THE CONCEPT OF MODULARITY IN COGNITIVE SCIENCE

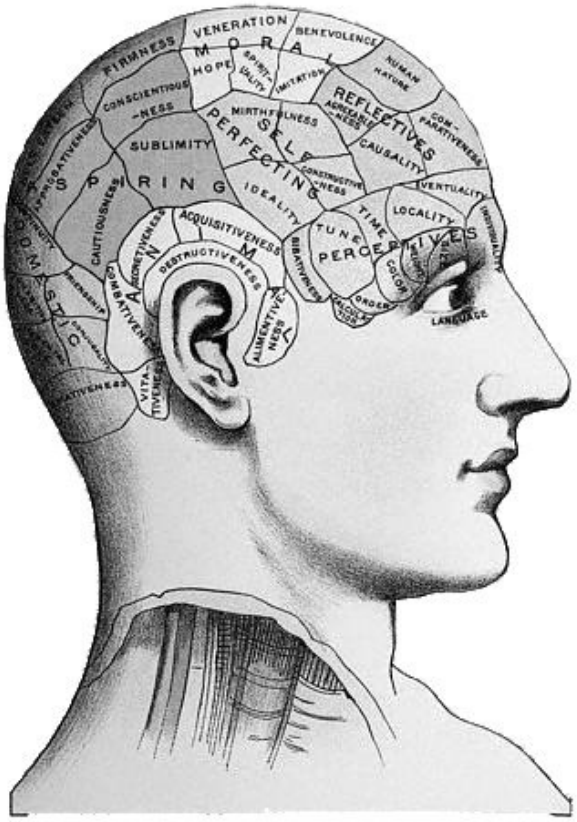
By

Amol R. Sarva

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
from Leland Stanford Junior University
2003

Doctoral Committee:

Professor Mark Crimmins, Chair
Professor Peter Godfrey-Smith
Professor Ken Taylor
Professor Lanier Anderson
Professor Susan Johnson



The Author's Mind

Image from Corbis

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Mark Crimmins, Principal Advisor

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Peter Godfrey-Smith

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Ken Taylor

Approved for the University Committee on Graduate Studies

_____

For the modules of my family and loved ones

Acknowledgments

Bigmouth strikes again. Apologies and many thanks to Mark Crimmins for supervising this work, and to Peter Godfrey-Smith and Ken Taylor for key inspirations. Tim Bloser read many drafts and helped me with this as much as anyone. Regards to some of my colleagues for their companionship: Patrick Scotto DiLuzio, Ben Escoto and John Eden.

My Columbia professors—John Collins, Taylor Carman, and Akeel Bilgrami—gave crucial encouragement. Some philosophers loomed large on the path ahead of me, Sidney Morgenbesser, Jerry Fodor and Robert Nozick. Thanks to Pierre Jacob and the CNRS for their hospitality in Paris.

Thanks to Pramila Sarva for calling me professor from an early age, and to Ramesh Sarva for his many examples of philosophical reflection. Chetan and Paraag, my brothers, have always made very clever sparring partners. Ursula Wynhoven let me stow away with her on the world tour during which this thesis was written. Great gratitude to my good friends for their fellowship, encouragement, and inspiration well apart from this one project: Matte Chi, Tom Sanford, Jason Lehmbeck, Joe Master, James Kearney, Darius Loghmanee, John Tantum, Roy Bahat, Jacob Waldman, Mustafa Khan, Jay Bhattacharya, Edward Rhee, Hugh Son, Gary Alperson, Julie Sheinman, Brett Knappe and many others who I have admired.

Preface

This dissertation is about the modularity of mind, and a network of related problems. Modularity is a familiar idea in the foundations of cognitive science, but less clear is just how modularity relates to tacit knowledge, nativism, and domain-specificity, as well as computationalism, neuroscience, connectionism, developmental psychology, and a dozen other progressive programs.

This dissertation is not "one long argument", but it is a work with one purpose: to clarify the picture of cognitive architecture that has been developing since Noam Chomsky's early work and particularly since Jerry Fodor's (1983) *The Modularity of Mind.* The big idea has certainly caught on: the mind is not one homogenous thing, but many component parts, interacting in various ways. Modules have joined with *computationalism* and *nativism* to form the conceptual trinity that supports a large part of cognitive science research in the last 20 years, and the launching point for a number of new directions. But the literature is full of varying uses, confusing key elements of the theory with those that are merely peripheral. This had led to problems in empirical theorizing. The aim of this dissertation is to take on a set of these confusions and offer solutions.

Table of Contents

Chapter 1. Introduction. The Concept of Modularity

1. Introduction

A central principal for much cognitive science is that the mind is modular, the thesis that it is composed of interacting but independent subsystems. The concept is nearly everywhere, in models of mind and individual cognitive phenomena and in the basic methodology for studying those phenomena. Chomsky (1966) launched the modern round of modularist thinking, isolating the language "faculty" as to be explained by an independent causal "organ" and distinct subject of study. Fodor's (1983) bold treatment set the agenda of issues for much of the contemporary debate. But while Fodor was principally concerned to state the limits of computational modular psychology, the emphasis has increasingly shifted to areas where he had less structure to offer: adaptationism, higher-order cognition, domain-generality, learned modules, and other phenomena bound up with whole fields of research such as evolutionary psychology, neuroscience, developmental psychology, and connectionism.

Where Fodor (1983) canonized a concept that Chomsky had pushed back to the fore of psychology, by establishing its historical connections and articulating the landscape of key ideas, the increasing popularity of modularity has begun to test its coherence. Fodor's modules were speculatively assigned a series of specific features, such as innateness or domain-specificity. Modularist models of mind have appeared that defy nearly every one

of these assignments, producing pictures of mind that vary widely from the archetype. Karmiloff-Smith's (1992) suggests a cognitive architecture that only becomes modular through the result of experience and development, but which does not contain modular structure innately. Farah (1994) attacks the assumption that neurally localized regions will correspond to unique cognitive functions. Barkow, Cosmides, and Tooby (1992) present a collection of views where modularity is a feature not only of "input systems" but of every cognitive process, even high-level reasoning. Theorists working with folk psychology have put high stakes on the distinctive features of so-called "theoretical" modules as distinguished from "mechanical" modules, a theretofore invisible distinction. Wrapped around these and other debates is a contest over what it means to claim something is a module. There is no consensus on a clear definition: so far, modularity is a rough-and-ready concept with a family of probable but non-dispositive features.

Clarifying the concept of modularity is the central aim of this dissertation. Part of this requires a proposal for how modularity should be understood, and what should count as a cognitive module. The existing literature presents only a tangle of inter-related issues where a unique account of modularity is needed. One wrong conclusion that this could suggest is that modularity as such is not a distinct hypothesis about the mind to be tested empirically. On this view, "module" just means a cognitive capacity meeting one or another list of more concretely defined features, such as innateness, domain-specificity, computational implementation, or others. We should not draw this conclusion, because modularity is indeed a freestanding and interesting doctrine about the mind. Its fortunes move independently of many of the concepts to which it is closely confederated, and

there is a common body of features that the literature typically invokes by talking of modules. This dissertation makes the case for this view by arguing for a basic definition of modularity as a theory of "informationally isolated" subsystems, a new term introduced for a moderately revised version of Pylyshyn's (1984) "cognitively impenetrable" and Fodor's (1983) "informationally encapsulated".

Fodor (2000) wisely refrains from the impulse to "legislate" a meaning for modularity in his review of recent developments, and the intention here is similar. While the thicket of modularist views certainly contains an array of clearly implausible or mutually inconsistent formulations, pure theoretical pruning still leaves many options. Empirical results will confirm whether the mind is modular, and how so. The goal here then is taxonomic. Modularity theories are assembled from conceptual parts, and those are the first subject of this dissertation. The thorniest issues lie with these components, and their resolution promises the more valuable results.

Explaining what modularity is only begins to untangle a complicated network of issues, since an interesting feature of modular psychology is its intimate relations with other big and controversial ideas in cognitive science. Among these are tacit knowledge, nativism, and domain-specificity, all concepts on which much depends and concepts which are frequently applied to modules. Tacit knowledge states are intentional mental states with belief-like functional properties, but without the epistemic status, conscious accessibility, or inferential integration. Nativism is the doctrine that at least some of our cognitive endowment is present before any learning takes place. Domain-specificity is a feature of

some cognitive modules that are specialized to function on certain ranges of inputs. Nearly every discussion of modularity runs across these three ideas, often with deep dependence.

In this introductory chapter, let me offer a few lines about the three essential debates around these concepts that this dissertation takes up, issues broached after offering an overview of modularity and its history in the modern literature. The debate of Chapter 3 concerns tacit knowledge, a concept made respectable by Chomsky's declaration that grammars are tacitly known or "cognized". The consequence of this proposal, in conjunction with a computational model of mind, is that the cognitive "mechanisms" characterized in physical and biological terms can be nothing other than species of tacit knowledge. The dichotomy between a purely mentalistic usage of "knowledge" and the physically observable "mechanism" is not sustainable under the contemporary paradigm of the computational theory of mind. The consequence, as we consider closely with respect to the debates over folk psychology, is that two "types" of modules are really just two ways of describing one. Any physical system will implement some intentional module. Module does not permit of two distinct types, the intentional module or the mechanism module; rather, it just means a mechanism that implements a body of tacit knowledge.

The second major issue links the fates of modularity with another controversial doctrine: nativism. The polemical landscape has regularly found nativism and modularism on the agenda of one side, and their denial on the other, an especially vivid polarity since

Chomsky. Yet there is no explanation for this, and there does not appear to be any necessary connection between the two concepts. The argument of Chapter 4 is that there is indeed a link between the two programs in cognitive science. But it has an unusual foundations; the link is taken to be methodological or epistemic, part of how we come to confirm either view involves buttressing the arguments for its confrere. The two arguments go together, though empirical results could one day separate them.

The third big issue pertaining to modularity springs from domain-specificity. This is a ubiquitous concept often identified as the signal characteristic of cognitive modules, sometimes a key part of nativist doctrine, and also important to evolutionary psychology and other research programs. Despite its wide deployment in the literature, the concept of domain-specificity is too slippery to be useful. Modules are called domain-specific on the basis of little more than intuition, casting doubt on the other concepts it underwrites. The only productive route, according to Chapter 5, is to adopt explicitly a background informational account of mental states, and then characterize domain-specificity as a formal criterion on bodies of information. Domains are coherent, maximal, and eccentric sets of information, and domain-specific cognitive capacities consist of tacitly held knowledge which is both coherent with the domain and actually relevant to its content. Seeing a module as a body of information in this way lets us constrain the range of its logical application without smuggling in messy terms about which subject matters are relevant or unrelated.

The present chapter will present an overview of what modularity is, and defend a few key points on how it should be deployed. The next chapter will link modularity with its partner concept, innateness, and their respective histories in psychology primarily since Chomsky. The third chapter will look at the problem of distinguishing tacit knowledge from psychological mechanisms, with close attention to the recent debates over folk psychology. The fourth chapter will consider how innateness and modularity are related, since they are often implicated as having deep interconnection, and suggest that they share evidentiary bases. The final chapter considers domain-specificity, and the dangers of some of its treatments as compared with a purely informational account.

2. Elements of Modularity

Accounts of modularity generally take a list-like form, identifying a series of properties that describe the nature of cognitive modules. In some cases, one or a few of the concepts on this list are privileged as the constitutive characteristics of modularity. Identifying some of these elements therefore at least provides a guide to the major views, and a simple of way of focusing the issues.

This chapter takes a position on what a minimal account of modularity should look like. Claiming that a cognitive capacity is implemented by a module is to claim, at the very least, that the capacity is *independent* in the sense of *informationally isolated.* Independence is the key intuition about modules, and the argument here is that it should be interpreted in terms of the information flows between the subject capacity and others in the system. Informational isolation is the claim that the function which defines a module is fixed or rigid *whatever* the informational states elsewhere in the overall system

or in other modules. Modules rigidly map inputs to outputs. Informational isolation is a moderate revision of two existing and more familiar ideas, cognitive impenetrability and informational encapsulation. With these ideas in hand, we have a perfectly adequate understanding of what modularity implies for psychology. All the other features are completely optional at the theoretical level, though it may surely turn out *empirically* that some or all modules have such-and-such features from the list. This is the nutshell synopsis of the view I espouse in this chapter.

While I offer this summary statement of modularity up front, the key aim of this dissertation is not to re-define concepts already in wide use by researchers studying the mind. Modularity and affiliated concepts already carry well-established sense. The key aim here is to focus on unseen issues embedded in these widely used concepts, and suggested resolutions. With that in mind, the aim of this first chapter and the next is to set up the landscape in which modularity theories have flourished, and to establish this context as constraining our interpretation of problems further downstream. The key elements of modularity, the meaning of nativism, and the basic computationalist framework will play major roles in Chapters 3, 4, and 5; they will be set out in these first two chapters. But a result is that you will not see a radical attack on Fodor or Karmiloff-Smith or some other key theorist of modularity; hopefully, we will broach new directions.

What Modularity Is. Setting aside now the nutshell synopsis I provided a few lines back, let us introduce the concept of modularity from the most fundamental point. Modularity is first of all a claim about cognitive architecture, and in principal can be debated

independently of other types of questions about mind such as nativism or computationalism. In practice, this is not true, since arguments for modules typically involve assumptions about the nature of those modules or evidence by which we discover them. Modularist theories typically reject more about Descartes' view than merely the notion of a single, unitary mental space of operation. As a result, this claim about cognitive architecture is sometimes bound up with a claim about development, such as innateness, or a claim about neuroscience, such as brain localization. We will look at both purely architectural as well as non-architectural features.

The various features attributed to modularity by different authors constitute a menu from which we could define modularity. Perhaps such-and-such variety of localization is *the key* constitutive feature of a modular theory, or perhaps it is another feature. Rather than take on such views, such as those advocated by particular researchers, let me offer an alternative plan for proceeding. Understanding the modularity debate depends on understanding the elements of modularity on the menu of options turning up in various contexts. The following discussion will work through the entire list of options and consider some of the issues that arise around each.

The discussion starts, however, with the most basic and widely-accepted element of the modularity concept. Modularity is always a claim about the independence of cognitive systems. It is my polemical position in this chapter to take this basic feature as fundamental; all considerations beyond independence merely add detail to the constitutive criterion. That part is common ground; the polemical position advocates a

particular interpretation of independence as informational isolation. The resultant criteria of modularity leave it as a more-or-less characterization; systems are modular to varying degrees.

