mechanism is perception-like, whereas interpretive access accounts argue that our mentalization mechanism is interpretation-like. A recent inner-sense account can be found in Goldman (2006b),[32] and one version of an ISA account is defended at length by Peter Carruthers (2009, 2011).

One source of evidence drawn on in deciding between these accounts comes from the evolutionary history of our mentalization system (Carruthers, 2009). In particular, the *order* of the selection history for successful *other*-mentalization and *self*-mentalization is used to provide evidence for one account over the other. The order of selection history matters because the core mentalization mechanism that evolved did so under the initial selection pressures, and through exaptation it is redeployed, largely in its original form, so that it is used for mental state attribution to both oneself and to others. The mentalization system, in other words, is optimized for what it was originally selected for, and so by appealing to the optimal system given the initial selection pressures one can defend either inner-sense or interpretive access theory of the core mentalization system.

Consider first the case where the selection pressures for self-mentalization precede those for other-mentalization. This might have occurred if self-mentalization proves advantageous in metacognition, for planning and reasoning about uncertainty and one's beliefs, and the like, and does so before there is any advantage to mentalization of others. From this order of selection pressures, we might expect that the evolved mechanism is much like an inner-sense. The expectation that an inner-sense mechanism would have evolved is supported because a direct, unmediated inner-sense mechanism would be optimal for monitoring one's cognitive attitudes, unlike a "best guess"

[32] In fact, this is the "introspective code" discussed in Chapter 3. But the details of these architectures are not important for my present purposes.

mechanism. One supporting consideration has to do with the advantages of being sensitive to certain kinds of error that are only made possible through an inner-sense; much like we tend to be introspectively aware when our perceptual states are uncertain or ambiguous (as in poorly lit visual conditions), we would expect to be directly sensitive to certain kinds of error or ambiguity in the mentalization inner-sense. Confabulation and error in self-mentalization would thus not occur in a range of cases that we would expect it to if it were interpretive, as interpretive systems do not provide introspective access to the quality of their input.

Alternatively, the evolutionary pressures may have appeared the other way around, so that the evolutionary pressures for other-mentalization precede those for self-mentalization. Why might the selection order be this way around? Consider the Machiavellian intelligence hypothesis, according to which the capacity for mentalization is selected for its advantages in social competition, such as for tactical deception and to detect deception from others (Byrne, 1988). It may also be that self-attribution and keeping track of one's own propositional attitudes only becomes advantageous after a certain level of sophistication in others' mentalization capacities, i.e., when others are already routinely attributing mental states to oneself (Carruthers, 2009).[33] Together these would suggest that the selection pressures start with other-mentalization.

An interpretation system would be optimal in this latter case because an inner-sense would be less efficient: "because neural connections are costly to build and maintain, some distinct evolutionary pressure will be needed to explain why the metarepresentational [=mentalization] faculty should have acquired the input channels necessary to monitor the subject's own propositional attitudes" (Carruthers, 2009, 128), and in contrast an interpretive mechanism does not require such unused, costly components. A second consideration has to do with performance in competitive situations, the original selective context as per the Machiavellian hypothesis. Inner-sense theories arguably have drawbacks here, as mentalization of others through an inner-sense mechanism would involve a sort of simulate-then-introspect architecture, and negative affect mental

[33] Accordingly, other systems are able function as performance monitors, which are sensitive to uncertainty and other features, but these are all done by first-order processes rather than ones that concern mental states as such.

states attributed to the other would likely 'leak' into one's own motivational systems due to simulation of these affective states. For example, if I simulate the perceptual states for disappointment in order to inner-sense the attitude and attribute it to another, such a feeling would temporarily interfere with my own affective and motivational systems. An inner-sense mechanism would be optimal if the selection pressures start with self-mentalization, but not if they start with other-mentalization, and vice versa for interpretive access mechanisms (summarized in Figure 4.1).

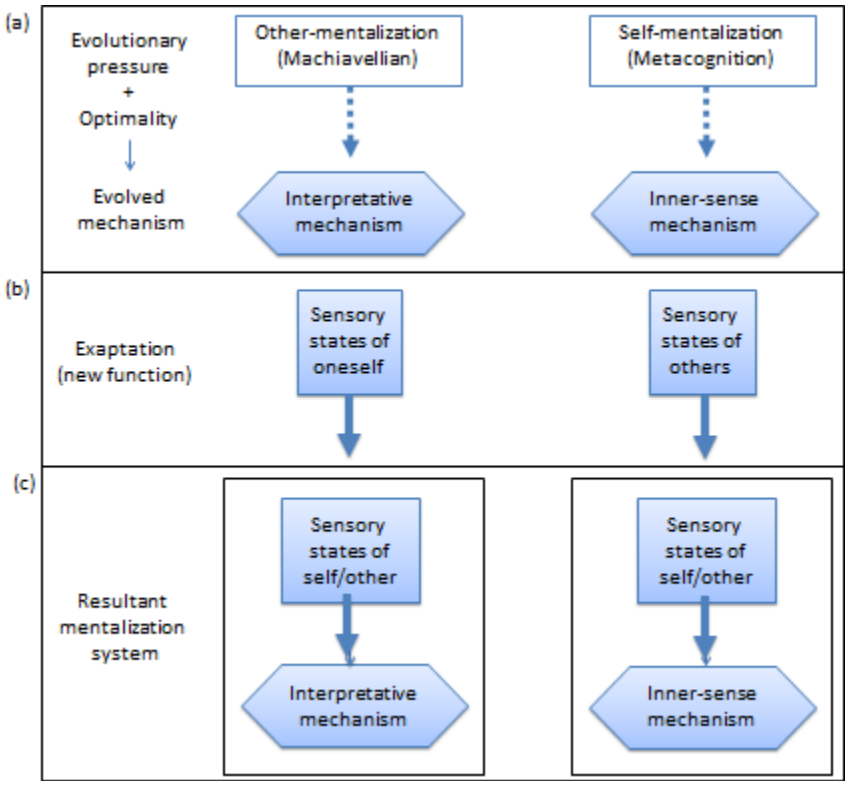

Figure 4.1

Summary of evolutionary arguments concerning the mentalization architecture. (a) evolutionary pressures and considerations of optimality are used to predict the evolution of one system over the other; (b) the evolved mechanism takes on a new input so as to have a role in mentalization of both self and other (exaptation); (c) the resultant mentalization system for both self- and other-mentalization employs a mechanism similar to that predicted by the initial evolutionary pressures.

Optimality considerations can go some way in providing evidence for the architecture of the mentalization system, but the strength of these considerations

depend on certain prior assumptions about the potential variation in ancestral populations. In order for selection pressures to support the above predictions about the architecture for the mentalization mechanism, one must assume that one mechanism could have been selected over the other by these selection pressures. And obviously, a precondition for selection of heritable variation in a population is the generation of such variation in a population in the first instance. So both mechanisms must be within the range of phenotypic variation in order for selection pressures to have selected for one over the other. This is an assumption about the potential for generation of heritable variation (concerning *evolvability*), which is a prior step in the evolutionary process to the selection for these mechanisms (adaptation).

Two stages of evolution

1. The generation of heritable phenotypic variation in a population (*evolvability*)
2. Selection of heritable variation in the population (*adaptation*)

The current line of reasoning sketched above focuses on step 2, and is sound only if step 1 (the generation of heritable variation) allows for both mechanisms to developmentally originate in the population (Amundson, 1994; Brigandt, forthcoming). Otherwise, optimality does not tell us anything about the structure of the mentalization mechanism, because the sub-optimal mechanism would have been selected for all the same.

Developmental considerations are decisive here (Hendrikse et al., 2007; Wagner, 2000; Müller and Wagner, 2003; Gerhart and Kirschner, 2003). The particular mode of development of a species determines what phenotypes (including cognitive structures) result from the genotype, and thus what phenotypes can result from genetic changes. Novel structures originate in evolution from modifications of the species' mode of development (that produces the structure). The existence of relevant developmental constraints would prevent optimality considerations from being a deciding factor in the overall structure of the core mentalization system. Developmental constraints in this sense are not so much constraints on an individual's development, but rather constraints on transgenerational variation and morphological evolution due to development:

> A developmental constraint is a bias on the production of variant phenotypes or a limitation on phenotypic variability caused by the structure, character, composition, or dynamics of the developmental system. (Maynard Smith et al., 1985, 266)

To illustrate, suppose that *self*-mentalization is what mindreading mechanism is selected for accomplishing in the first instance, but that an inner-sense mechanism cannot be generated due to the internal causal structure of the developing organism. In such a case, an otherwise suboptimal mechanism of interpretive-access which *is* present would be selected for all the same. Only if both mechanisms are present within the overall species can one say that optimality decides between which of the two wins out over the course of adaptation. But is the assumption that both architectures could be exposed to selective pressures justified? A fully satisfactory answer to this would draw on the causal-developmental structure of the cognitive and neural systems in place prior to the evolution of the mentalization mechanism in order to determine what phenotypes (that can be used for attributing mental states) can and are likely to develop.

A positive answer to this question would identify two such mechanisms (one of which is inner-sense like, one of which is interpretation-like in its architecture), where each could be used for mentalization. A negative answer to the question would be more difficult, as requires arguing that there are no other developmentally plausible mechanisms. Establishing either answer with much certainty is difficult as one would want to consider multiple organismal levels in order to have a fully adequate account. My aim in the next section is to use the default network in understanding the evolution of the mentalization system. The standard that is operative in making empirical claims about the default network and mentalization system is plausibility, rather than aiming at a comprehensive synthesis.

The mentalization system is homologous to the default network

Recall the default network and how it is used in mentalization tasks. The default network comprises a number of temporal, parietal and frontal regions and contains two sub-systems, which in Chapter 2 I suggested might be serially

homologous default network activities. These subsystems are associated with a number of tasks (Figure 4.2)

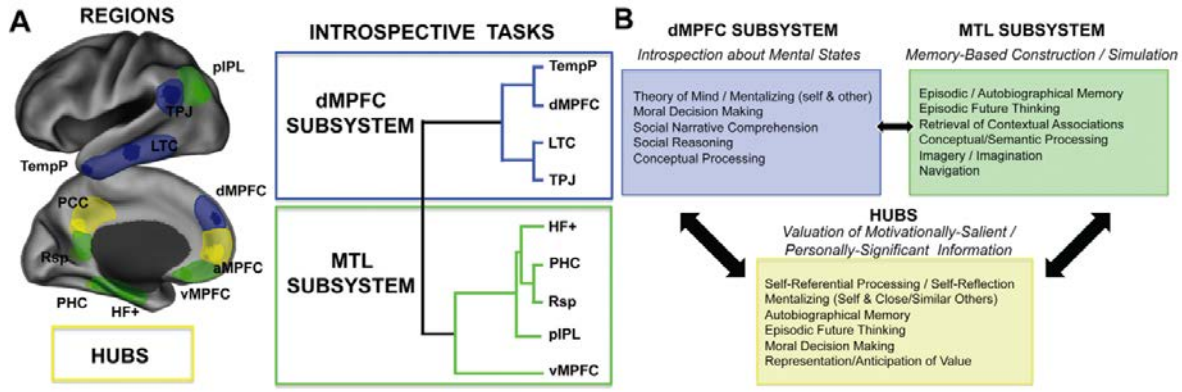

Figure 4.2

(a) Two subsystems of the default mode network, comprising of the dorsal medial prefrontal cortex system and a medial temporal lobe subsystem, as determined by connectivity in introspective tasks; (b) Summary of tasks that these subsystems and hubs play an important functional role in. (See Andrews-Hanna, 2012, for a detailed discussion.)

In particular, the dMPFC subsystem is engaged in moral decision making, social narrative comprehension and reasoning, and mentalization (especially of distant others). In contrast, the MTL subsystem is engaged in episodic memory and future thinking, imagery and spatial navigation. Each system is strongly coupled with the hubs but are not strongly coupled with each other. Two relevant distinctions are often made in the mentalization literature in psychology: (1) between the 'classical' mentalization system and mirror neuron systems, and (2) between cognitive and affective mentalization.

The first distinction is between mirror neuron systems and the "classical" mentalization system. For instance, Teufel et al. (2010) distinguish between these systems as follows:

> Any simplified generalization will cut across categories, but two broad distinctions can be discerned from the literature. The classical areas associated with ToM involve a distributed network including the [whole] mPFC and the TPJ. This system is related to visual perspective taking, higher-order belief reasoning and the attribution of representational mental states (perceptions, knowledge, beliefs). …

> More recently, the mirror system has been discussed as a second component of the ToM system, subserving a more implicit understanding of other people. It spans a wide range of cortical areas including the insula for the understanding of others' emotions, the secondary somatosensory cortex for the understanding of bodily sensations, and the premotor and parietal areas for the understanding of actions. A characteristic of all of these areas is that they have shared circuits for processing of one's own and other people's emotions, bodily states and actions. (Teufel et al., 2010, 337; cf. Van Overwalle and Baetens, 2009)

This first distinction between a mirror system and a classical mentalization system has also been proposed in a recent comprehensive meta analysis of mentalization (Van Overwalle and Baetens, 2009). However, Van Overwalle and Baetens (2009) also stress the importance of the precuneus in the mirroring system (which refers to a region overlapping with PCC/Rsp in the MTL subsystem as in Figure 4.2). Mirroring is normally associated with mirror neurons specifically in the frontal cortex related to action planning and execution, comprising a "fronto-parietal mirror network" (Rizzolatti and Sinigaglia, 2010). Yet the conception of a mirror system present here is not restricted to a fronto-parietal mirror system but also includes sensory circuits relevant to processing information concerning one's own and others' mental states. However, recall the structure of the connectivity among nodes within the precuneus/PCC (Figure 2.4) – this connectivity suggests that adjacent regions within the precuneus are integrated with bodily and visual processing systems along with the fronto-parietal mirror network, providing a second common factor in Teufel et al.'s mirror systems. Interestingly, these mirror neuron systems seem to be *complementary* systems in that one tends to be active in the absence of the other, suggesting that one or the other will generally be sufficient for accomplishing mentalization task (Van Overwalle and Baetens, 2009).