The criterion of informational isolation suggests a way to clarify one of the basic assumptions of all modularity theories. But it does not settle the issues on which they typically disagree—about whether modules must be innate, or neurally local, or other such issues. Instead, the view I'm advocating here just says that any modularity theory will rely on independence, and that independence should be interpreted as informational isolation. The remaining issues will still be important, and adjudicated on other grounds. Let's now consider the landscape of issues underlying modularity.

2.1 Subsystems

The fundamental aspect of the modular mind is that it has "isolable subsystems" (Shallice, 1994). Nearly any post-Cartesian view of mind at all will admit that the mind has parts, and does not operate as a single, undifferentiated *res*. The earliest 19th-century results of neuroscience suggested a physically compositional structure to the brain, first with regions of functional specialization (Gall and Spurzheim, 1824; Wernicke, 1874) and then even specialized cells (e.g., Golgi, Ortega y Casals). Even earlier, perhaps in Descartes (Hatfield, 1999), views of mental function suggested functionally distinct cognitive faculties. This fits with the principal method of investigation in neuroscience and often in biology more generally: *decomposition* (Zawidzki and Bechtel, in press).

Subsystems hide complexity from the architecture of the overall system. For example, a camera relies on the flash to provide illumination to the subject at the correct moment. An engineer may face this task and determine that there are many complex ways of solving it. One way may involve a power supply, a charging system, a trigger, amplification for the bulb, and a notification that the flash is not yet ready. Another way may simply trigger a solid-state flash bulb with its own fuel. The flash subsystem can be internally complex in many ways; but when we diagram the camera's overall architecture it is sufficient to simply designate an atomic entity to somehow produce the right outcome: a flash.

By contrast, mere parts are themselves individually simple in their performance of tasks that compose into the overall system's complex function. The chain in a bicycle functions as part of a more complicated assembly, but itself only translates force directly along from one gear to another without internally applying any conditional logic or complexity. Parts count as what Fodor calls "functionally individuated cognitive mechanisms", and are too simple to be a module (2000:56). The subsystem is like the overall system itself: complex. The subsystem is itself assembled, and hides some of its internally machinations from the overall system. The overall system just wants a particular input to yield a particular output; it does not care how. A camera's flash can be removed and replaced by a wide variety of others, constructed to perform their basic function in many different ways, because the flash is a subsystem. Its role in the larger system is prescribed to meet particular criteria which are neutral to how those multi-part, composed tasks are

actually carried out, and sensitive to certain special instructions about timing or brightness that they might receive.

Modules are more like subsystems than bare parts. The modules share many of the features of the overall system—complex, assembled, somewhat self-contained.[1] Stillings (1987) labels this "strong modularity", dismissing mere simple parts as totally uninteresting. The cognitive system comprising language ability can look cleanly divisible from the broader mix of mental functions, but is itself a set of complicated and interconnected abilities with a broad domain of application. The temptation to homoncularize individual cognitive functions comes partly from the surprisingly robust sophistication of linguistic or visual subsystems; there is no need for "little men" just to implement AND-gate logic. It is the sophisticated, module-scale tasks that feel so complicated that only a mind could implement them. At the other extreme of the continuum, a non-modular mind composed of only very simple parts, could still itself be highly complex. A classically associationist picture of mind composes mental functions out of two atomic elements: ideas and associations. The complex whole could consist of no distinguishable subsystems. The extent to which a mental part is a subsystem is graded or more-or-less, where clearer division of labor and isolation of internal complexity more sharply signal subsystems.

---

[1] "Assembled" just means that the task has component steps. This is different from the odd way Fodor uses "assembled" in his 1983 p. 37.

The notion of subsystem is theoretically neutral about *what* we say the mind is. If we focus on knowledge and its application, we might characterize the mind as composed of specialized subsystems of information, like books in a library. Chomsky's picture of the language systems usually talks only about a body of specialized knowledge, not about processors and algorithms. If we focus on the sensory modalities, each subsystem may be a format dependent processor (like compact disc player vs. record player), as in the case of Fodor's enumeration of classic modules like "hearing" or "vision". Yet another view might suggest that modules are specialized on classes of computations (like arithmetic vs. logarithm computation; rather than sensory modalities or subject matters of knowledge), with each module functioning like a sub-processor in a larger computing device. This view would draw attention to families of functions, possibly more in line with connectionist models. With the important caveat about the neutrality of how we have formulated subsystem, the premise that modules are subsystems is a basic and universal supposition among those suggesting modular architectures.

2.2 Independence

Modules are to some degree independent of each other and independent of non-modular regions of the mind such as "higher cognition". While the hierarchy of a subsystem architecture may put a particular module into a highly specific role, subordinate to some other process for instructions or inputs, the module itself is at least partly autonomous in fulfilling that role, "as nearly independent…as the overall task allows" (Marr, 1982:102). This autonomy is in the subsystem's rigid implementation of a particular core set of procedures, regardless of the external conditions. This rigidity is a matter of degree. Optical illusions demonstrate the strong rigidity of the vision system in interpreting

incoming data with respect to edges, overlap and relative position, even when the mind has explicitly contradictory information available. Object identification, meanwhile, seems less rigid, e.g. experimental evidence shows subjects are more likely to spot water in a picture if they are thirsty.

Subsystems can be independent in different ways. The emphasis might fall on the function implemented. Here we might say a module is independent only when it reliably produces the same *function,* the same input-output pairing, regardless of the states of other cognitive modules. A different view might put the emphasis elsewhere, such as on *resource autonomy*. The early neurologists clearly emphasized the operational independence of neural assemblies, since their evidence focused on the sensitivity of various faculties to neural damage. Destroying the brain's language centers might cut off nutrients to the object identification centers though there is no logical relation between the two areas activities. Cognitive psychologists studying error-rate patterns or task-completion speeds emphasize the independence of processing system or dependence on common resources (like working memory or channel bandwidth). Some functionally independent systems will be implemented on shared computational resources, just as two computer programs on a single machine share processor time and temporary disk space even though neither program experiences any state-changes due to the other. So there is a continuity between the early and contemporary views that focus on resource usage. There are also other options. The key point is to say that any modularity theory provides that the subsystems are independent in some important respect and to some degree.

Modules can only range from moderately independent to not at all independent. The perfectly independent module is unlikely to exist. Even in computer science, the design of perfectly non-correlated functions is notoriously difficult, especially in fields that emphasize independence from the environment such as key-encryption or random number generation. The mind itself is highly dependent on external input, such that most of its function appears to be designed to process the external world and produce actions that manipulate it. Of course, the mind and any module will rely on external sources for nutrients and so on. Even more strongly, any module so far proposed is tightly integrated into the functioning of many other components of the mind. The linguistic system has no contact with pure sound waves, for example, relying on interpreted signals from the early auditory system. This is more striking with higher-order modules, such as cheater-detectors or agglomerative counting mechanisms. However well we subdivide the mind's faculties and operation, the overall system displays well-coordinated sharing of information and division of labor.

Taken together, subsystems and independence are the basics of modularist psychology. They are the key characteristics of a cognitive module. This is moderately polemical. Typical discussions in the literature seem to be rather taken by subtler ideas, such as encapsulation or domain-specificity, as the hallmark features of modularity, or a further list of features that a module must have. But it is not polemical to say that every major view of modularity relies on independent cognitive subsystems as a starting point. In my view, it is important to set the base only here and recognize that we start with only a broad characterization. Fodor's (1983) precedent has been frequently misread as

enumerating conceptual *requirements*, while instead it was quite explicitly a speculation about empirical features to be eventually proved. There are many conceptually sensible ways to advocate modularity.[2]

2.2.1 Cognitive Impenetrability

Pylyshyn (1980, 1984) proposes a straightforward test of a module's independence. Cognitive mechanisms treat a particular input in a given way. The vision system interprets a certain contrast condition as an edge, for example. The test is simply whether varying the states of other mental systems has an effect on this outcome. If we add a belief to the belief box that the given image is a specially-designed optical illusion, does the vision system still call the detected contrast an edge? When cognitive conditions elsewhere in the mental environment change the mechanism's operation in this way, the mechanism is cognitively penetrable (we ignore non-cognitive phenomena, like nutrient availability or tissue damage). Where the mechanism is unchanged, it is cognitively impenetrable. For Pylyshyn, this is an indication that a cognitive faculty is implemented by a module.

---

[2] Another source of confusion comes from evolutionary psychology, where writers like Hirschfeld and Gelman (1994) or Sperber (unpublished, 2000) refer to "domains" or cognitive abilities interchangeably with "modules". This is a somewhat loose usage. They take the core observation to be that there are distinct faculties, identified by their specific domains of operation. Sperber speculates that the two basic proposed here are not needed, since we can make do only with finding those bare cognitive parts that have been specifically selected for by evolution. This, like most similar attempts to pick a single key concept aside from the two I have been discussing, simply smuggles them in. Subsystems and independence are always features of the evolutionary modules discussed by evolutionary psychologists.

The restriction to cognitive considerations focuses us on those conditions in the brain that count as mental states. We can ignore nutrient flows or the speed of a computation. Where we understand those mental states as computational, we have a standardized currency for assessing the relevant conditions. All cognitive states will have available some characterization in terms of propositions and inference rules. Where modifying a proposition or rule outside a module has an effect on its outcome, we can see the module's penetrability clearly.

This proposal articulates a characteristic of perceptual systems that drove their identification as modular: they are bottom-up processors of inputs. Information available to upstream systems, as in cases of optical illusions, cannot influence the way certain stimuli are perceived. These downstream systems are cognitively impenetrable with respect to certain upstream data sources. This one of the classic uses of "modular" in the literature, to describe psychological systems like vision or hearing that are not sensitive to changes in the subjects knowledge or beliefs.

This picture of modular independence oversimplifies the variety of ways information can flow between modules. The simple picture assumed here is that a module takes inputs, applies its core procedures, and dumps the outputs. At the mind's periphery, visual stimuli are a natural example of a sensory system's inputs. Somewhere upstream, the vision system delivers various products, such as 2.5-dimensional diagrams or lists of recognized objects. In between, the module's core procedure—a mix of inference rules

and their proprietary database of stored propositions—transforms the input information "intelligently", i.e. by enriching the raw data with inferences (Fodor, 1985c).

Inputs will come from other modules as well, even in cases of sensory systems drawing information mostly from the external world. A command from upstream may tell a given system to pay special attention to some detail (as might be the case where thirsty subjects are more likely to find water in a picture), or may shift the frame of analysis completely (such as the expectation of speech causing random auditory input to be analyzed as speech). It would be surprising if modules did not have to accept instructions from neighboring systems as they analyzed their inputs, so we should not expect that any module will be perfectly impenetrable. Pylyshyn's formula for assessing a module's penetrability leaves us with only a binary judgment to make, and most modules will simply turn out to be penetrable.

There is a further presumption that modules are organized hierarchically, that input just means information flowing upstream to the module. The question about penetrability is intended to assess the module's independence from modules *above* it, where higher means further away in cognition from external stimuli. This picture of overall cognitive architecture seems keyed to the models of vision developed in the 1980s, but not in accordance with less rigidly one-directional flows of information as suggested by developmental psychologists or evolutionary psychologists. Some information flows loop back through modules (folk psychology), branches out through two channels at once (separate paths for auditory/speech stimulus), or run backwards (as in imagined

visualizations). Rather than advocate either architectural model, it is worth inoculating the notion of independence against either assumption as we will do with informational isolation.

2.2.2 Informational Encapsulation

Taking Pylyshyn's very clear test for a module's independence, we can see how Fodor's (1983) notion of informational encapsulation improves its versatility. Fodor links his discussion to Gall's notion of independence, something like that of the other 19th century neurologists. For him, independence was about resource autonomy, such as memory space or processors. Fodor distinguishes the informational nature of the implemented functions, so that two functions might be very independent even while they share resources at the implementation level. So a cognitive module has a characterization as a function on certain types of inputs and producing certain outputs. That characterization is quite apart from its resource-consuming implementation in the brain's machinery.

The main question for independence, then, is whether information is flowing between these systems. An encapsulated module is one which is not drawing information from elsewhere, and so therefore is also impenetrable from above. By focusing on barriers to communication between modules, we avoid the automatic assumption of hierarchy in the flow of information. To be modular means to have closed channels rather than simply to be a one-way, upstream-flowing processor. Whereas penetrability stops with the yes-or-no assessment of whether there has been change to a module's function due to a change in some higher-level belief, the informational encapsulation criterion permits us to look a

bit more closely. It focuses precisely on the restrictions to information flow, not on simply on the testable consequence.

Fodor (1983) does not make a big deal of the difference himself. Pylyshyn's concern was to provide a test for models like Marr's (1982) which treated inputs only with fancy calculations; the aim was to rule out the downward penetration of more "intelligent" processes into these models of early vision or other sensory faculties. Cognitive penetration is the lynchpin test of that issue. But the underlying fact to explain this phenomenon is meant to be restrictions on information flow.

The merit of casting independence in terms of restricted information flow is that modules do not need all the information available in the mind. Only some types of information will be relevant. A panther-detector only needs panther-related facts, and even then only particular panther-related facts ("Is there are panther here?", but not "Panthers have toes"). So modules need to be encapsulated *against* useless information, since they should not even be forced to sort through all the loosely relevant information. Saying just that certain information does not in fact change a module's operation leaves it open that the module somehow receives and dismisses it. Rather, the point about encapsulation is that there are barriers to such flows.

Fodor obviously has in mind the types of barriers presented by restricted communication channels. Perhaps visual information is turned into beliefs somewhere higher in cognition ("That's lightning!"), but that such beliefs cannot be transformed into a format readable

by the hearing system (which only accepts sound waves). So while both represent information, there is a format problem in converting "that's lightning" into information the audition system can interpret as "expect thunder". Alternatively, there could be something equally crude going on, like a severing of the corpus callosum or some other information conduit. The barriers exist in both directions; not everything in the module is free to be read out into another system.

2.2.3 Informational Isolation

An isolated core. Focusing on the restrictions of information flow to establish independence can miss those aspects of a module which are indeed rigidly autonomous. The language module, for example, will have access to information about anything a person can talk about. We can talk about rules for non-human languages for example. Purely in terms of access to information, the system normally has nearly no restrictions. So we might conclude oddly that the language system is not at all encapsulated. A parallel case might involve a case of top-down influence, such as "knowing you are in France" influencing the way your language faculty parses incoming word streams. Again, this would be to miss something important about how in fact there is an independent and rigid process at work.

Looking at the internal operation of the module we will find it to be computing its outputs by applying a body of inference rules to the inputs, in combination with drawing on a body of fixed axiomatic propositions. One step down from what Marr called the "computational" level, there is the algorithm by which the function is actually realized.

That algorithm is rules of analysis that draw on stores of data (Peacocke's 1986 "level 1.5"), together sufficing to perform the function.

Consider a simple word-detector. Presented with phonemes or letters, the detector applies rules to parse the stream of input into likely word strings. It then applies rules for matching the word strings against the list of words in its lexicon. A bilingual speaker might maintain two separate lexicons of words, sometimes checking both databases until having determined safely to only expect words from one of the lexicons. Yet the command "expect only French words", for example, could also come from outside the word-detector. When you are in France, you expect any speaker's first words to be French; an expectation formed at a higher-order cognitive level that reasons about travel, culture, and so on. In such a case, the word-detector is doing its job as normal, but higher-order knowledge restricts a single key operation: the choice of lexicons to search.

Typical cognitive modules are likely to have at least this level of penetrability, and there will also be many cases of information flowing between modules we suspected to be rather independent. The trick then is to say something about such a module's independence. My suggestion here is to find a subset of rules and propositions that characterize this module, eliminating those that can be toggled on or off and maintaining the core body of procedures that is isolated from external modification. The word-detector probably implements a rigid search algorithm – implementing a bubble sort, say, rather than a simple sort as it searches through long lists. No instruction changes that aspect of the module's procedure. In fact, insofar as the *only* penetrable aspect of the

capacity is the lexicon choice, its implemented procedure simply has a disjunction built into it. The word-detector's core procedure is "such-and-such a procedure run on the English lexicon or French or both."

The module has a "core" procedure, as well as an associated apparatus for implementing different types of instructions depending on the particular case. This core is how the module is independent. Chess skill may not have any core at all; every bit of it can be deleted and revised by instructions or practice. Edge-detecting may consist mostly of core procedures.

The core is a rigid function, but we may not always find the system to produce rigid input-output mappings. The first reason for this is that the core function may work alongside non-modular elements. My lexicon search device implements a rigid search algorithm on the words that are in my word list. But the results it produces depend on what words are on the list. The first time I hear the French word "*jamais*", the lexicon search will return NULL. When I learn the word, the lexicon search will yield the Mentalese meaning, "*never"*. The content of the lexicon is not rigid or modular. Yet we can still study the core rigidity of the look up procedure that is constant no matter what words I have learned.

There is a second reason we may not see a rigid function produce *apparently* rigid mappings. The core procedure may be wrapped in conditional logic that is sensitive to triggers from other systems. The module will treat these instructions from elsewhere as if

they are simple inputs. Some inputs, like any individual phoneme in the stream, will constrain the ultimate output simply by setting an empty variable. Those are ordinary inputs. But other inputs, like a lexicon-choice command, will influence a particular next step in the procedure. Those are more like *instructions* than simple inputs. But the core procedure is set up with the conditional "If in France, do A; if in England, do B." So the instruction is just setting one of the empty variables for the system. Data flowing upstream from the environment and commands flowing down from the higher-order modules will simply be taken as inputs to an elaborately conditionalized computation. Yet the function can still be rigid, since in every case it performs are certain fundamental job. Whatever calculation one performs on a pocket calculator, there is a set of arithmetic functions that are perfectly constant in the machine. So while the *apparent* behavior may not be obviously rigid, looking at the broader pattern we will see how the apparent behavior is indeed rigidly following a conditional rule.

The goal of this account is to shift from looking at the information flows to the description of the activity of the module itself. The core set of rules and data deployed by the module in every case is what we should assess in looking at the module's independence. We should end up with a description of the rules embodied by the module, such as Chomsky's Universal Grammar as a description of what the syntax faculty does. It is a feature of the rules and data themselves that non-human language information triggers nothing. As a contrast, there will also be parts of the mind that are highly receptive to learning, perhaps along the lines of associationists' models. In such cases, there will be no limits at all on what the core rules are capable of engaging.

So the principle for measuring independence is the extent to which the informational core of a system is isolated. The key to independence, on this view, is that the core of the module's computation—a mix of rules and propositions or "data"—is isolated from influence by any other information at all, upstream or downstream. While it processes a fluid array of inputs, the core itself is rigid. Nothing overwrites it or revises it. Informational isolation is the durability of a module's core procedures in spite of whatever informational states develop.

Logical or de facto isolation. Understanding modules as bodies of information, we can analyze two different ways in which they may be independent: first, because the information is contingently isolated from other modules, or second, because it is necessarily so. The first case is where constrained information flows prevent the interaction between one module and another, perhaps because the two modules are simply not connected. A variation on this case is where accidental facts about the world collude to leave any connection untested. A person who never learns a second language will never test the ability of the where-am-I-detector to pass lexicon-choice instructions to the word-detector.

The second way to be independent is via properties of the information itself, the logical inter-relation of the rules and facts. Some bodies of information will contain no inter-linking rules, regardless of the channels of communication available. Just as two entirely unrelated programs on a personal computer can run simultaneously without calling each

others' functions, a module may be isolated in virtue of the eccentricity or singularity of its subject domain. Two bodies of information might exist independently in the same belief box.

The significance of these varieties is that looking only for *barriers* as the encapsulation view suggests will not be enough. We may discover logically isolated modules of knowledge or cognitive capacity that are in a widely readable format or information store. There could be modular characteristics to tasks performed by central cognition, that region of cognition which is fully connected to information flows throughout the mind. For example, complex skills like musical ability or chess playing may be modular in this way – no non-musical fact has any bearing on how a musician names a particular pitch.

As a conclusion to this section, informational isolation is an attractive way to formulate the fundamental notion of independence associated with any view of modules. This view treats modules as functions implemented by bodies of rules and data. Independence is the rigid implementation of a core set of procedures in a wide range of circumstances.

2.3 Domain-specificity

Modules are usually thought to be specialized for performing a particular job. A major benefit of modularity is supposed to be that individual modules have unique equipment for dealing with their proprietary domains of expertise, a kind of expertise that cannot be accumulated without sacrificing generality. Some theorists give this feature such importance that they take it to be the single crucial feature of modules (Hirschfeld and Gelman, 1994; Coltheart, 2000).

Domains might either be subject matters of information such as biology, formats of information such as visual data, or simply "processors" (Carruthers and Smith, 1996). To be specific to a domain means to be restricted in how many domains the module relates to, though a module might be domain-specific and still apply to many domains (but not every domain). Getting beyond the merely intuitive notion of specialized modules is a tricky task (see Chapter 5). All the present views are badly inadequate sketches for a more robust view.

The approach advocated here proposes to define domains purely in virtue of the inter-relatedness of bodies of information. Closely interconnected sets of sentences are defined as a domain where their connections trail off. A particular module is also defined as a set of rules or propositions (rules like "if wavelength X, then red", and facts like "wavelength X exists"). A module is more specific when its contents connects to fewer domains, and less specific when it connects to more. The account ignores "natural" domains like cheater-detection or folkbiology, unless these domain exhibit the right informational structure.

Domain-specificity, on the treatment I suggest, is then a merely optional feature of modules. Gall's horizontal modules, such as memory, will handle information from any domain, yet they will still have the independence properties I've suggested are essential to modularity.

2.4 Intentional Modules, Mechanism Modules

Commentators typically have two ways of describing modules, as mechanisms or as bodies of knowledge (Segal, 1996; Samuels, 2000). The mechanisms view is taken to describe engineering-style analyses of a cognitive system (e.g. Marr, 1982), where the cognitive task is broken down into discrete steps performed by simple mechanisms. Computational models are seen as elaborate versions of the mechanism view. The key contrast is with modules that consist of "knowledge" or other mental representations. These "intentional modules" are broadly in the Chomskyan vein, who first proposed to explain a cognitive capacity purely in terms of the knowledge of a domain it represented. More recently, developmental psychologists have applied this approach to thinking about a wide variety of further capacities, such as folk physics or naïve sociology. Connectionist modelers or those cognitive psychologists deploying computational models typically resist assigning any such knowledge or mental representations in explaining the capacity.

It is difficult to see how anything but a computational mechanism will implement the types of knowledge attributed by Chomskyan theories. Conversely, the computational mechanisms likely qualify as mental representations (see Chapter 3). Regardless, the literature frequently makes this distinction when taxonomizing modules.

2.5 Nativism and Developmental Regularity

Nativism is the thesis that at least some of our cognitive capacities are endowed to us before any learning or experience takes place. Closely associated is the observation that

certain capacities develop in regular and predictable ways, suggesting that they are maturing on an inborn plan. Fodor and Chomsky both advocate the innateness of the modules they identify. As a result, it has come to be seen as a constitutive feature of a modular view of cognitive architecture. If something is learned, then one doubts its modularity. There may be good empirical reasons to think this, if the only ways to be modular involve innately present conditions in the mind. This is unlikely, however, from a purely conceptual perspective (see Chapter 4).

A number of theorists have persuasively questioned the necessity that a module be innate. Learned skills have many of the characteristics of more typical modules. Color naming is a learned ability, yet the Stroop task for color-word naming shows that it is highly impenetrable. We cannot stop ourselves under certain conditions from saying "green" when a color-word like "red" is written in green ink (Stillings, 1987).

2.6 Input Only

Fodor has persistently argued that only input-systems (the perceptual faculties plus language) can be modular. This is a result of an ambitious line of argument about the limits of computational psychology. On his view, any task with high-level criteria for success such as those of rationality, simplicity, or analogical coherence cannot be implemented by a computational system. Leaving alone his main argument, we might at least discover that a wide range of non-perceptual tasks are carried out to much lower standards (Gigerenzer and Todd, 2000; Stein, 1999; Stich, 1985). In cases where we reason incompletely and systematically ignore relevant information, such as in our naïve reasoning about probability or risk, it is hard to see the impact of Fodor's argument. Such

faculties, though "higher order", are not rational and so might turn out to be modular regardless of how Fodor's argument fares.

Evolutionary psychologists have begun to argue for a wide array of higher-level cognitive modules which perform the types of tasks Fodor has attempted to prohibit (Sperber, 1994; Pinker, 1998; Barkow et al. 1992). On their view, *all* the mind's functions are modularly executed. So the literature is not at all agreed on whether modules can do more than gather inputs.

2.7 Neurally Local and Characteristic Breakdown

The classical faculty psychology assumed physically well-defined parts of the brain performed unique functional roles. Pre-modern writers typically assigned gross functional roles to distinct substances, such as the variety of humours' roles for the spirit. Gall and Spurzheim suggested that cognitive faculties were neurally localized to specific regions of the brain in all humans (1825; Gall, 1818). Most subsequent neuroscience has proceeded from this basic assumption, that functions can be isolated by studying physical damage to the brain.

The relevant point here is that results about where in the brain things happen are likely to be orthogonal to the purely cognitive level issues driven by the functional interactions between distinct operations or tasks. While most writers seem to think that functions will turn out to be neurally localized, this does not at all bear on their functional independence. Farah (1994) attacks this "locality assumption" persuasively; the results of contemporary neuroscience do suggest that processing is occurring more widely. Yet, as

Farah points out, this does not rule out "division of labor" between functional parts, and so does not bear on the principal notion of modularity we are interested in.

A closely related point is that when the functional structure of the capacity maps to discrete physical parts reliably, the breakdown structure will also be reliable. Damage to a given neuron will reliably disable a particular function in the system. This is the pattern of much neuroscience, but there are important dissensions.

2.8 Mandatory and Fast

Fodor is the principal advocate of the speculation that all modules will be mandatory and fast. Any time relevant input is presented, they will simply kick off obligatorily and rapidly calculate their result. This is closely tied to his assumption that modules exist for processing perception, and that perception *ought* to work this way if it is to work at all (Fodor, 1985). The only things, however, that are *not* mandatory and not fast are those non-demonstrative reasoning activities of central cognition. And since Fodor's model has the path from external stimulus to central cognition paved with a continuous row of computational modules, it makes sense that there be no delays on this route. If there turn out to be higher-level modules, the evolutionary psychologists will need to explain why that class of reasoning takes so long. So far, these features have played an innocuous role, though the underlying question of what aspects of cognition Turing-style computation can explain is deeply contested, mainly by Fodor (1983, 2000).

2.9 Nestable

Hirschfeld and Gelman (1994a) raise the question of whether modules can be nestable, and conclude that they cannot be. A module would be nested if one module contained another. Nothing much turns on this, as far as the polemic has gone. Unfortunately, there does not seem to be any good argument to resist nestability. If we depart from the independence-driven account I am advocating, and suggest that domain-specificity is a key feature of any module, vagaries of that concept may make it difficult for a sub-module to be domain-specific to a sub-set of the parent module's domain.

As I am advocating it here, however, independence is purely a fact about a modules relations to the external world and the "subsystem" structure means it has to have a complex interior structure. The two features together suffice for a cognitive faculty to count as a module. On the face of it, just as the mind itself is a complex structure with component modules, there should be nothing preventing modules from having similar composition.

2.10 Summary

The view advocated here is that to be a module is to have these features: subsystems, more-or-less independent in the sense of informationally isolated. The rest are optional: cognitively impenetrable, informationally encapsulated, domain-specific, innate and developmentally regular, input only, neurally local and characteristic breakdown, mandatory and fast, nestable.

The next chapter will set up the broad polemical context, and do more to illustrate several major modularist views.

Chapter 2. Modularity and Nativism in Chomsky

1. Background

Modularity is an empirical claim about the structure of the mind. It stems from the observation that our various mental faculties appear to differ in their nature and application. Characterizing our expressed faculties as the products of distinct mental entities is a way of explaining their heterogeneity; the mind is not one unitary thing, but many interacting things, and so we observe their operation as our varying faculties.

This story about mental architecture typically involves a second plank: these modules of mind are more or less innate. This is partly because the diversity of mental functions that motivates modularity is there from the ontogenetic start; diversity doesn't just appear after sufficient experience has accumulated. But it is also because these diverse mental functions appear universally, in similar form across nearly all human individuals. Such universality suggests the presence of similar, innately specified biological modules in all humans.