How do these two systems relate to the subsystems of the default network? One plausible suggestion is that they correspond to each other, in the following way. The mirror neuron system is accessible through the MTL-subsystem activity of the default network (as seems to be the case with mental imagery and episodic thinking). In contrast, the 'classical' mentalization system corresponds to the hubs and dMPFC subsystems. However, this classical mentalization system can be further subdivided into systems for mentalizing of close/similar others and distant others. In particular, the dMPFC subsystem

(which is involved in narrative comprehension and moral decision making) is involved in mentalization of distant others, whereas the hubs are central to mentalization of close/similar others. The idea behind this distinction is that self-referential processing and mentalization of close/similar others use the same simulative processes accessed via the MTL-subsystem, whereas mentalization of distant others relies on processes involved in social reasoning and narrative comprehension.

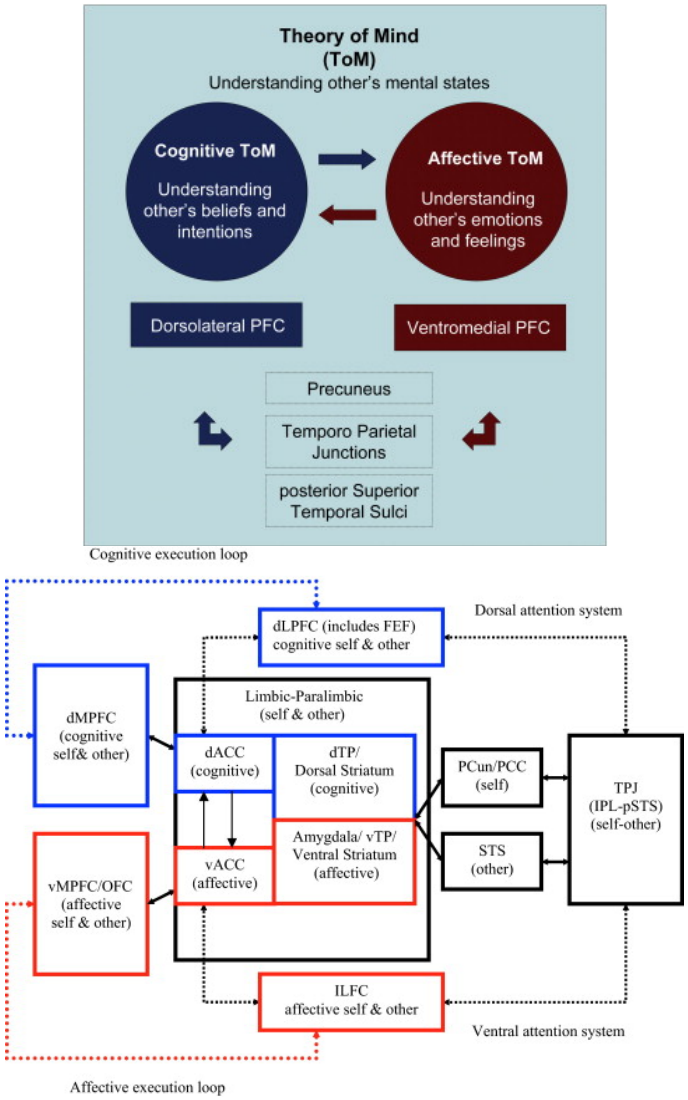

Figure 4.3

Cognitive and affective mentalization systems. Left: Summary of the two systems for understanding others' mental states (Poletti et al. forthcoming). Right: An architecture for affective and cognitive mentalization for self- and other-mentalization (Abu-Akel and Shamay-Tsoory, 2011).

The second distinction is between cognitive and affective mentalization (Figure 4.3), so called because different systems tend to process cognitive mental states (beliefs and intentions) and affective mental states (such as emotions). Central to the distinction between cognitive and affective mentalization is activities along the dorsal-ventral axis of brain regions, such that dorsal and ventral components of various anatomical regions are associated with cognitive and affective mentalization respectively (Figure 4.3, right). *Dorsal* medial prefrontal cortex, anterior cingulate, striatum and lateral prefrontal cortex are involved in cognitive mentalization, whereas *ventral* medial prefrontal cortex, anterior cingulate, striatum and inferolateral frontal cortex are involved in affective mentalization. Moreover, the precuneus/pCC along with the inferior parietal lobule (of the tempoparietal junction) are associated with self-mentalization, whereas the superior temporal sulcus along with *posterior* superior temporal sulcus (of the tempoparietal junction) are involved in other-mentalization.

Affective and cognitive mentalization relate to the subdivisions of the classical mentalization system. In particular, cognitive mentalization mechanisms show considerable correspondence with the hubs and dMPFC subsystem, whereas affective mentalization mechanisms have high correspondence with the hubs and many areas assigned to the MTL subsystem.

Recall that what is at stake in saying what 'the mentalization system' is is an account of the *core* mentalization mechanism. There is a boundary problem in characterizing the core mentalization system because one has to distinguish between core and peripheral cognitive activities involved in mentalization. Because cognitive systems are softly assembled, one expects considerable variation in component cognitive activities assembled for accomplishing a given mentalization task. Aside from the two hubs of the default network, there does not seem to be any set of activities that are always present in standard cases of mentalization. (To be sure, this is suggestive that the core mentalization system shares at least the same two components as the hubs of the default network.) The relevant considerations come from the evolutionary and developmental origins of our mentalizing capacities (and what occurs in ecologically valid/representative forms of mentalizing).

Now let's consider the evolutionary history of cognitive and affective mentalization. It is not implausible that affective mentalization is evolutionarily older than cognitive mentalization, and this seems to be consistent with studies on nonhuman animals (especially nonhuman primates) that are highly responsive to the emotional/affective states of conspecifics, in contrast to beliefs (Preston and de Waal, 2002; de Waal and Ferrari, 2010; Decety and Sveltova, 2011). The high correspondence in the architectures for affective and cognitive mentalization given by Abu-Akel and Shamay-Tsoory (Figure 4.3, right) is highly suggestive that cognitive and affective mentalization activities are serially homologous. In terms of the default network, this means that the MTL subsystem is likely evolutionarily prior to the dMPFC subsystem (an expected consequence, given the tasks associated with the dMPFC subsystem versus the MTL subsystem, e.g. social narrative comprehension versus spatial navigation). Moreover, the duplication of affective mentalization networks in dorsal regions of the anatomical parts of the affective mentalization network is suggestive of the corresponding subsystems of the default network being serially homologous default network systems.

What about developmental considerations? Recall that homologues are homeostatic property clusters that maintain homeostatic integrity through a number of developmental mechanisms. Importantly, these developmental mechanisms include cognitive activities. The cross-cultural neuroscience discussed at the end of Chapter 2 highlighted how cultural practices and one's history of cognitive activity affect the coupling of cognitive activities. The existence of a particular cognitive homologue (such as the parieto-frontal inhibition of emotional processing in Asian but not European American adults) owes its homeostatic integrity to the way in which this activity is facilitated through cultural practices during maturation. To put it somewhat paradoxically, cognitive systems are part of their own developmental mechanisms; they maintain their own homeostatic integrity through their parts being repeatedly deployed in unison.