The conjunction of modularity and nativism is by no means mandatory, but the compulsion toward the pair is evident in contrast with the main competitor: varieties of developmental empiricism, i.e. "blank slate" psychologies typified in the present age by connectionist models. Taking anti-nativism as a starting point, such views describe

complexly structured, though not modular, minds as developing through the accretion of rich and varied experience. Experience is not considered to accumulate into logically distinct domains; experience results in something more like a "web" of interconnected beliefs. So the learning-driven conception of mind has not been thought to display any domain- or content-related divisions. Typically, blank slate psychologies are neither nativist nor modular, while nativist views also include modularity. A third option is not much explored at all: modularity is rarely offered as an architecture developing from experience (but see Karmiloff-Smith, 1992). Cowie (1999) has argued that there is a deep reason for this. The historical empiricist tradition was not just peripherally but centrally concerned with denying modularity, and the polemic between nativism and empiricism has been driven by *what* and not *how much* is within. The *what* question is to decide between a theory with one domain-general learning system and a theory with many small, specialized modules; contrary to this is the *how much* question of innate versus learned faculties which is the more conventionally characterized focus of the polemic.

The dominance of thorough-going empiricism in the form of associationism, behaviorism, and the neurologically-motivated connectionism (at least in many fields), has long left the above characterization of modularity and nativism as sufficient for contrast. The present situation, however, is becoming increasingly unstable with only these bare characterizations to rely on. This is mainly due to the transformation in the status of nativism and modularity from opposition party to governing coalition, especially since Chomsky (1959).

Chomsky (1980), Marr (1982), and Fodor (1983) presented differing yet each highly influential accounts of cognitive structure and operation that reached the heart of mainstream cognitive psychology. Modularity and nativism play key roles in the newest work on various methodological programs, investigations of particular cognitive functions, and macro-level theorizing. A diverse critical literature has developed since the early 1980s including specific challenges on how to think about modularity itself from theorists working in:

developmental psychology (e.g. Karmiloff-Smith, 1992),

connectionist modeling (Elman et al., 1996),

neurologically-oriented research (Quartz and Sejnowski, 1994),

and others fields (Sperber, 1994).

At the same time, a dense growth of theories in various domains have bloomed to apply similar arguments to research on specific cognitive faculties, such as:

theory of mind (Perner, 1991; Segal, 1996),

folk biology (Atran, 1994);

folk physics (Spelke, 1990, 1991),

naïve sociology (Hirschfeld, 1994), and other areas.

The framework has fueled macro-level theorizing about the origins of human mentality and cognitive function (Deacon, 1997; Mithen, 1996; Donald, 1993). Highest profile of all, of course, has been the deployment of extreme modularist and nativist theses alongside strong adaptationism by evolutionary psychology (Barkow et al., 1994; Pinker, 1998; Sperber, 1994; Plotkin, 1997).

The heavier burden borne by modularity and nativism has spawned a flourishing of different and mutually incompatible uses for each of these terms. The empiricism/nativism controversy is very old, and so has endured such confusions before. Nonetheless, a number of recent authors consider the proper role for innateness in contemporary psychological theorizing (Ariew, 1998; Cowie, 1999; Samuels, forthcoming; Griffiths, forthcoming). Modularity, only come to prominence in the present era of psychology and perhaps the parasitic concept, has not received similarly systematic attention. This chapter attempts to make a start at analyzing the present state of play.

2. Chomsky's Mentalism

Three crucial revivals in contemporary psychology are widely credited to Chomsky: mentalism (Chomsky, 1959), nativism (Chomsky, 1965, 1966), modularity (Chomsky, 1980). Let us begin on the big picture with mentalism--what Chomsky (1984) calls "the ontological question" about mind--and with what I take to be the related ideas of representationalism and computationalism, before descending into details about the latter two.

2.1 Anti-behaviorism

The immediately preceding period in study of the mind was dominated by anti-mentalist backlash against Cartesian dualism and its essential problem of mental causation. Rylean logical behaviorism specifically reproached any concept of mental states that made them anything more than purely relational properties. Ascriptions of "mental state or process to an organism is semantically equivalent to the ascription of a certain sort of dispositional

property to that organism." (Fodor, 1981b: 3). In practice, this meant behavioral dispositions which reliably tied particular stimuli to express responses. So "knowing a language" is not other than being disposed to respond to utterances in appropriate ways with verbal behavior of your own. On the one hand, the philosophical version of this view suffered from a number of technical problems: one with explaining statements about mental dispositions without appealing to further dispositions; and another with admitting the existence of intra-mental causal states without express behavioral results, such as with a "creeping doubt" you choose to disregard (Fodor, 1981b).

The latter finds expression in Chomsky (1959) as two deep criticisms of Skinner's program and its application to language in *Verbal Behavior* (1957). The first is that the behaviorist picture is hopelessly oversimplified; the second that this oversimplified conception cannot explain language acquisition. This latter point criticizes the familiar idea that training, correction, and rewards play a central role in conditioning language learners as they acquire linguistic ability. Chomsky cites a host of systematic errors in the behaviorist methodology to undermine this contention, along with a substantial body of evidence suggesting that such training neither happens in fact nor is required in principle for language learning to take place.

The former point, however, is that "the behaviorists' anemic conception of linguistic competence needs to be replaced by a more robustly mentalistic account." (Cowie, 1999: 162). This is argued on the basis of three points. The first is that language is *stimulus-independent*. It does not display the rigid and predictable relation between stimulus and

response expected, but rather involves internal and cognitive factors essentially in its explanation. The second that language is *productive*, or that there are boundlessly many and completely novel utterances expressible in a language, a fact that conditioning cannot explain. Finally, language is *systematic*; its innovations are highly constrained by semantic, pragmatic, and syntactic rules whose efficacy cannot be explained without appeal to the real existence of unobserved governing processes (Fodor, 1975, 1981b).

In Chomsky's middle-period work (e.g. 1980), Quine is occasionally the behaviorist under scrutiny. Quine is skeptical of explanations that rely on abstract entities lacking clear, externally observable identity criteria: "no entity without identity," as the slogan goes. Quine (1972) objects specifically that Chomsky's account cannot distinguish between rules that merely *fit* the behavior and rules that actually *guide* the behavior. For any finite set of sentences, there are always many extensionally equivalent grammars that fit the sentences. But the Chomskyan account claims that one grammar rather than another is the actual set of rules in use, explicitly without recourse to the behaver's conscious knowledge of or assent to such rules. By Quine's lights, the weakness of such a theory is just that it posits the rules not in a purely descriptive role as a demarcation of the totality of grammatical sentences, but as "themselves part of the objective linguistic reality to be specified." (105)

Chomsky, however, sees an objectively real grammar as playing just the required role: rules are a complex of unobserved entities and states whose machination *best* explains a mass of linguistic phenomena (Chomsky, 1980: 10, 12*ff.*). Indeed, Chomsky

characterizes the generic methodological considerations in favor of behaviorist anti-mentalism as no more than plain empiricist dogma. The more successful theory of verbal behavior, e.g. Chomsky's Universal Grammar, relies heavily on abstract entities. Since the success of the theory should be the primary consideration--as it is in physics, where similarly "Galilean" abstract theories make use of unobserved phenomena towards the best explanation of observation--the theory itself should vindicate the use of mentalistic terms.[3] Indeed, language should be considered the unobservable entity, since the entirety of uttered sentences does not constitute the full language. Only generative grammars are available for scientific investigation (Chomsky, 1984); the grammar is the finitely specifiable "organic unity" that is responsible for linguistic behavior (Chomsky, 1966).

One version of the behaviorist methodological critique of mentalism relies on an argument about the risk of "private languages" and the need for meaning externalism, a version of which is debated by Kripke's Wittgenstein. Chomsky connects this debate to the broader issues of the explanatory role of mental representation. Quine rejects the possibility of private language partly on grounds that it relies on a false theory of

---

[3] The behaviorists are not anti-realists about mental entities, suggesting that mental states are like Reichenbach's *abstracta*, viz. calculation-bound entities like instantaneous velocity. Rather, mental states are defined by identical with the behavioral causes and effects without remainder. Chomsky's argument shouldn't be misread as an argument about realism of these entities: in the first instance it argues that they are at least real in the manner of *abstracta*; but in the second instance, pending further results in psychology, they are probably even real in the manner of *illata* (really existing but difficult to observe, e.g. atoms). See Dennett, "The Intentional Stance" (1987) for a discussion of Reichenbach's abstracta/illata distinction.

meaning, "the myth of the museum." The museum myth is that meanings are mental objects to which names are attached, typically through experience. Quine's "gavagai" case is meant to raise difficulties for such a theory by showing that the meaning-object is scientifically unstudiable. Explanation by appeal to methodologically unobservable entities, on Quine's view, is irresponsible. An account of language-mastery should be an account of the observable facts of experience. Chomsky considers the Quinean argument as one proposal for avoiding the risk of private language. Denying the museum myth for this methodological reason, in Chomsky's assessment of Quine's argument, implies a general argument denying mental objects categorically, since they can never have external individuation conditions.

Chomsky (1980) takes the above methodological argument against meaning-objects to have a strong and general implication for psychology. If a theory cannot trade in meanings that are in-the-head to explain language, then equally it cannot trade in mental representations for explaining any other faculty. Any psychological explanation that appeals to the existence of an internal mental state is vulnerable to an analogous "museum myth" argument. For example, Marr's account of how the vision system detects the edges of objects relies on a number of internally stored and represented rules for handling visual input data. Certain conditions are systematically treated as edges while others are not, sometimes cross-cutting the distinction between real and false edges. The explanation for this systematic behavior thus cannot be that the world's edges directly trigger "edge-sightings" in the vision system; the edge detector systematically misidentifies certain cases. The sightings correspond to a rule about edges that is stored

in the head, not to any naturally distinct set of cases in the world. These rules are taken by Marr's theory to be internal mental objects that interact with data transduced from the external world.

According to Chomsky, the argument that rejects meaning objects must also reject Marr's edge-detection rules. Just as the Quinean can demand externally-observable criteria for the existence of a particular meaning-object, he can demand that the edge-detection rule be observable. Thus, the methodological component of the private language argument forces out any psychological theory requiring mental representations, an unacceptable result. The best theory of how the vision system identifies edges relies on there being some physically-incarnate mental structures of a particular sort. How it manages to be implemented is not essential to the explanation. What is essential, though, is the existence of some such mental object which interacts with the incoming data. Yet there is not yet direct evidence for the existence of some material that embodies these rules, or perhaps there can never be empirical evidence for a rule (as with Kripke's Wittgenstein). So Chomsky's argument is that if the Quinean behaviorist line denies private languages *tout court*, then so too does it deny all mental representations in psychological explanation of any capacity.

Fodor (1985c) describes this line without direct appeal to a concept of representations. Mental faculties are distinctively cognitive in virtue of how "smart" they are. The basic material of behaviorist conditioning, reflexes, are simple mappings from stimulus to response. They are *noninferential* in the sense that they involve no intermediate states

elaborating the relationship between input and outputs. They are not “smart”. Cognitive faculties such as perception, on the other hand, are inferential precisely because "a lot of inference typically intervenes between a proximal stimulus and a perceptual identification" (Fodor, 1985: 197).[4] How should we interpret this claim about *inferential* faculties?

At a minimum, Fodor's statement requires that an input pass through at least one step before transformation into an end-condition. But even the simplest human reflexes involve the serial operation of more than one internal steps[5]. Fodor's appeal to intervening inferences should be taken, instead, as requiring elaboration on the input, perhaps by bringing additional information or rules to bear. Marr describes the vision system's edge-detection as a process which adds visual stimulus to the antecedent rule that "such-and-such light pattern is an edge" to infer the existence of a particular edge. In Chomsky's case, to borrow from Searle (1974), sentences such as "I like her cooking" present no surface structural cues to disambiguate among their various possible meanings (e.g. I like the food that she cooks, I like her to be cooking, I like the fashion in which she

---

[4] Consider even a simple reflex like a spinal reflex. The sensation bit itself is at least several steps (see next note). More importantly, it is at least a few more inferential steps to *identifying* that impulse as “in my leg” and “caused by the needle”. Those additions are not part of the single long neuron connecting kneecap to spinal cord.

[5] Very simple reflexes, such as spinal reflexes, go only from the nerve ending to the spine (well short of the brain) and then return a signal via a motor neuron to the muscle. The entire loop involves just a few neurons: the sensory neuron, the interneuron, and the motor neuron. Even so, this simplest case is complicated enough to be more than a single-step link.

cooks, or even, I like the fact that she is being cooked). Rather, their perception requires the selection from one of several models of their underlying syntactic structure; the very existence of this extra step meets Fodor's requirement. The substance of this extra step is the contribution of the pre-existing mental representation, an entity methodologically outlawed by the pure Quinean behaviorist line.

In a separate Chomskyan line of argument, evidence for such "smart" processes follows on "arguments from the poverty of the stimulus" (APS), a distinct line of argument from the methodological considerations so far described. The methodological considerations suggest that no successful model can be constructed without positing internal states. Unlike these, the general direction of an APS is to suggest that observed external inputs simply do not provide the relevant inputs for completing the types of inferences we see people make. As such, there must be some mental states, which exist before or independent of any external phenomena, in position to aid in perception of linguistic input. A behaviorist theory of mind can neither use internal states to explain complex mechanisms nor to explain the origin of unlearned knowledge.

The upshot of Chomsky's arguments is a revival of mentalism, a view that essentially treats mental states as having real, intrinsic properties, contrasting with the materialist view that the study of mind is just the study of physical states like behavior[6]. That these

[6] Psychological mentalism here is contrasted with psychological materialisms like behaviorism; mental realism could be another name for the view. It is not meant to be confused with Berkelean immaterialism or

states are intentional in character is a point to come later, but it is the basis for Fodor's campaign for Intentional Realism. Mere mentalism (or mental realism) is a bare commitment, which Chomsky shares with full-bore Cartesian substance dualists (they also think the mind is a real, causally efficacious object of inquiry). The more refined doctrine that is presently with us is more appropriately called "cognitivism"--the doctrine that there are psychologically real cognitive states and processes at work (in the brain) which are the subject of cognitive psychology and the other cognitive sciences (a view which permits that there are perhaps other, non-cognitive mental states as well). This view is also typically representationalist, which just means that psychology's cognitive states have worldly content and that operations on those states explain the function of our cognitive capacities.

2.2 Chomsky's Representationalism

The first main point about mentalism, just discussed, is that Chomsky's revival was anti-behaviorist and the basis for modern cognitivist psychology. The second main point is that at least some cognitive function—the system underlying language competence, in particular—is representational and also likely computational.