When these considerations are applied to mentalization, there is a strong case for the default network being its own homeostatic mechanism and as well as that of the mentalization network. The default network is characteristically active

at rest, and whenever there are no sufficiently cognitively demanding tasks, mammals engage in default network activity. This tendency to enter default mode processing whenever possible certainly owe to the existence of a number of developmental mechanisms (as do biases in the content of default mode processing). However, the activity of default mode processing also has the effect of maintaining the homeostatic integrity of itself in its various character states. Additionally, a strikingly large portion of humans' default mode processing revolves around social cognition/mentalization; "human beings have a predisposition for social cognition as the default mode of cognizing which is implemented in the robust pattern of intrinsic brain activity known as the 'default system'" (Schilbach et al., 2008, cf. Schilbach et al., 2012). Consequently, the default network is a major defining component of the developmental mechanisms that maintain the integrity of the mentalization system. In summary, the mentalization system consists in a redeployment of the default network and maintains its homeostatic integrity through this continual redeployment.

The core mentalization system can be understood as consisting in activities that form a homeostatic property cluster that are maintained by the default network. Since this system is also a *redeployment* of the default network, on my account it is also *serially homologous* to the default network; that is to say, the core mentalization system is the same system as the default network. As a result, to give an account of the mentalization system is just to give an account of the default network as it is engaged in mentalization.

Homology and mentalization system's structure

Now to revisit the optimality argument and whether the structure of the mentalization system is interpretation-like, or whether it instead resembles an inner-sense. The optimality argument draws on what kind of structure for a mentalization system would be optimal given the initial selection pressures (for either self-mentalization or other-mentalization), and says that it will resemble this initial system even after it is used for mentalization of both oneself and others. This depends on there being two systems (an inner sense and interpretive system) that can develop so as to be subsequently acted upon by natural selection, such that the optimal system will win out over the course of evolution.

(It also depends on these two systems being redundant, so that there are costs with maintaining both systems.)

It is difficult to prove that there is only one system that could arise much more readily, but I have not argued that this is so—rather, the optimality argument depends on there being two such mechanisms that can develop in order for selection to favor the optimal one. What I *have* argued is that one candidate system is the default network, which already existed prior to the evolution of mentalization capacities. The optimality argument says that the system that is present prior to a general mentalization system for both self- and other-mentalization will likely be highly similar to its present state in humans, and a similar consideration (but with a focus on development) also applies here: the structure of the mentalization system will likely be highly similar to that of the default network as it is engaged in facilitating similar tasks. So one can use the structure of the default network to infer the likely structure of the mentalization system, and that is to say the question of whether the mentalization system is interpretation-like or like an inner-sense depends on whether the default network can be described as such.

Yet the inference is not altogether straightforward. For what is the structure of the default network? The default network, as a cognitive homologue, can take on various character states, and these character states can be organized differently. For instance, it could be that the default network is organized as an inner-sense mechanism in the mentalization character state but in every other character state it is organized as an interpretive mechanism. The structure of the mentalization system depends on the organization of the character state the default network has in mentalization.

Importantly, there may be more than one mentalization-facilitating character state for the default network. In particular, there are at least two character states for the default network, one corresponding to each subsystem of the default network. Both of these subsystems are engaged in mentalization, depending on the kind of mentalization task at hand. Mentalization of close/similar others employs the character state of the MTL subsystem (which modulates sensory and motor systems). Mentalization of distant others, in contrast, corresponds to the character state of the dMPFC subsystem (which is

also used in social narrative comprehension and social reasoning). These systems are dubbed affective and cognitive mentalization systems, respectively. If these two subsystems correspond to the two character states for the default network and are used in mentalization, the varieties of architectures for the mentalization system can be inferred from the structure of these character states.

Now to briefly consider whether these character states resemble an interpretive system or an inner-sense. By way of illustration, the simplest way to go about this is by analogy to the structure of other cognitive mechanisms in the same character state. The architecture itself is what really matters here, but one would need to describe in detail what the spatiotemporal architectures of these subsystems actually are and to then compare them with an independent account what makes a structure an inner-sense mechanism or an interpretive mechanism. Recall that interpretive systems are modelled after the operation of folk psychological theories, whereas inner-sense systems are modelled after sensory systems. One hallmark feature of inner-sense systems that distinguishes them from interpretive systems is that they stand to be sensitive to certain sorts of error present due to the way things directly appear.[34] Or at any rate, a phenomenological component will be present. One can also tell, by introspection of this phenomenological component, what the content's of one's inner-sense is (as in cases of seeing-as, such as the duck-rabbit or Necker cube).

Consider first the cognitive mentalization network, which corresponds to the dMPFC-subsystem character state of the default network. Characteristic tasks for this system include social narrative comprehension, social reasoning involving social conventions and other rules, and moral decision making (e.g., trolley problems). These are all best characterized as interpretation-like. Take moral decision making in particular. It is well established that people do not have any introspective access to the reasons operative in moral deliberation; when asked to give reasons for moral judgments people provide a best guess rather than accurately reporting the operative reason underlying the judgment (Lanteri et al., 2008). This is what is expected of an interpretation-like system, as

[34] This feature is in analogy with sensory systems; one can sometimes tell that one's visual system may not be presenting things accurately due to how they look: an object in the dark may look like, say, an animal rather than a garbage bag (even though it is a garbage bag), but such uncertainty by the visual system is apparent in the way the object looks.

an inner-sense would allow for direct access to the content's of one's moral decisions. In addition, there does not seem to be any relevant phenomenology associated with moral deliberation. So it seems that cognitive mentalization may be interpretive rather than an inner-sense and will share these features.

The affective mentalization network employs the MTL-subsystem used in memory-based scene construction and simulation. Characteristic tasks involve mental imagery, spatial navigation, episodic autobiographical memory and episodic future thinking. These are best characterized as being like an inner-sense. Take mental imagery in particular. There are obvious resemblances to sensory systems here, including a relevant phenomenological component that can be introspected such that the contents can be directly read off of the phenomenology. For instance, one can directly report on one's inner speech by introspection of what is present in the phenomenological character of the inner speech, and these reports are not subject to the kind of confabulation found with reporting on the content of one's moral judgment. So it seems that affective mentalization resembles an inner sense rather than an interpretive mechanism.

Although these conclusions on the structure(s) of the mentalization system are tentative at best, they illustrate how cognitive homology can figure into arguments about the structure of cognitive mechanisms/systems. The contribution of thinking about cognitive homology here is that it highlights that there are four rather than three possibilities with respect to the core mindreading mechanism's structure as one of interpretive access or an inner sense: neither, the interpretive, inner-sense, or *both*. The important point given by the cognitive homology framework is that one and the same evolutionary character can vary in its character state, so one and the same cognitive mechanism can vary in its architecture.