A computational system, such as a digital computer, implements rule-governed formal transformations on syntactically-structured representations (Fodor, 1975; Haugeland, 1985). Classical cognitive science (Fodor and Pylyshyn, 1988) takes the idea that

---

idealism, also called philosophical mentalism, the idea that *only* mental states are real objects of inquiry and that all else is illusory.

cognition is computation to be a fundamental insight, an idea that Fodor and others attribute first to Turing. This requires a theory of the mental that is stronger than simply adding bare mentalism to materialism, a move that simply ends up looking like Cartesian substance dualism. It has to be stronger in two ways. First, Turing's idea requires that the mental states are symbolic representations, perhaps positing brain states that have semantic content in virtue of their causal connections to the world. This is how mental states count as real intentional states (and not just as states that are consistent with an intentional stance, as on Dennett's view). Second, it requires that the interaction of these symbols fall under purely formal rules, as opposed to rules that depend on what the symbols actually mean. Such a system can, for example, physically represent any statements of the form P→Q and P, and then infer from their structure that Q. So claiming that the mind is computational is claiming not only that the mind traffics in representations, but that its processes are entirely explained as rule-governed interactions between these representations.

Chomsky's critique of the methodological behaviorism of Skinner and Quine can be taken to favor a computational view for at least some parts of the mind. Language acquisition and use is best explained by appeal to mental representations of linguistic rules and concepts as entities independent of linguistic experience (Carruthers and Botterill, 1999). Most famously, this is the case for a Universal Grammar, or a set of syntactic rules for the construction of valid linguistic expressions. But Chomsky also suggests that there are proprietary rules for the appropriate interpretation of phonological input that allows us to parse and recognize speech, and that there is likely to be some

special function semantic system that explains the remarkable acquisition speed and size of our vocabularies (Chomsky, 1980:54). Marr's (1982) account of vision and Newell and Simon's (1972) account of heuristic reasoning also put forward empirically successful theories which rely on computational models of cognitive functions. Their empirical success at modeling the relevant competence is a formidable, non-demonstrative consideration in their favor considering the failure of behaviorism in these areas.

The computational view is the major option as a representational theory of how the brain implements a cognitive system, but it is not the only one. Connectionist neural networks are also representational (Fodor and Pylyshyn, 1988; Kosslyn and Hatfield, 1984; Hatfield, 2001, 1991). On these models, however, representations are not each discrete, physical symbols but global states of the system. While there may be features of the psychological phenomena that favor one view over the other--e.g. compositionality of language, as Fodor has repeatedly urged--the general arguments for representationalism as against behaviorism are not of this nature.

The Chomskyan story about language can be happily agnostic about the mechanisms by which it is implemented. The intentional states this story requires need only be representations constituting facts about linguistic syntax or phonology. How they manage to function in this way is beside the point. (For a dissenting view, see Buller and Hardcastle, forthcoming.)

Non-representationalist theories of mind which are otherwise consistent with Chomsky's arguments in that they do posit internal states, on the other hand, do seem to pose a problem. If it is possible to reject behaviorism and be a mentalist without accepting a theory of mental representations, we should re-examine the Chomskyan arguments. There are some contemporary views that attempt this line. Van Gelder (1995) offers that dynamical systems theory presents just such a challenge. On this view, a complex system, such as a Watt Governor invented to regulate steam flow in 19th Century railroad engines, can be accurately described without any appeal to representational states.[7] The mechanism works to regulate steam flow without having elements that specifically represent aspects of the operation to be performed. (By contrast, a modern air conditioning system has specific parts corresponding to temperature and air flow.) Van Gelder's point is that a mind could be similar, a complex system whose mentalistic properties simply spring out of a mechanism with no correlates for the mental states they implement.

It is unclear, however, that this dispenses with more than just the language of representations (Chemero, 2000). One important caveat to the issue of representationalism involves the realism of the computationalist picture. Van Gelder (1995) raises the issue of what it is to describe cognition as computation, in responding to

---

[7] The Watt Governor is a mechanism invented during the 19th century to regulate the flow of steam through a rail engine in proportion to the desired speed of travel. The operator moves a lever to a certain position, and the appropriate amount of steam flow is permitted. But this is not a directly proportional relation. Cf. Van Gelder (1995).

which Chemero (2000) distinguishes an ontological from an epistemological point. The former is a point about the nature of mental states. Classical cognitive science maintains that features of the brain itself can be shown to have the structure of a Turing machine. The latter point, however, suggests only that such structures are the most effective models for understanding cognition. Separating the two theses, a computational theory may provide the best explanation for the mind's operation by use of theoretical entities that do not exist anywhere in actual heads. It seems likely that Chomsky's view is the latter type, at least at present, though Fodor has repeatedly urged that the absence of alternatives is to be taken as evidence for the ontological accuracy of the computational theory of mind.

The representational theory of mind, and even the strong thesis of the computational theory, are key thrusts of the early Chomskyan critique of behaviorism and now lay at the heart of broader cognitivist psychology. Even where connectionists or other objectors claim differences with the classical computationalist picture, nearly everyone ends up supporting a theory of mind where mental states are implemented by physical symbols bound together in rule-governed interactions.

2.3 Summary

So far we have reviewed a number of arguments due to Chomsky regarding the nature of mind and how best to understand it. On this view the mind consists of syntactically-structured, semantically-evaluable, internal mental states and their interactions. Insofar as his work has set the agenda for certain lines in cognitive psychology, many present debates take basic anti-behaviorism, cognitivism and representationalism as common

ground in their psychological theorizing. For this reason, the present review is useful in understanding the more contemporary arguments.

This approach is limiting, however, in that significant issues remain unsettled by Chomsky's work. For example, while he makes extensive arguments for the existence of intentional states, their physical nature as biological entities is left completely open. Even on the psychological level, Chomsky himself does not help us decide between computational, connectionist or other representational theories of mind. To those who have adopted the Chomskyan paradigm for explaining mental competences, the issues of biological constitution and computationalism remain serious concerns. The situation with respect to nativism and modularity will be similar.

3. Nativism

The concept of innateness in psychology is typically introduced in contrast to the nativist's opponent, "blank slate" empiricism in the form of associationism, behaviorism, or connectionism (Cowie, 1999; Fodor, 1983). Where the latter emphasize the role of learning and experience in the acquisition of mental structures, the nativist thesis highlights the limits of this approach. Nativism is, minimally, the thesis that at least some mental structures are partly endogenously specified and not entirely the products of external stimuli. Innateness, then, means being endogenously specified. This sort of independence can be demonstrated by "isolation experiments" which involve separating an organism from its normal environmental stimuli and observing the inhibited development of the relevant cognitive faculty, a strategy Lorenz emphasized in his foundational work in behavioral ethology (Lorenz, 1963; Ariew, 1999).

For clarity, we should distinguish explicitly between two types of empiricism, since nativism is so often characterized merely as anti-empiricism. *Developmental empiricism* is the blank slate view that identifies learning as the source of all knowledge. This is the view relevant to the present discussion. *Epistemological empiricism* is the other variety that denies the possibility of *a priori* knowledge. Famous historical empiricists typically espoused both views. In the *Essay*, Locke argues for the developmental thesis as a means toward the epistemological one. Nativism and empiricism as *developmental* questions are the issues of this paper, since psychology is concerned with explaining how capacities come to be and how they work.[8]

The simple characterization of nativism above is inadequate for most uses, as many commentators from different perspectives have pointed out (Samuels, 2002; Griffiths, 2001; Cowie, 1999; Ariew, 1996, 1999; Wimsatt, 1999; Bateson, 1991; Oyama, 1990; Stich, 1975). Complications to the treatment of nativism in biological disciplines outside of psychology have left a very difficult situation within cognitive psychology. Some critics have attempted to reduce the confusion by elevating one aspect of the nativist position to the foreground. Samuels (2002) claims that innateness is a concept used to describe psychologically primitive phenomena, those which cannot be explained by internal features of scientific psychology. Cowie (1999) suggests that domain specificity is what distinguishes a paradigmatically innate mechanism from an empiricist one. Frank

[8] The philosophical impulse to read empiricism and nativism (or "Rationalism") as theses about epistemological justification should be avoided in what follows.

Keil (2000), in a review of Cowie's book, argues that nativism names a multifarious collection of distinct concepts. Griffiths (2001) urges us to dispense with innateness altogether, as it is both "irretrievably confused" and an artifact of our own folk essentialism.

The present confusion around the concept of innateness will not help our discussion of Chomsky's nativism. However, Chomsky's arguments for nativism are largely responsible for the current interest in developmental theories appealing to internal structures. Fodor interprets his view as one about propositional attitudes. What the child knows is most closely analogous to beliefs, and if not quite beliefs then cognized propositions. Though this is consistent with Chomsky's way of putting things, it does add more technical precision than is already there. In elaborating his own view, Chomsky sees an intimate involvement with the historical nativist tradition. Understanding the view he has developed is crucial to finding a bearing on nativism's present situation and its origins.

The aim in this paper is to look at Chomsky's own use of nativism and modularity, and look at the standard arguments for them. This is meant to be useful exactly because Chomsky's usage has been so influential, so it would be unproductive to refrain from taking a view about the key concepts. A basic thesis underlies all variants of nativism, and that is the view I advocate as the core formulation of the view: at least some mental structures are endogenously specified and therefore present before learning or experience take place. (See Chapter 4 for a detailed discussion of this account.) This view will guide the discussion to follow.

Moreover, there is a polemical goal in the way the discussion is structured. The single classic argument for nativism popularized by Chomsky is the Argument from the Poverty of the Stimulus. I describe a broad historical tradition and varied family for this argument. Cowie (1999) has recently made a case for distinguishing a second important type of argument for nativism: the Impossibility Argument, which focuses on the unlearnability of a type of knowledge. An aim of this chapter is to suggest a third and equally significant body of argument: the Argument from Fixed Capacities. This argument focuses on the universality of certain features of human cognitive systems. For this argument, the focus can be variously on their common architecture, developmental path, distribution of specialties, or yet other shared features. I will argue below, however, that it is worth distinguishing and carries its own proprietary line of argument.

3.1 Historical Sources

In presenting his arguments against developmental empiricism, Chomsky has repeatedly appealed to certain historical sources: Descartes and Leibniz chief among them. From this Rationalist tradition, Chomsky draws not only the *a priori* existence of particular bodies of knowledge, but also certain other features of the mind, including the anti-mechanistic and creative nature of language use. In the following sections we look at the main historical sources for Chomsky's "Cartesian linguistics" as well as the leading metaphors invoked in explaining the concepts.

Chomsky claims that a significant part of language acquisition is due to innate, purely internal structures. The argument for this claim requires at least two steps. The central

claim of his *linguistics* is that "the general features of grammatical structure are common to all languages" (Chomsky, 1966: 59). That is, there is a single Universal Grammar underlying all natural human languages. The further, *psychological* claim is that these features "reflect certain fundamental properties of the mind." Here, the fundamentality of these features is precisely that they are "not learned" or "innate"; these features must exist "if data is to lead to knowledge" (59-60). The Universal Grammar is itself innately cognized, if only partially.

The linguistic claim evokes a position familiar from historical nativism: some particular feature of our mentality is identified as an object of universal consent or common endowment. Nearly all humans are said to believe in God, or such-and-such morality, or that basic mathematical claims are true, or that phrase-structure dictates syntactic construction. That very fact is taken as evidence for the innateness of the mental feature. An argument from universal consent is not decisive on the subject of innateness. One problem is that it does not rule out non-nativist explanations, e.g. everyone believes that unsupported objects fall toward the earth, though this may not be innately known, as Boyer, 1994 has pointed out). A second problem is that universality claims usually face exceptions, such as children or foreign peoples, whose ignorance of certain knowledge must be explained. But the argument from universal consent does set the stage for stronger arguments.

The following sections review the main lines of argument for nativism, in the Rationalist tradition and also in Chomsky. The historical arguments for nativism generally fall into

three categories: poverty of the stimulus (APS), impossibility of learning, or the growth of fixed capacities. For the first two, I follow Cowie (1999). In distinguishing a third argument, I offer a different interpretation of extant arguments with the suggestion that they should not be lumped together with APS and impossibility. Chomsky provides arguments principally from APS and fixed capacities.

3.2 Poverty of the Stimulus

One simple way to divide the sources of knowledge is between "internalist" and "externalist" (Godfrey-Smith, 1994). Some explanations of psychological phenomena will appeal to facts *external* to the mind. Empiricism as a developmental doctrine is paradigmatically externalist in this way, emphasizing the role of the environment in stimulating sensations. The traditional empiricists, such as Locke, do not deny that there is some psychological structure receiving this input, and even some basic faculties; but, overall, they think "there is nothing in the mind that was not previously in sense" (Godfrey-Smith, 32) and that inputs develop into mental structures purely due to the pattern of sensory stimulation. To the nativist, it is little more than a faith that stimuli will "just coalesce into beliefs" (39).

The internalist gives the internal psychological structure itself an important role, in virtue of some innate character it has prior to any interaction with the environment. This innate character could be many things. It might simply be knowledge or ideas that a mind naturally possesses. But it is sometimes described as a set of dispositions to particular beliefs, or as "structure" that organizes the raw incoming data into ideas. Admittedly, the simple distinction we will use here leaves many issues about *what* is inside the organism

unaddressed. Nor does it mean that empiricism is perfectly externalist; proponents of any empiricist or nativist view will appeal to internalist and externalist elements in their explanations of development. The emphasis is what is different. Some of these issues will be addressed later.

The interest of this section is to catalog the types of arguments *for* the nativist conclusion. The internal/external distinction is useful for this purpose. One obvious way to argue that an idea has an internal origin is to say that it does not exist anywhere external to the psychological system. Essentially, this is the Argument from the Poverty of the Stimulus. The nativist examines the external world for a particular idea. If the world can be shown to lack this concept contingently or necessarily, then the concept must come from the mind or from whatever created it. The Poverty argument turns on a claim about the world, that it contains no source for a particular concept. Taken with the basic internal/external dichotomy, it constitutes an argument for innateness.

3.2.1 Historical Poverty Arguments

Plato's argument from the *Meno* is the oldest source in this argument's history, and Chomsky frequently calls the problem of language acquisition simply "Plato's Problem". In the *Meno*, Socrates demonstrates that a slave boy can answer questions about geometry if questioned appropriately. Since the slave boy has never received instruction, it must be that these ideas were within him from birth, waiting to be remembered with the aid of Socratic questioning. This version of the argument appeals directly to the unavailability of relevant empirical information. The boy has received no instruction in geometry, which Plato takes as sufficient to show that he could not have learned certain

sophisticated facts about geometrical concepts (i.e. the Pythagorean Theorem holds for right triangles).

Descartes makes the argument from the poverty of the stimulus in several places. He uses it to establish that at least some ideas are innate, where ideas can be thought of as concepts. To do this, he argues that worldly phenomena are too impoverished--and our concepts about them too rich--to be the origins of our knowledge. Two famous cases concern the idea of God and the idea of triangles.

For each of these, Descartes' burden is heavier than Socrates'. It is not enough to deny that no one has taught us the particular concept. Call this simple denial the poverty of *instruction* argument, which allows that the concept is out there but not explicitly communicated to the subject. Implicit in Socrates' poverty of instruction argument is that we cannot *learn* such sophisticated facts any other way. Encountering triangles or even knowing what triangles are is not alone sufficient.

For the concepts in Descartes' formulation, however, it seems that the relevant concepts are out there and that one could acquire them from the world *without instruction*. They are not so sophisticated. That is, raw experience of the features of the world--powerful thunder claps or the magnificent adaptedness of species--might be sufficient to cause the idea of God within us. Descartes needs to make a poverty of *environment* argument, a claim that the world does not contain anything which could cause the idea or belief.

Descartes' arguments about God are presented in the Third Meditation and in *Comments on a Certain Broadsheet*. The argument takes two forms. In *Comments*, he claims that the senses can only provide ideas via pictures and sounds (apparently the other senses do not provide ideas, for Descartes). But we know much more about God than can be learned by paintings or utterances of his name. So, it must be that "everything over and above these utterances and pictures which we think of as being signified by them is represented to us by means of ideas which come to us from no other source than our own faculty of thinking" (Cowie, 1999: 34; Descartes, 1985 I:305).

The second version, from Meditation Three, relies on the "formal reality" of the idea of God. Very simply, an idea cannot be the result of a cause with less reality or perfection than it. Since the idea of God is supremely perfect, it cannot have the world or Descartes himself as its cause. Each are too far imperfect. So it must be God himself who is the cause. But then Descartes concludes that it must be that "in creating me [God has] placed this idea in me" (Cowie: 35; Descartes II:35).

Descartes runs a similar argument about mathematical ideas, featuring triangles in particular. Essentially, he argues that there are no perfect triangles in the world. Both Chomsky (1966) and Cowie cite parts of the following passage:

> although the world could undoubtedly contain figures such as those the geometers study, I nonetheless maintain that there are no such figures in our environment…Hence, when in our childhood we first happened to see a triangular figure drawn on paper, it cannot have been this figure that showed us how we should conceive of the true triangle studied by geometers… (Cowie: 36; Descartes II:262).

We cannot gain the idea from experience because the authentic triangle is simply unavailable in nature. Indeed:

> the true triangle is contained in this figure, just as the statue of Mercury is contained in a rough block of wood. But because we already possess within us the idea of a true triangle, and it can be more easily conceived by our mind than the more complex figure of the triangle drawn on paper, we, therefore, when we see that composite figure, apprehend not it itself, but rather the authentic triangle (Descartes, cited in Chomsky, 1966: 69).

Locke and the empiricists have typically responded to this argument by appealing to a faculty of abstraction or generalization. From many imperfect triangles, this faculty can draw out the essential form of the triangle. An approach like this has also been tried in the modern discourse. Chomsky thinks linguistic knowledge is already possessed for a reason very much like Descartes', because linguistic data alone does not distinguish between correct and incorrect rules. Contemporary empiricist arguments--relying on the connectionist extraction of statistical information encoded either in normal linguistic experience or in Motherese--have taken a Lockean approach in response. Connectionist models have been designed to demonstrate abstraction from assorted linguistic inputs to rules (Rumelhart and McClelland, 1986; cf. Pinker 1999 for a review of the "past tense" debates of the late 1980s).

3.2.2 Chomsky's Poverty Arguments

Chomsky argues that primary linguistic data (*pld*) is insufficient for learning the rules of grammar. There are “vast qualitative differences between the impoverished and unstructured environment and the highly specific and intricate structures that uniformly

develop" (1983: 34); and while a child needs nutrition to grow, it does not come to resemble its food. The external world is too impoverished to explain what develops in the mind, so something inside must do much of the work. One version of this argument begins with some specific rule or fact about language. It then examines the likely *pld* to conclude either that there was no relevant data, or that the data cannot discriminate between several correct and incorrect rules.

Cowie (1999) discusses the famous example of polar interrogatives, or yes-no questions formed from declarative sentences (also Chomsky, 1988:1-35; Pinker, 1994:40). The formation of these sentences requires attention to the sentence's internal phrase structure. The child hears these declarative and interrogative forms during the process of learning language:

(1a) Ali is happy.

(1b) Is Ali happy?

From this evidence, the child might infer that one of several possible rules governs such constructions. One rule simply observes the surface structure changes involved in converting (1a) into a question:

(H_1) Move the first occurrence of "is" to the front of the sentence.

A second rule might specify the conversion with respect to the underlying phrase structure of the sentence:

(H_2) Move the first occurrence of "is" that follows the subject noun phrase (NP) to the front of the sentence.

Both rules dictate moving "is" before "Ali".

H_1 is not the right rule, as its application to trickier cases shows:

(2a) The book that is on the table is blue.

(2b) *Is the book that on the table is blue?

Rule H_1 chooses the wrong "is" and forms an ungrammatical sentence. The phrase structure rule gets it right by observing the NP:

(2a) [$_{NP}$ The book [that is on the table]] is blue.

(2c) Is [$_{NP}$ the book [that is on the table]] blue?

Perhaps some other rules could also give the right construction here, and present linguistic theory has modified this example's somewhat old analysis, but that is not the essential point.

The essential point is not about how the child constructs the question, but about how the child *learns* to make the construction: "the fact that without instruction or direct evidence, children unerringly use computationally complex structure-dependent rules rather than computationally simple rules that involve only" the position of particular words in a linear ordering (Chomsky, 1986b: 7). The available *pld* makes H_1 a perfectly reasonable candidate for the correct rule. But children *never* say anything like (2b). If the child were learning H_2 from experience, some children would at least sometimes entertain H_1, a very simple rule that requires no overhead phrase structure concepts. The fact that children do not entertain H_1 cannot be explained by anything in the stimulus. Indeed, it must be that something--perhaps the NP concept--is already available prior to any *pld*. Research confirms that children do not make such errors (Cowie, 1999; Crain,

1991:602; Crain and Nakayama; 1987). There has been research to confirm also that parents do not provide specific instruction (positive or negative feedback) for the vast majority of cases; they seem to ignore the child's errors (Pinker, 1994).

In addition to the specific cases constructed around particular syntactic rules, there is the family of remarkable pidgin language cases reported in Bickerton (1981) and Singleton and Newport (1993). Bickerton compared two groups of migrants to Hawaii. One group emigrated as adults and spoke only a broken, limited form of English after many years on the island. The other group arrived as children and learned this pidgin English as they were growing up. Their adult language, Hawaiian Creole English, had a "systematic grammar utilizing many of the syntactic elements that Hawaiian pidgin lacks, including articles, auxiliaries, sentential embeddings and relative pronouns" (Cowie, 1999: 302). The pidgin language became *creolized* by the children that grew up with it. Similar results have been shown for deaf children raised among hearing communities; they develop sophisticated sign languages with many of the features of spoken natural languages or official sign languages (e.g. American Sign Language), all without anyone to teach them and without any auditory linguistic input. These cases highlight the rules that the children acquire in a context where the *pld* does not even observe these rules.

These arguments are intended to demonstrate both poverty of *instruction* and poverty of *environment*. With polar interrogatives, linguists have argued that parents do not instruct children with these rules. It is impossible that they should, as most adults cannot even state rule H_2 upon reflection. This is a perfectly empirical claim, of course. The

interrogatives also show that the available evidence cannot teach the correct rule. Many possible rules are consistent with the observed evidence; the observed unerring learning is not explained by what is available externally. The same goes for another famous example from developmental psychology, Jean Berko's (1958) study of children's invention of plural forms for novel words. When shown one animal called a "wug", the children are asked what to call a small group of the animals. Their invention of the term "wugs" is roughly similar to the formation of interrogatives above: since the word is novel, there has not been any instruction; and since plurals are unpredictable, the relevant rule is not obvious. Yet children reliably add "s" rather than add "es" or change the ending to "x".

The creolization cases are even stronger. The pidgins do not even exhibit the correct rules, as the *pld* of (1a) and (1b) does. The pidgin *pld* is full of ungrammatical sentences. The children go *beyond* the *pld* by "creating" or "inventing" new grammar rules, thereby enriching the language. This cannot be due to any instruction or cues from linguistic data since there isn't any (the deaf child Simon from Singleton and Newport's 1993 case, in particular, had no interaction with other deaf persons).

Linguistic arguments based on a wide variety of phenomena have been presented to support a similar conclusion. The general point is that the *pld* fails to display features of the grammar exhibited in the behavior of language users. A result of Chomsky's program is that the basic discontinuity between surface and deep structure ranges across linguistic

data in principle. It is not an accidental feature of some specialized cases. As Bever (1974) puts it:

> the calculus which defines the sentence-space involves formal devices which appear to have no physical basis at all, and thus to be attributable to structures within the child's mind…the crucial fact is that *no* sentence form represents explicitly its inner form; thus any theory which requires that inner knowledge be extracted from explicit data cannot explain the existence of a transformational grammar in the human mind. (149)

Though in some respects vague, these arguments are taken to have a specific output: the innateness of intentional states about language. For Plato and Descartes, the conclusion of the Poverty arguments is that the mind innately contains what can literally be called *knowledge*. Chomsky thinks this is perfectly natural as an account of what is innate in the case of language, calling nativism one of the "two general lines of approach to the problem of acquisition of knowledge, of which the problem of acquisition of language is a special and particularly informative case" (1965: 47). Under criticism from Kitcher (1977) and others, Chomsky (1980) wheels back off this strong line, preferring the epistemically neutral (but still cognitive and intentional) attitude "to cognize" rather than "to know". These true beliefs about language which children cognize are loosely considered tacit knowledge, though I will not investigate this categorization here (see Chapter 3). It is relevant, however, that the innateness argument takes psychological objects as part of its conclusion, intentional states like knowledge or at least cognized propositions. Some careful commentators prefer to treat Chomsky's results as more neutral, implying only "pre-existing structures" (Godfrey-Smith, 1994). But the historical discourse and Chomsky's own comments occur entirely within the psychological level--

there is no suggestion of sub-psychological mechanisms or "bins" for automatically sorting incoming data. As such, we should take the Poverty arguments as explicitly claiming the existence of intentional states, not merely vague "structures".

3.3 Impossibility Arguments

Let us return to the broadly "internalist" characterization of nativism suggested by Godfrey-Smith. An argument for nativism is an argument for the internal origin of some or all knowledge. The Poverty argument focuses on the external environment of the organism during development. Socrates and Descartes are concerned with what there is in nature, and what sorts of concepts it can reasonably be said to contain. The difficulty with this argument arises when an argument for the poverty of the environment is not strong enough to establish that the concept is unlearned, as is the case with Locke's proposal that abstraction distills perfect geometrical triangles from imperfect drawings.

An alternative approach for the internalist focuses on how ideas get into the mind. Like the Poverty argument, it begins by identifying mental contents that at least some people possess. It is then argued that this content--concepts, beliefs, knowledge, or what have you--cannot possibly be learned. This approach, the Impossibility argument, relies only on facts about the nature of the mind and the nature of the possessed knowledge, concluding from this that such knowledge is not learnable.

Consider a different case for internalist/externalist explanation. Some under-aged students are discovered in a nightclub after it has opened for businesses. How did they get in there? There are two broad types of explanation. The externalist option suggests that

they came from outside the club. They are like any other of the nightclub's guests, off the street and in one of the doors. The internalist tack suggests that they are somehow affiliated with the club itself—perhaps they are with the DJ, or the bartender, or some other of the staff. They were there before the club opened, and only now have been detected.

To demonstrate the students' internal origin, we might appeal to a Poverty argument. While students are quite unusual inside this nightclub, they are equally unusual outside the nightclub. There are no universities around, and so it seems quite unlikely that any passersby outside might have been students. Given that, it must be that the students are affiliated with someone inside the club especially. In contrast to this type of argument, an Impossibility argument moves focus away from who is outside the nightclub. Instead, the emphasis is on the conditions of entry. Here we might emphasize the club's rigorous door check policy, and its consistent record of preventing under-aged entry. Perhaps all other doors are kept locked. If no students could have entered once the club opened, they must have been there from the start. How exactly they could have been there all along is unclear, since the Impossibility argument establishes only that any externalist explanation must fail.

3.3.1 Historical Impossibility of Learning

As Cowie (1999) describes this second type of argument, when applied to psychological states, the Impossibility argument denies the very possibility that any concepts are learned. The Impossibility argument is consistent with the Poverty argument, though they proceed from independent premises. Plato's theory appears to enlist both types of

argument. In the *Meno*, Socrates invokes his theory of *anamnesis* and his demonstration with the slave boy as a response to a paradox cited by Meno (80d-e). Briefly, the paradox is that "if you don't know what you're inquiring after, you won't be able to recognize it when you find it; but if you do know what you're inquiring into, your inquiry is superfluous" (Cowie: 13). Socrates here concedes to Meno that acquiring the concept of justice by experience is indeed impossible, then adding that we need never learn ideas at all but only remember them. In effect, Socrates concedes an Impossibility argument against the learnability of justice. Recall, though, that he uses the same slave boy to argue for the Poverty of the Stimulus in the boy's demonstrable knowledge of mathematical concepts. In a discussion in the *Phaedo*, Socrates again argues that certain knowledge we possess could not possibly have been acquired via the senses, then inferring that the soul must exist before life begins.