The second important point is with respect to evolvability the optimality argument. The argument depends on there being two mechanisms generated in the population so that selection can favor the optimal one, something that depends on developmental constraints. But it also requires that one can be generated in the *absence* of the other—another sort of developmental constraints—and if they cannot then selection cannot select for when one is present in the absence of the other. Instead, optimality will operate so that the

mentalization system is deployed in the right way at the right times (e.g., detect if another is a close/similar other, then simulate, otherwise deploy the cognitive mode, and so forth).

The third point is also with respect to the optimality argument. For natural selection must not only be able to select for one system's presence in the absence of the other, but *willing* to do so. It depends on it being the case that having mostly redundant systems is sufficiently costly so as to outweigh the perks of having two systems that are optimal for different situations (*viz*., self- and other-mentalization). What cognitive homology highlights here is how these redundant systems may contribute to other functions. One cannot just measure adaptive value of a system by dividing the gains afforded by it contributing to *one* function by the costs of having this system; one has to measure adaptive value with respect to how the costs of maintenance of one system weighs against the gains afforded by *all* of the functions it can be used for. For example, the simulative mode of the default network is used in spatial navigation, so even if it should barely contribute to mentalization capacities it will still be selected for because of its role in spatial navigation, and selection will favor using this system at the right time.

Conclusion

This thesis contributes to philosophy of cognitive science by evaluating the prospects of viewing psychological kinds in terms of cognitive homology over a functionalist view. Whereas according to functionalism psychological kinds are functional roles which can be implemented in a variety of biological systems, homology approaches to psychological kinds view them as biological systems which can have a variety of functional roles. The standard for deciding between these is how these stances foster good scientific theorizing. Homology has previously been argued to be superior to functionalism because it captures deep causal commonalities in the context of the emotions. In contrast to homology, functionalist approaches to psychological kinds do not tend to adequately capture deep causal and computational relationships that are important for good scientific theorizing about psychological kinds (such as emotions) in cognitive science. Homology can do so because unlike functional kinds, homologues are historical homeostatic property clusters which are hierarchically composed of further historical, homeostatic property clusters.

I aimed to expand the scope of homology in philosophy of cognitive science to a wider range of psychological kinds than just the emotions. The past focus on emotions is partially attributable to the current understanding of the homology concept as centered on phylogeny: relations of homology between individuals and even species. In contrast, ontogeny is central to understanding relationships among psychological kinds within an individual, but the corresponding developmental (serial) homology concept is not as well theorized. This less well understood notion of serial homology and the nature of cognitive activities as spatiotemporal organizations present difficulties for homology thinking, as the homology concept is largely understood in terms of phylogenetic homology concerning morphological structures. By reformulating Remane's operational criteria in terms of serial homology and cognitive activities, I showed how cognitive homology can overcome these obstacles, and by using concrete cases as illustrations I showed how cognitive homology can established from empirical data in cognitive science.

Three general types of cognitive homologues emerged from considering cognitive activity at a number of levels of organization: local activities, intrinsic networks, and soft assembled systems. First, there is cognitive homology for very local activities performed by particular neural circuits, as exemplified by theories of neural reuse. For these activities, a large part of what is of interest is serial homology of token cognitive activities, because serial homology at the type level requires two non-numerically identical types of activities—something which is not present in many cases of neural reuse.

Second, there are intrinsic cognitive networks such as the default network. Intrinsic networks are interesting because these large networks can be viewed as evolutionary characters, but they are considerably complex in the way they and vary in form (as well as function) within an individual. I focused on the default network as a paradigm case of such an intrinsic network, and discussed serial homology of the default network's character states by way of its two subsystems.

Third, there are soft assembled systems that are assembled on the fly in order to accomplish some task, and I argued that these too are in fact amenable to homology thinking. The worry was that if homology thinking does not apply to soft assembled systems, a lot of psychological kinds under the purview of mechanistic explanation will be left out of the picture. But soft assemblies are arguably stabilized by their own developmental histories of being used in coordination, and given such homeostatic mechanisms, they can be viewed as cognitive homologues which are considerably dependent on environmental input, cultural norms, and the like.

In addition to causal depth captured by homology, cognitive homology is valuable because of the conceptual independence of cognitive homologues from the functional roles they are used for. Because of this independence cognitive homologues can ground causal and computational similarities in the face of variation in functional architectures, fulfilling the explanatory goal of common coding attributions. I argued that this was so using common coding explanations of similarities between imagination and corresponding non-imagination representational states. Moreover, I argued that cognitive homology can also form a basis for further understanding representational codes/formats more

generally, and did so by drawing on a number of desiderata. Homology thinking accommodates the independence of codes from representational content, their characterization according to the typical contents they have, and how codes exhibit spatiotemporal hierarchies. Homology thinking about codes also positively contributes to clarifying the relationship codes have to their neural substrates and internal structure: codes map one-to-one on to activity functions, but many-to-one onto neural substrates. Moreover, codes are not identical to their internal organization, as this can vary, but rather codes are constituted by them.

Cognitive homology is also valuable because it is a concept that relates to the intersection of evolution and development by way of evolvability. This allows for addressing the part of the evolutionary process concerning evolutionary potential: what kind of traits can originate in a population in the first place, which can subsequently be acted upon by natural selection. Having the right evolutionary potential is precondition for the success of optimality arguments about the probable architecture of cognitive systems, as what is selected for as optimal depends on which traits can in fact developmentally originate. Moreover, they have to be able to originate independently of each other and not be used for other reasons (so that there is a cost in maintaining a mostly functionally redundant system). I considered the optimality argument with respect to initial selection pressures on the core mentalization system's resultant architecture being either interpretive or like an inner-sense. However, from the perspective of cognitive homology this system does not need to be one way to the exclusion of the other. If the mentalization system is a cognitive homologue, it is possible that it varies in character state so that it sometimes is interpretive and sometimes like an inner sense. I argued that if the default network is homologous to the core mentalization mechanism (as is not implausible), then we can look at its character states to see whether this system can be interpretive and/or like an inner-sense. Given the two sub-systems as reflecting serial homologues of the default network, it is also not implausible that mentalization system can be described as both interpretive and inner-sense. In one form, the mentalization system seems to have the features of an interpretive

system, whereas for this very same system in another state, it is more like an inner-sense.

In cognitive science, there is increasing interest in biological approaches to cognition, and the recent interest in homology as a fundamental concept for thinking about cognition no doubt owes to this. Theorizing about cognitive homology is in its infancy, but if I have been successful I have shown that it has great promise: it can ground thinking about psychological kinds in a way that can be fruitfully used to understand causal and computational relationships among psychological kinds as part of an evolutionary developmental cognitive science.

Bibliography

Abu-Akel, A., and Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* 49, 2971–2984.

Amundson, R. A. (1994). Two concepts of constraint: Adaptationism and the challenge from developmental biology. *Philosophy of Science* 61, 556–578.

Amunts, K., Schleicher, A., and Zilles, K. (2007). Cytoarchitecture of the cerebral cortex: More than localization. *NeuroImage* 37, 1061–1065.

Anderson, M. L. (2007a). Massive redeployment, exaptation, and the functional integration of cognitive operations. *Synthese* 159, 329–345.