Descartes also argues that some knowledge cannot be attained sensibly, again using triangles as his example. From Descartes' *Notes Directed against a Certain Program*, Chomsky cites the following remarkably modern passage:

> …any man who rightly observes the limitations of the senses, and what precisely it is that can penetrate through this medium to our faculty of thinking must needs admit that no ideas of things, in the shape in which we envisage them by thought, are presented to us by the sense. So much so that in our ideas there is nothing which was not innate in the mind, or faculty of thinking, except only these circumstances which point to experience--the fact, for instance, that we judge that this or that idea, which we now have present to our thought, is to be referred to a certain extraneous thing, not that these extraneous things transmitted the ideas themselves to our minds through the organs of sense, but because they transmitted something which gave the mind occasion to form these ideas, by

> means of an innate faculty, at this time rather than at another. For nothing reaches our mind from external objects through the organs of sense beyond certain corporeal movements…but even these movements, and the figures which arise from them are not conceived by us in the shape they assume in the organs of sense…Hence it follows that the ideas of the movements and figures are themselves innate in us. So much the more must the ideas of pain, color, sound, and the like be innate, that our mind may, on occasion of certain corporeal movements, envisage these ideas, for they have no likeness to the corporeal movements…I should like *our friend* to instruct me as to what corporeal movement it is which can form in our mind any common notion, e.g., the notion that *things which are equal to the same thing are equal to one another,* or any other he pleases; for all these movements are particular, but notions are universal having no affinity with the movements and no relation to them (Chomsky, 1966: 66-7; Haldane and Ross, 442-3).

Here Descartes lays out an entirely adequate argument for innateness: (a) our only experience comes through our senses, (b) senses can only carry particulars, (c) but ideas are universals, which do not resemble these particulars, (d) so our ideas must be innate. Indeed, *all* our ideas are thus innate.

Leibniz issues some related arguments as well, many drawing directly on Descartes. One argument, relying on his monadic metaphysics, is of less interest to us here. He argues that mental and physical substances are causally independent, and cannot interact. As such, it is not possible that the world can imprint ideas upon the mind. This directly rules out learning (Cowie, 1999: 50-52).

There is a second argument, in the *New Essays*, that at least some ideas are innate. Leibniz divides knowledge into intellectual or necessary truths and truths of fact. The former are those knowable by reason alone, such as mathematical or geometrical truths (Broad, 1975). We would not come to think them unless prodded by sensible experience, but their truth "must have principles whose proof does not depend on instances nor, consequently on the testimony of the sense, even though without the sense it would never occur to us to think of them" (50). The justification of these beliefs can only be internal, and that is what demonstrates their innateness: "the way for [innate principles] to be rigorously and conclusively proved is by its being shown that their certainty comes only from what is within us" (76). Leibniz takes for granted the idea that there are such intellectual truths, which are entirely *a priori* in justification and are the issuance of "the light of nature" or, as Locke calls the faculty, "Reflection" (52). Once we know that such truths exist, then we know that something must be justifying their truth. If he can have the supposition that such truths are justified by reason directly and not by sensible experience, then his argument appears to show that the existence of intellectual truths demands the existence of innate knowledge of these truths.

Both these arguments can be interpreted as requiring particular conditions for knowledge of intellectual truths. These truths cannot be acquired by sensible experience. As such, their innateness is a condition on possessing them; there can be no acquiring such truths by experience, you can only start out with the ideas themselves or the resources to justify them. This connects with another comment from *Discourse on Metaphysics* which appeals to the requirement of certain concepts as conditions for their own acquisition,

"nothing can be taught us of which we have not already in our minds the idea" (Chomsky, 1966: 63).

Kant is also implicated as continuous with a tradition of philosophers that hold certain mental preconditions as necessary for knowledge or experience of particular types. Chomsky does not cite from Kant, but he appeals to Kant in a general way. At a very sketchy level, Kant's transcendental inquiry aims to identify "conditions of possibility of our first-level knowledge of objects in space and time…renouncing all claims on the supersensible, and redirecting our attention rather to the necessary conditions which make possible natural scientific knowledge" (Friedman, 2001: I, 5). Kant's argument is not meant to fall within the polemic of nativism and empiricism, but Chomsky finds the approach of giving conditions on the learnability of certain types of knowledge to be broadly similar.

The purpose of these historical sources is to ground and contextualize Chomsky's nativism. In some cases, it is possible that Chomsky's stated forbears may be designated misleadingly. The argument from impossibility is a candidate for this type of confusion. Chomsky draws the borders very broadly: nearly any view that posits psychological preconditions on knowledge is a species of nativism. It is not clear how to assimilate the diverse "nativist" arguments--Kant's categories, Leibniz's monadic relations, Descartes' concepts, or Plato's *a priori* knowledge--into one category. Indeed, virtually all *empiricists* will also fall into the category of theorists requiring *some* preconditions on learning (e.g. Locke requires a faculty of Understanding). We already noted above that

externalist-internalist does not neatly categorize the polemical divide; the present argument does not either. This very broad category, “requires preconditions”, does not seem useful in analyzing the contemporary situation in psychology of language. I will not argue here for a better classification of arguments in this tradition, but note only that Chomsky’s aim is not to present a balanced history of the debate between nativism and empiricism. Rather his wide-ranging sources are marshaled to demonstrate that his “Cartesian linguistics” have a long history which “originated and in part were revitalized in the ‘century of genius’ and which were fruitfully developed until well into the nineteenth century” (1966:72). His targets are Skinner and Piaget (Chomsky 1980: 40).

3.3.2 Impossibility Arguments for Language Learning

While Impossibility arguments have a major role with nativism's main historical proponents, they do not play a big role in Chomsky's nativism. He does not himself provide arguments for the logical impossibility of learning language from purely empirical bases, as he does with the empirically-grounded Poverty arguments. Of course, his position is not inconsistent with this type of line. Fodor has focused on presenting an argument of this type for the innateness of concepts, a position he thinks is required for any kind of linguistic nativism. There is also an argument in psycholinguistics for the unlearnability of language due to the logical impossibility of presenting "negative evidence". I will consider each of these briefly.

Knowing how to use language requires possessing some basic concepts about language, such as concepts about noun phrases or plural words. Since such concepts name the basic terms over which syntactic rules operate, a language user could not get by without them.

This is the case both for empiricist and nativist accounts of language; it's the linguistic knowledge that makes use of concepts however that knowledge was acquired. The question of concept acquisition, however, is a general issue distinct from the main linguistic issues.

Fodor (1975, 1981, 1998) has argued that the empiricist cannot explain concept acquisition. The essence of the issue concerns *atomic concepts*, the subset of concepts including those that are not constituted out of any other concepts. A paradigm example is RED, the color concept named by "red". How do we come to have the concept RED? One way, induction over experience of red things, is ruled out already by Meno. You cannot recognize something as being red unless you already know what red is. An empiricist's inductive explanation of concept acquisition might work for complex concepts like FLURG (which is identical to “GREEN or SQUARE”), since it only requires the association of existing concepts with a new term. But this empiricist approach does not work for primitive constituent concepts such as RED or NOUN PHRASE. Fodor argues that such atomic concepts cannot be acquired purely in virtue of experience; some innate element must be implicated.

On his view (1981), we innately have a mechanism which red things trigger, causing within us the concept RED. That innate mechanism, though not embodying a concept in the absence of the empirical trigger, is essentially the innate RED-learning-mechanism. A variation on this argument can be used to suggest that language learning is impossible without at least some endogenous linguistic-concept-learning mechanisms. For example,

without an innate noun-phrase-learning mechanism which can be triggered by the mere occurrence of a noun phrase in the primary linguistic data, a speaker can never acquire the concept NOUN PHRASE. And if this concept is essential to any learner's acquisition of phrasal syntactic rules, then the possibility of natural human language acquisition demands at least some innate mechanisms.

This argument for concept acquisition is rather removed from the mainstream psycholinguist's interest in mechanisms for language learning. There is, however, a more empirically grounded argument for the impossibility of language learning. This is a second set of Impossibility arguments for language. Cowie (1999) reviews a recent crop around the "no negative evidence" problem, the problem that the grammar for a language cannot be learned empirically from positive instances only. Children learn language by exposure to the language itself, including ungrammatical utterances, and very few or no specific instruction about what is *not* appropriate linguistic performance.

An empiricist account of language learning describes the learner as generalizing certain rules from the body of sentences encountered. Simple memorization will not do, since learners go on to produce novel sentences and since the body of grammatical sentences is very large. But moving from instances to generalizations is a tricky thing, especially if we assume that the learner encounters only normal, mostly grammatical speech. The data in such a case will support many hypotheses about the appropriate grammar for the language. Lack of negative evidence is a problem of having too many hypotheses, and positive evidence does not help trim those back to the single useful set of rules. The

literature especially focuses on the absence of *negative evidence*, in the form of corrections from other speakers or other specific statements about the grammaticality of particular forms. As a result, there will always be many grammars compatible with the presented data (think of the polar interrogatives case from Section 3.2.2). And almost all of these grammars will be incorrect, containing sentences not permitted by the true grammar of the language (Pinker, 1984; Hornstein and Lightfoot, 1981; Cowie, 1999). Purely empirical language learning is impossible because induction on the observed data is insufficient. Since a relatively compact set of rules produces a very large set of sentences, observation of the sentences alone typically underdetermines hypotheses about the grammar.

Cowie (1999) is right to complain that this argument looks too strong. Any inductive learning suffers, in principle, from the lack of negative evidence the way the argument is presented above. All we have for most hypotheses is confirmation; only occasionally we get falsifying observations. So Cowie thinks the "no negative evidence" problem simply points out a ubiquitous and familiar problem for inductive learning. If this problem is merely Hume's problem of induction, that observations are never definitive when generalizing from inductions, then we should be skeptical of its conclusion. Learning what a curry is from empirical observation cannot possibly rely on an exhaustive survey of all things that are curries, so in that way it is like language learning; further, there is no reliable source of negative evidence about curries for most learners. We just see some things on the menu called curries, and others not. Yet, we seem able to learn what a curry is from only a few instances and without an innate curry-learning-mechanism. (There

probably *is* negative instruction about food names; but let's assume Cowie's example works.) Cowie takes this situation to be analogous to the situation with language. If we think the lack of negative linguistic data implies nativism about language, then the same argument should also imply that we have innate knowledge about curries. But surely this is absurd, on Cowie's view, and so the argument must fail.

Unfortunately, the gesture at a *reductio* of the anti-empiricist learning argument makes two mistakes. First, the essence of the linguistic argument isn't the lack of negative evidence. The crucial point just is that the linguistic data significantly underdetermines the grammar. The child's positive observations are not in fact sufficient for deciding the many linguistic rules that apparently are learned. It may be that no finite amount of positive linguistic input is sufficient, but that an infinite amount would do the trick (Gold, 1967). In that case, we can say of any child, *a priori*, that its experience has only been finite and therefore insufficient. So the core of the nativist's argument is that no reasonable amount of positive evidence could be sufficient, and therefore we can infer that any child has insufficient evidence to learn language.

Note that this is an argument that has been made specifically for the case of language (recall the example of polar interrogatives above), though it could perhaps also be made for other subjects. Learning does often happens on the basis of positive instances alone, as when I introduce a child to Fred by saying, "This is Fred." Usually, I don't have to point out what is not Fred. Even complex learning can work this way, as when a child learns how to play football from observing others and receiving advice on good moves to

make. Even where negative instruction is provided, it may be that the child *could* have learned with purely positive instruction as well.[9]

However, the nativist argument needs to show further that *negative* evidence either would be insufficient or is actually unavailable. Here is where negative evidence starts to matter. Since there is no negative evidence, nativism must be true, say the nativists. But Cowie mixes up her challenge to the argument. With language, we know that finite positive evidence is not enough, so the empiricist would have to show the availability of negative evidence. Cowie's case of curry, however, does not meet this premise. A few instances of positive observation should be plenty for determining what curries are; no special argument has been provided to show why not, as it has in the case of language.[10] Indeed,

---

[9] A more difficult problem arises if we say that nothing can be learned from positive instances alone. Kripke's Wittgenstein seems to say this, saying we can never know which rule is being followed simply by observing that rules employment. As such, no positive instances would ever be enough to learn what rules of football-playing or language-using to deploy. Let us leave this aside. But note that the natural successor would be the question: would adding negative instruction be enough to learn such rules?

[10] Is it *in fact* true that no negative evidence is available for learning curries? Probably not. Let's use Cowie's example anyway, though.

Mark Crimmins makes a more provocative, related point: even positive instances of curry carry with them some negative instruction. The cases chosen were surely chosen carefully to suffice as a demonstration of curry. The fact that the teacher didn't choose some other cases is important, and that is a kind of negative evidence. This *other* thing not chosen is not a curry.

learning what a curry is should be the same as learning what a car or dog or other named object is. So the fact that there is no negative evidence is indeed irrelevant. There is no negative evidence, yet we know what curries are just from the positive evidence. Language is a different case, since confirming, positive observations would never suffice for children to learn language rules (according to nativist arguments like Gold, 1967). While some psycholinguists have advertised this as a "logical problem" of language acquisition, it seems to be beholden to the availability of specific arguments: negative evidence matters when a concept cannot be learned on positive instances. It may be an empirical issue to determine when this is the case.

There is a second problem with Cowie's criticism. CURRY is a complex concept, like FLURG and not like RED. CURRY is some combination of constituent concepts like FOOD, SAUCES, SPICED, MIXTURES, and others. Nobody thinks complex concepts like CURRY or FLURG need dedicated innate mechanisms for their acquisition. They are assembled from their atomic components. So if there is no negative evidence about CURRY, you still have no need to appeal to a curry-learning device. Essential to Cowie's *reductio* is that any argument that justifies a curry-dedicated learning device must be

---

I think the way to address this interesting point is to step away from it. If there is such a thing as learning by positive instances, curry or not, then Cowie wants to use that as an example where something was learned without negative evidence. If Crimmins is right that there is no such case, then Cowie's argument is over very early. If there is some such case, then I contend it is not relevant anyway. The language case is one where positive instances are demonstrably *not* sufficient, only after which does the question of negative instances matter.

absurd. Anyone can learn about curry by assembling it from its constituent concepts, like FOOD and SPICED. Cowie's line of argument does apply, however, to atomic concepts such as NOUN PHRASE or other linguistic elements. Those cannot be assembled from parts. And if we focus only on primitive concepts, which we presume are relatively few and basic, Cowie does not get her *reductio*. Even if we end up with a dedicated device for each of them, there are relatively few concepts in all (Matthews, 1999 makes a generally similar point). Cowie's argument depends on forcing the nativist to proliferate a learning-device for every concept for which there is no negative evidence, but the best she can hope for is to proliferate learning devices for every such *atomic* concept. Those are surely far fewer, and CURRY is not among them.