Anderson, M. L. (2007b). The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology* 20, 143–174.

Anderson, M. L. (2008). Circuit sharing and the implementation of intelligent systems. *Connection Science* 20, 239–251.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33, 245–266.

Anderson, M. L., Richardson, M. J., and Chemero, A. (2012). Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science*. doi: 10.1111/j.1756-8765.2012.01211.x

Anderson, M. L., and Penner-Wilger, M. (2012). Neural reuse in the evolution and development of the brain: Evidence for developmental homology? *Developmental Psychobiology*. doi: 10.1002/dev.21055

Andres, M., Di Luca, S., and Pesenti, M. (2008). Finger counting: The missing tool? *Behavioral and Brain Sciences* 31, 642–643.

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist* 18, 251–270.

Assis, L., and Brigandt, I. (2009). Homology: Homeostatic property cluster kinds in systematics and evolution. *Evolutionary Biology* 36, 248–255.

Bach, T. (2011). Structure-mapping: Directions from simulation to theory. *Philosophical Psychology* 24, 23–51.

Balari, S., and Lorenzo, G. (2008). Pere Alberch's developmental morphospaces and the evolution of cognition. *Biological Theory* 3, 297–304.

Balari, S., and Lorenzo, G. (2009). Computational phenotypes: Where the theory of computation meets evo-devo. *Biolinguistics* 3, 2–60.

Barsalou, L. W. (2008). Grounded cognition. *Annual Reviews Psychology* 59, 617–645.

de Beer, G. R. (1971). *Homology, an Unsolved Problem*. Oxford University Press.

Bergeron, V. (2010). Neural reuse and cognitive homology. *Behavioral and Brain Sciences* 33, 268–269.

Block, N. (1995). The mind as the software of the brain: An invitation to cognitive science. MIT Press.

Blumberg, M. S. (2012). Homology, correspondence, and continuity across development: The case of sleep. *Developmental Psychobiology*. doi: 10.1002/dev.21024

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61, 127–148.

Brigandt, I. (2002). Homology and the origin of correspondence. *Biology & Philosophy* 17, 389-407.

Brigandt, I. (2007). Typology now: Homology and developmental constraints explain evolvability. *Biology & Philosophy* 22, 709–725.

Brigandt, I. (2009). Natural kinds in evolution and systematics: Metaphysical and epistemological considerations. *Acta Biotheoretica* 57, 77–97.

Brigandt, I. (forthcoming). "From developmental constraints to evolvability: How concepts figure in explanation and disciplinary identity," in *Conceptual Change in Biology: Scientific and Philosophical Perspectives on Evolution and Development*, ed. A. C. Love. Berlin: Springer

Brigandt, I., and Griffiths, P. E. (2007). The importance of homology for biology and philosophy. *Biology & Philosophy* 22, 633–641.

Broyd, S. J., Demanuele, C., Debener, S., Helps, S. K., James, C. J., and Sonuga-Barke, E. J. S. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews* 33, 279–296.

Buckner, R. L. (2012). The serendipitous discovery of the brain's default network. *NeuroImage* 62, 1137–1145.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences* 1124, 1–38.

Bullmore, E., and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10, 186–198.

Byrne, R. W. (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Clarendon Press.

Caeyenberghs, K., Tsoupas, J., Wilson, P. H., and Smits-Engelsman, B. C. M. (2009a). Motor imagery development in primary school children. *Developmental Neuropsychology* 34, 103–121.

Caeyenberghs, K., Wilson, P. H., Van Roon, D., Swinnen, S. P., and Smits-Engelsman, B. C. M. (2009b). Increasing convergence between imagined and executed movement across development: Evidence for the emergence of movement representations. *Developmental Science* 12, 474–483.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32, 121–138.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-knowledge*. Oxford University Press, USA.

Charland, L. C. (2002). The natural kind status of emotion. *The British Journal for the Philosophy of Science* 53, 511.

Clark, J. A. (2009). Relations of homology between higher cognitive emotions and basic emotions. *Biology & Philosophy* 25, 75–94.

Clark, J. A. (2010). Hubristic and authentic pride as serial homologues: The same but different. *Emotion Review* 2, 397.

Clark, J. A. (2012). Intersections between development and evolution in the classification of emotions. *Developmental Psychobiology*. doi: 10.1002/dev.21063

Clark, J. A. (forthcoming). Integrating basic and higher cognitive emotions within a common evolutionary framework: Lessons from the transformation of primate dominance into human pride. *Philosophical Psychology*. doi: 10.1080/09515089.2012.659168

Coltheart, M. (2010). Lessons from cognitive neuropsychology for cognitive science: A reply to Patterson and Plaut (2009). *Topics in Cognitive Science* 2, 3–11.

Cracraft, J. (2005). Phylogeny and evo-devo: Characters, homology, and the historical analysis of the evolution of development. *Zoology (Jena)* 108, 345–356.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, USA.

Decety, J., and Sveltova, M. (2011). Putting together phylogenetic and ontogenetic perspectives on empathy. *Developmental Cognitive Neuroscience*. doi: 10.1016/j.dcn.2011.05.003

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience* 12, 43–56.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition* 44, 1–42.

Donley, J. M., Sepulveda, C. A., Konstantinidis, P., Gemballa, S., and Shadwick, R. E. (2004). Convergent evolution in mechanical design of lamnid sharks and tunas. *Nature* 429, 61–65.

Donoghue, M. J. (1992). "Homology," in *Keywords in Evolutionary Biology*, eds. E. F. Keller and E. A. Lloyd (Cambridge, MA: Harvard University Press), 170–179.

Ereshefsky, M. (2007). Psychological categories as homologies: lessons from ethology. *Biology & Philosophy* 22, 659–674.

Ereshefsky, M. (2012). Homology thinking. *Biology & Philosophy* 27, 381-400.

Fadiga, L., Craighero, L., and D'Ausilio, A. (2009). Broca's Area in Language, Action, and Music. *Annals of the New York Academy of Sciences* 1169, 448–458.

Farag, C., Troiani, V., Bonner, M., Powers, C., Avants, B., Gee, J., and Grossman, M. (2010). Hierarchical organization of scripts: converging evidence from fMRI and frontotemporal degeneration. *Cerebral Cortex* 20, 2453 –2463.

Fox, P. T., and Friston, K. J. (2012). Distributed processing; Distributed functions? *NeuroImage* 61, 407–426.

Foxe, J. J., and Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology* 2, 154.

Friston, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biology* 7, e1000033.

Gallese, V., and Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2, 493–501.

Gao, W., and Lin, W. (2012). Frontal parietal control network regulates the anti-correlated default and dorsal attention networks. *Human Brain Mapping* 33, 192–202.

García, C. L. (2010). Functional homology and functional variation in evolutionary cognitive science. *Biological Theory* 5, 124–135.

Gendler, T. (2011). "Imagination," in The Stanford Encyclopedia of Philosophy, ed. E. N. Zalta [http://plato.stanford.edu/archives/fall2011/entries/imagination/]

Gerhart, J. C., and Kirschner, M. W. (2003). "Evolvability," in *Keywords and Concepts in Evolutionary Developmental Biology*, eds. B. K. Hall and W. M. Olson (Cambridge, MA: Harvard University Press), 218–227.

Glahn, D. C., Winkler, A. M., Kochunov, P., Almasy, L., Duggirala, R., Carless, M. A., Curran, J. C., Olvera, R. L., Laird, A. R., Smith, S. M., et al. (2010). Genetic control over the resting brain. *Proceedings of the National Academy of Sciences* 107, 1223–1228.

Glenberg, A., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., and Buccino, G. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology* 61, 905–919.

Goldman, A. I. (2006a). Imagination and simulation in audience responses to fiction. *The Architecture of the Imagination* (Oxford University Press), 41–57.

Goldman, A. I. (2006b). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press, USA.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology* 3, 71–88.

Goldman, A., and de Vignemont, F. (2009). Is social cognition embodied? *Trends in Cognitive Sciences* 13, 154–159.

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences* 15, 20–25.

Goodman, N. (1983). *Fact, Fiction, and Forecast*. Harvard University Press.

Griffiths, P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press.

Griffiths, P. E. (2003). Basic emotions, complex emotions, Machiavellian emotions. *Royal Institute of Philosophy Supplement* 52, 39–67.

Hall, B. K. (2012). Homology, homoplasy, novelty, and behavior. *Developmental Psychobiology*. doi: 10.1002/dev.21039

Hall, B. K., and Olson, W. M. eds. (2003). *Keywords and Concepts in Evolutionary Developmental Biology*. Harvard University Press.

Hendrikse, J. L., Parsons, T. E., and Hallgrímsson, B. (2007). Evolvability as the proper focus of evolutionary developmental biology. *Evolution & Development* 9, 393–401.

Hubbard, E. M., Piazza, M., Pinel, P., and Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience* 6, 435–448.

Jabbi, M., Bastiaansen, J., and Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE* 3, e2939.

Katz, P. S. (2011). Neural mechanisms underlying the evolvability of behaviour. *Philosophical Transactions of the Royal Society B* 366, 2086–2099.

Keller, S. S., Crow, T., Foundas, A., Amunts, K., and Roberts, N. (2009). Broca's area: Nomenclature, anatomy, typology and asymmetry. *Brain and Language* 109, 29–48.

Kello, C. T., Beltz, B. C., Holden, J. G., and Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General* 136, 551–568.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.

Kitayama, S., and Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual Review of Psychology* 62, 419–449.

Klein, C. (2010). Redeployed functions versus spreading activation: A potential confound. *Behavioral and Brain Sciences* 33, 280–281.

Laird, A. R., Fox, P. M., Eickhoff, S. B., Turner, J. A., Ray, K. L., McKay, D. R., Glahn, D. C., Beckmann, C. F., Smith, S. M., and Fox, P. T. (2011). Behavioral interpretations of intrinsic connectivity networks. *Journal of Cognitive Neuroscience* 23, 4022–4037.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books.

Langland-Hassan, P. (2012). Pretense, imagination and belief: The single attitude theory. *Philosophical Studies*. doi: 10.1007/s11098-011-9696-3

Lanteri, A., Chelini, C., and Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics* 83, 789–804.

Lanyon, S. J. (2010). *A Saltational Approach for the Evolution of Human Cognition and Language*. University of New South Wales Press.

Laubichler, M. D. (2008). "Evolutionary Developmental Biology," in *Cambridge Companions to Philosophy* (New York: Cambridge University Press), 342–360.

Leech, R., Braga, R., and Sharp, D. J. (2012). Echoes of the brain within the posterior cingulate cortex. *Journal of Neuroscience* 32, 215–222.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review* 94, 412-432.

Lewis, D. (1980). Mad pain and Martian pain. *Readings in the Philosophy of Psychology* 1, 216–222.

Lickliter, R., and Bahrick, L. E. (2012). The concept of homology as a basis for evaluating developmental mechanisms: Exploring selective attention across the life-span. *Developmental Psychobiology*. doi: 10.1002/dev.21037

Light, G. A., Williams, L. E., Minow, F., Sprock, J., Rissling, A., Sharp, R., Swerdlow, N. R., and Braff, D. L. (2010). Electroencephalography (EEG) and event-related potentials (ERPs) with human participants, in *Current Protocols in Neuroscience*, eds. J. N. Crawley, C. R. Gerfen, M. A. Rogawski, D. R. Sibley, P. Skolnick, and S. Wray (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Love, A. C. (2007). Functional homology and homology of function: Biological concepts and philosophical consequences. *Biology & Philosophy* 22, 691–708.

Lu, H., Zou, Q., Gu, H., Raichle, M. E., Stein, E. A., and Yang, Y. (2012). Rat brains also have a default mode network. *PNAS* 109, 3979–3984.

Di Luca, S., and Pesenti, M. (2011). Finger numeral representations: More than just another symbolic code. *Frontiers in Psychology* 2.

Machery, E. (2010). Reply to Barbara Malt and Jesse Prinz. *Mind & Language* 25, 634–646.

Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., Villringer, A., Castellanos, F. X., Milham, M. P., and Petrides, M. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences USA* 106, 20069–20074.

Marr, D., Ullman, S., and Poggio, T. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. MIT Press.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology* 58, 25–45.

Matthen, M. (1998). Biological universals and the nature of fear. *The Journal of Philosophy* 95, 105–132.

Matthen, M. (2000). What is a hand? What is a mind? *Revue Internationale de Philosophie*, 653–672.

Matthen, M. (2007). Defining vision: What homology thinking contributes. *Biology & Philosophy* 22, 675–689.

Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D., and Wolpert, L. (1985). Developmental constraints and evolution: A perspective from the mountain lake conference on development and evolution. *The Quarterly Review of Biology* 60, 265–287.

McGinn, C. (2004). *Mindsight: Image, Dream, Meaning*. Harvard University Press.

McGinn, C. (2009). "Imagination," in *The Oxford Handbook of Philosophy of Mind* (Oxford, England: Oxford University Press), 595–607.

Michel, G. F. (2012). The concept of homology in the development of handedness. *Developmental Psychobiology*. doi: 10.1002/dev.21038

Moore, C. (2012). Homology in the development of triadic interaction and language. *Developmental Psychobiology*. doi: 10.1002/dev.21032

Moore, D. S. (2012). Importing the homology concept from biology into developmental psychology. *Developmental Psychobiology*. doi: 10.1002/dev.21015

Moore, D. S., and Moore, C. (2010). Neural reuse as a source of developmental homology. *Behavioral and Brain Sciences* 33, 284–285.

Müller, G. B. (2003). Embryonic motility: Environmental influences and evolutionary innovation. *Evolution & Development* 5, 56–60.

Müller, G. B., and Wagner, G. P. (2003). "Innovation," in Keywords and Concepts in Evolutionary Developmental Biology, , eds. B. K. Hall and W. M. Olson (Cambridge, MA: Harvard University Press), 218–227.

Murata, A., Moser, J. S., and Kitayama, S. (2012). Culture shapes electrocortical responses during emotion suppression. *Social Cognitive Affective Neuroscience.* doi: 10.1093/scan/nss036

Nichols, S. (2004). Imagining and believing: The promise of a single code. *The Journal of Aesthetics and Art Criticism* 62, 129–139.

Nichols, S. (2006). Just the imagination: Why imagining doesn't behave like believing. *Mind & Language* 21, 459–474.

Nichols, S. (2008). Imagination and the I. *Mind & Language* 23, 518–535.

Nichols, S., and Stich, S. P. (2000). A cognitive theory of pretense. *Cognition* 74, 115–147.

Nichols, S., and Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. Oxford University Press.

Nichols, S., and Stich, S. P. (2006). *The Architecture of the Imagination: New Essays on Pretence, Possibility, and Fiction*. 1st ed. Oxford University Press, USA.

Olivetti Belardinelli, M., Palmiero, M., Sestieri, C., Nardo, D., Di Matteo, R., Londei, A., D'Ausilio, A., Ferretti, A., Del Gratta, C., and Romani, G. L. (2009). An fMRI investigation on image generation in different sensory modalities: The influence of vividness. *Acta Psychologica* 132, 190–200.

Van Overwalle, F., and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage* 48, 564–584.

Owen, R., and Cooper, W. W. (1843). *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. Delivered at the Royal College of Surgeons, in 1843. Longman, Brown, Green, and Longmans.

Papaxanthis, C., Pozzo, T., Skoura, X., and Schieppati, M. (2002). Does order and timing in performance of imagined and actual movements affect the motor imagery process? The duration of walking and writing task. *Behavioural Brain Research* 134, 209–215.

Patterson, K., and Plaut, D. C. (2009). "Shallow draughts intoxicate the brain": Lessons from cognitive science for cognitive neuropsychology. *Topics in Cognitive Science* 1, 39–58.

Penner-Wilger, M., and Anderson, M. L. (2008). An alternative view of the relation between finger gnosis and math ability: Redeployment of finger representations for the representation of number. in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Austin, TX, July, 23–26.

Piccinini, G., and Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183(3), 283-311.

Platt, M. L., and Spelke, E. S. (2009). What can developmental and comparative cognitive neuroscience tell us about the adult human brain? *Current Opinion in Neurobiology* 19, 1–5.

Plaut, D. C., and Patterson, K. (2010). Beyond functional architecture in cognitive neuropsychology: A reply to Coltheart (2010). *Topics in Cognitive Science* 2, 12–14.

Ploeger, A., and Galis, F. (2011). Evo devo and cognitive science. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 429–440.

Poletti, M., Enrici, I., and Adenzato, M. (2012). Cognitive and affective Theory of Mind in neurodegenerative diseases: Neuropsychological, neuroanatomical and neurochemical levels. *Neuroscience & Biobehavioral Reviews* 36(9), 2147-2164. doi: 10.1016/j.neubiorev.2012.07.004

Polger, T. W. (2012). Functionalism as a philosophical theory of the cognitive sciences. *Wiley Interdisciplinary Reviews: Cognitive Science* 3, 337-348. doi: 10.1002/wcs.1170

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron* 72, 665–678.

Preston, S. D., and de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25, 1–20.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6, 576–582.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *PNAS* 98, 676–682.

Remane, A. (1952). *Die Grundlagen des Natürlichen Systems, der Vergleichenden Anatomie und der Phylogenetik: Theoretische Morphologie und Systematik I*. Leipzig: Geest & Portig.

Rieppel, O., and Kearney, M. (2007). The poverty of taxonomic characters. *Biology and Philosophy* 22, 95–113.

Rizzolatti, G., and Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience* 11, 264–274.

Robinson, R. (2011). Different paths, same structure: "Developmental systems drift" at work. *PLoS Biology* 9, e1001113.

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* 52, 1059–1069.

Scharff, C., and Petri, J. (2011). Evo-devo, deep homology and FoxP2: Implications for the evolution of speech and language. *Philosophical Transactions of the Royal Society B* 366, 2124–2140.

Schilbach, L., Bzdok, D., Timmermans, B., Fox, P. T., Laird, A. R., Vogeley, K., and Eickhoff, S. B. (2012). Introspective minds: Using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. *PLoS ONE* 7, e30920.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., and Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition* 17, 457–467.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *The Journal of Neuroscience* 27, 2349–2356.

Shubin, N., Tabin, C., and Carroll, S. (2009). Deep homology and the origins of evolutionary novelty. *Nature* 457, 818–823.

Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., and Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia* 45, 2802–2810.

Solomon, K. O., and Barsalou, L. W. (2004). Perceptual simulation in property verification. *Memory & Cognition* 32, 244–259.

Spaulding, S. (2012). Mirror neurons are not evidence for the simulation theory. *Synthese*.

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* 8, 418–425.

Spreng, R. N. (2012). The Fallacy of a "task-negative" network. *Frontiers in Psychology* 3.

Teufel, C., Fletcher, P. C., and Davis, G. (2010). Seeing other minds: Attributed mental states influence perception. *Trends in Cognitive Sciences* 14, 376–382.

Tomasi, D., and Volkow, N. D. (2011). Association between functional connectivity hubs and brain networks. *Cerebral Cortex* 21(9), 2003-2013.

True, J. R., and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evolution & Development* 3, 109–119.

de Waal, F. B. M., and Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences* 14, 201–207.

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding Accounting for the mismatch negativity. *Journal of Neuroscience* 32, 3665–3678.

Wagner, G. P. (1992). Homology and the mechanisms of development, in *Homology: The Hierarchical Basis of Comparative Biology*, ed. B. K. Hall (San Diego: Academic Press), 273–299.

Wagner, G. P. (1996). Homologues, natural kinds and the evolution of modularity. *American Zoologist* 36, 36–43.

Wagner, G. P. (2000). What is the promise of developmental evolution? Part I: Why is developmental biology necessary to explain evolutionary innovations? *Journal of Experimental Zoology* 288, 95–98.

Wagner, G. P. (2007). The developmental genetics of homology. *Nature Reviews Genetics* 8, 473–479.

Walton, K. L. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Harvard University Press.

Wilson, R. A., Barker, M. J., and Brigandt, I. (2007). When traditional essentialism fails: Biological natural kinds. *Philosophical Topics* 35, 189-215.

Wimsatt, W. C. (1986). Developmental constraints, generative entrenchment, and the innate-acquired distinction, in *Integrating Biological Disciplines, (*ed. William Bechtel), 185–208.

Wouters, A.G., (2003). Four notions of biological function*. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34, 633–668.

Yablo, S. (1993). Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53, 1–42.

Zanto, T. P., Rubens, M. T., Bollinger, J., and Gazzaley, A. (2010). Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage* 53, 736–745.

Zanto, T. P., Rubens, M. T., Thangavel, A., and Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience.* 14, 656–U156.

Zeman, A. Z. J., Della Sala, S., Torrens, L. A., Gountouna, V.-E., McGonigle, D. J., and Logie, R. H. (2010). Loss of imagery phenomenology with intact visuo-spatial task performance: A case of "blind imagination." *Neuropsychologia* 48, 145–155.