3.4 Fixed Capacities

A third type of nativist argument focuses neither on the learner's environment nor the learnability of the relevant knowledge. Most taxonomies of nativist arguments stop only there. Here I want to suggest a third proper category of argument with an independent empirical background and polemical tradition. Such an argument begins by articulating a specific picture of the psychological faculty implicated by the mental objects under study. From demonstrated facts about the mental system itself, especially features of the system's internal regularities, we are meant to draw conclusions about the substantial role for internal or native mechanisms in psychological development. These fixed capacities have certain universal characteristics wherever they are found in a species or in the history of a species.

In the internal/external dialectic so far used to frame these nativist arguments, the argument from developmental mechanisms is the first positive argument from evidence for innate mechanisms. Poverty and Impossibility focus on ruling out the empiricist option. Here, the emphasis is on the direct evidence for internal structure. The more structure and processes established, the less plausible it becomes to maintain the associationist's picture of a general-purpose learning device. This makes the empiricist's job harder. In the explanandum's simplest form, she need only explain how experience can cause a learner to acquire that specific psychological ability. But if there are characteristic developmental pathways, dissociation and impairment effects, and other regularities, then the empiricist must again look to the environment to provide more sophisticated and structured cues for the explanandum's more structured behavior. By making the empiricist case less plausible, the developmental argument provides non-demonstrative evidence for a nativist solution. The capacity operates in its unique way not because of some external trigger or correlate driving the process, but because this operation directly exhibits the capacity's internal structure.

Consider the example of the students in the nightclub. One strategy against the externalist relies on refuting his candidate mechanisms, arguing that students were unlikely to be present outside the nightclub and that they couldn't get in if they were. The present strategy focuses on the internal situation. Several kinds of facts are relevant. It may be that tonight's DJ is himself a student, or that students have accompanied him on other occasions. The students may have been sighted moments before the opening, confirming that they were there from the start. Or they may be in staff uniforms, suggesting that they

were hired as bartenders. Any of these scenarios suggests a non-externalist mechanism for the observed phenomena. But short of ruling out any other explanation, such evidence does force a more subtle explanation from the externalist. It will not do to say simply that students can enter the club, because something further must explain why they were there before the opening and wearing staff uniforms.

The developmental argument can draw on various types of evidence. Each of them works to establish the independence from experience of the ontogenesis of a particular mental function. While the argument is being called "developmental", it does not always relate directly to features of the organism's psychological development. Distinct aspects of a particular mental function can underpin a developmental argument:

Developmental Rigidity. The regular ontogenesis of a mental capacity typically progresses through a sequence of discrete phases with characteristic timings and ranges of ability. Children progress from first saying "went" as the past tense of "to go", to saying "goed", and then again to saying "went". To the extent that this diachronic sequence of states is rigid with regard to specific parental instruction, lack of exposure to particular linguistic data, and so on, we can infer that internal processes are operating independently of inputs from experience.

Universality. Certain features of mental capacity are consistent across wide ranges of individuals, despite the diversity of their individual experiences. Speakers of languages as diverse as French, Hindi, and Korean all employ phrasal structure in construction and

understanding of their utterances. This commonality across extreme diversity can suggest that there is some common factor, though it may be a deep feature of the empirically extant languages rather than of human brains (e.g. everyone believes that unsupported objects fall to the ground, but this is not decisive evidence for an innate theory of gravity). Universality is a pattern in need of explanation. When individuals in highly diverse environments share bodies of concepts or patterns of behavior, we have to look to their shared background to explain this. Of course, their own biological endowment is only one aspect of what they share. In some cases, their common exposure to gravity, or carbon-rich environments, or sunlight, will suffice to explain even high-level aspects of psychology, as many writers have noted (Boyer, 1994; Samuels, 2002). Still, universality is an important non-dispositive piece of evidence.

Articulated Structure. Information about the steps through which cognitive systems develop in children, the characteristic patterns of breakdown due to neural damage (dissociations), or the details about psychological processes can illuminate the component structure of a particular faculty. Cases of Broca's Aphasia display relatively good speech comprehension though very limited ability to produce speech. Such dissociation implies that the underlying mechanisms of these abilities are at least somewhat independent. Similarly, the persistence of optical illusions implies that certain cognitive systems do not have access to information available to other systems (i.e. the fact that the image is an illusion) and are therefore distinct psychological systems. This sort of evidence contravenes the empiricist suggestion that a single learning device accumulates data into a materially homogenous body of associations on various topics.

The empiricist is required to provide an explanation how such isolated bodies of "associations" can develop from the relatively undifferentiated empirical input of language comprehension for input and for output.

Origins and Causes. Postulated explanations for why internal mechanisms have their particular hypothesized structure, such as God's creation or genetic predisposition shaped by evolution, increase the plausibility that they indeed have such structure (Cowie 1999). Insofar as accounts of the phylogenetic development of mental structure become plausible, the idea of innate structure gains credibility. The modern version of this is typified by Chomsky's frequent appeals to as-yet-unknown genetic structure that dictates the development of the language faculty. If it is accepted that genes carry endogenously specified programs for development, then it seems plausible to appeal to them as bearers of programs for developing innate mental structures of great complexity.

In all these varieties of argument from the character of the developmental process, the common nature of the cognitive capacity plays a central role. From facts about the capacity that are shared in all language users, for example, the empiricist's job becomes more difficult. The empiricist must show that a corresponding phenomenon exists in nature, not just sometimes, but generally enough to explain the uniformity of the capacity under consideration. Development rigidity, universality, articulation, and origins all give structure to the capacity, raising the bar for the empiricist's counter-explanation.

3.4.1 Historical Arguments for Fixed Capacities

The classic historical argument for nativism is from "universal consent", as Hume called it. If everyone, everywhere is found to hold some view, this is taken as evidence that the view itself is innate rather than merely learned. Consent to an idea such as the existence of God, or in the dimensions of space and time, requires the possession of a complex mental structure of some sort. But to possess something so complex, and therefore unique, in a world that permits of so much variation in experience and environments, is an unlikely coincidence. Universality, therefore, is evidence that the environment does not create the capacity through experience, but that it is inborn.

Universal consent puts a high standard of proof for demonstrating that a given mental structure is present universally; the nativist must meet the Quinean criteria that a speaker assent to any proposition he can be said to believe. Leibniz admits that universal consent does not guarantee that a belief is innately held; quite the opposite, consent is not a reliable guide to whether an idea is held at all. While consent indicates little, lack of consent also:

> A principle's being rather generally accepted among men is a sign, not a demonstration, that it is innate;…the way for these principles to be rigorously and conclusively proved is by its being shown that their certainty comes only from what is within us. As for your point that there is not universal approval…even if they were not known they would still be innate, because they are accepted as soon as they have been heard. [And] everyone does know them…we use the principle of contradiction (for instance) all the time, without paying distinct attention to it. (76)

Leibniz emphasizes the disposition to assent over the actual assent. The innate knowledge is not the knowledge or principle itself, but the underlying inclination to believe it:

> This is how ideas and truths are innate in us--as inclinations, dispositions, tendencies, or natural potentialities, and not as actions; although these potentialities are always accompanied by certain actions, often insensible ones, which correspond to them. (52)

It may not be easy to detect an innate endowment, since it may not take the form of the knowledge it eventuates. In Leibniz's famous example of the statue of Hercules, the black veins in the marble function as constraints on what the sculptor can do but simultaneously as the rough outline of what will be the finished statue.

Descartes gives a similar account of what innate knowledge might be like, using the analogy of a congenital disease. A disease may be present from birth but not evident, displaying no symptoms, until it makes an explicit appearance. Stich (1975) develops this a bit further as a "dispositional" account of innate knowledge. The knower has something, not the actual knowledge, but a positive inclination to develop the full knowledge.

Arguing for innate knowledge includes the claim that such knowledge will be found universally. Descartes and Leibniz each provide formulations of innate knowledge that avoid a typical empiricist criticism: that children do not exhibit some knowledge, or that certain groups do not. In either case, the knowledge can be defended as latent but inactive.

Furthermore, each argument supports the principle of innate knowledge without suggesting that some particular knowledge is so. Descartes and Leibniz use different arguments to establish that triangles or ideas of God are innate. But by characterizing the nature of the mind as the sort of thing that might have latent diseases or seams of black marble, they contradict the idea that everything is learnable.

3.4.2 Modern Arguments for Fixed Capacities

Chomsky frequently characterizes linguistic ability in a way meant to make nativism of its features more plausible. He says language "grows" in the mind of the child, and says linguistic capacity is the function of a "language organ". Nobody disputes the possibility that bodily organs like hearts and kidney "mature" in the fulfillment of a biological program of development, largely of internal and innate provenance. Chomsky characterizes the situation of language similarly: language matures like an arm (Chomsky, 1984). And therefore many of its principal structural features are present from the beginning. On the one hand, this is a metaphor for what nativism could mean for mental capacities; but at the same time it is an argument *for* nativism. It says that mental capacities are of a kind with bodily organs, and share their developmental processes.

Lennenberg (1960/1964) leans heavily on the presumption that cognitive capacities must be fundamentally similar in their provenance to bodily organs; and he suggests that the methods of evolutionary biology are appropriate for identifying which cognitive capacities are species-typical and likely to be innate. He considers two dimensions. When all members of a species at a certain time have a trait, this suggests innateness. Similarly,

when all members of a group (say, humans) have the same (cultural) trait over the history of that group, this is also evidence for innateness.[11] In both cases, the species or the cultural group, the criterion for group membership is a shared biological endowment. Throughout the species at a time, or in the particular group over a stretch of time, many aspects of the environment will vary. So if a particular trait is observed to be constant, it is better correlated with the common genetic endowment than it is with the environment.

There are two other types of arguments about the rigidity of mental capacities. First, there is evidence of dissociations. Disorders or physical insult to the brain cause various types of psychological impairments, disabling narrow types of function. This dissociation between functions is evidence for articulated structure in the overall cognitive system. Insofar as this structure is judged typical of all humans, this constitutes a kind of black vein in the marble, limiting the ways in which cognitive architecture takes shape. Of course, the physical material need not constrain the functional structure. But when physical damage disables a particular function and not another, we see that the two functions are relatively independent. Second, there is substantial evidence for regular schedules of psychological development. Children babble during their first year, start using words around their second year, and so on. Generally observed courses of behavior such as this suggest that the language organ is more like the maturing arm than the blank slate of experience.

---

[11] It is suggestive of innateness. Though other factors could also explain it, like tradition.

3.5 Summary

In this section we considered a variety of historical and modern arguments for nativism. There are three categories: poverty of the stimulus, impossibility of learning, and the argument from fixed capacities. All three have various incarnations in the standard historical nativists, and have been used by Chomsky and allied theorists in favor of the language faculty.

A number of famous debates in psycholinguistics fall squarely into the category of Poverty arguments: Chomsky's debates with Skinner and the behaviorists; creolization evidence; specific rules like auxiliary fronting, is-contraction, or past tenses of verbs; and most evidence drawn from early in development, such as Motherese detection.

Several fall into the category of Impossibility arguments: Chomsky's debate with Putnam on whether there are general learning rules; Gold's Theorem; "no negative evidence" arguments; and Lennenberg's argument from the impossibility of other species' learning language, and therefore the innate human-typicality of language.

In the third category, arguments from Fixed Capacities, I suggest we separate out: dissociation evidence; existence of linguistic universals; critical period effects; Lennenberg's arguments about the universality of language across all current peoples and cross-temporally through human history; and Chomsky's Universal Grammar argument.

In Chapter 4, I detail the technical structure of these arguments in considering how they are related to each other and to modularity.

4. Modularity

The concept of modularity has less illustrious history behind it in psychology or philosophy than the heavily debated issue of nativism. As a serious concept for cognitive architecture, there is no precedent until Chomsky himself. The historical roots of modular cognitive architectures are certainly prominent in 19th Century neuroscientific thought, but there are deep differences with its present deployment. As we have noted, the basic arguments in Chomsky for nativism are well pedigreed. The structure of modularity arguments has less substance to draw from its historical antecedents, but here we will see some of the connections between the major theories of modularity.

The essence of modularity in contemporary psychology is the notion of *independence*. Modules perform cognitive functions independently of each other. In so carrying out their apportioned functions, they are subsystems of the overall mind that isolate groups of activity into distinct functional units. This basic notion ties together the contemporary and historical traditions.

Framing modularity in the contemporary context forces it to work with a broadly computational theory of mind, a framework in which earlier sources do not fit. There is, of course, a long tradition of explaining the mind by decomposing its distinct functions or faculties as evidenced in sources as diverse as Plato, Locke or Leibniz. Distinguishing one faculty from another, for these views that break down the mental activity into

distinctive types, is primarily to distinguish functional types of activity. *Intellect* is different from *sense*, say, in Descartes' view, where only intellect has direct access to metaphysical essences. They are different types of activities, the way addition and subtraction are different from each other. But they are also restricted to different domains, as vision and hearing treat distinct types of input.

The uniqueness of the contemporary view, however, is that it provides a more precise way to interpret independence, as a fact about the information flows in and between the functional subsystems. I have advocated *informational isolation* as the account of independence to apply to cognitive subsystems. A system is informationally isolated when its core functions are rigid or fixed regardless of the inputs to that module. That function or set of functions constitute the information processing heart of the module. This treatment of modularity is roughly consistent with Chomsky's, and so a useful way to draw a line from Chomsky through the various contemporary theorists who rely on modularity in their views of the mind.

This minimal treatment explicitly rejects overloading modularity with a laundry-list of hypothesized features. Fodor (1983) popularized a family of characteristics which *informational isolation* does not incorporate. Furthermore, it is important to make the conceptual distinction between *cognitive* modularity of interest to us here, and fully distinct concepts such as anatomical (neural) localization or functional specialization. Cognitive modules may or may not be localized, and equally they may not be "specialized" to any function (a general purpose device could be a module though it were

not specialized). Chapter 1 discusses some reasons for this minimal approach. Chapter 4 details the types of arguments available for modularity. In discussing some of the paradigmatic statements of modularity, however, we will see that *independence as informational isolation* will be completely adequate for the present aim.

4.1 Chomsky's Modularism

Chomsky sees his view as dramatically different from Descartes' view of cognitive modularity, a view he summarizes as "that there is no modularity" (1984: 15). For Descartes, the mind is a single, unitary thing, "there is within us but one soul, and this soul has not in itself any diversity of parts…The mind is entirely indivisible" (15). Descartes is not a materialist, of course, and so the physical notions "divisibility" and "parts" are impossible to apply to mind. He would surely reject a view of the mind as composed of mental organs—functional units that map onto distinct chunks of brain matter. Descartes says "we cannot think of a body but as divisible, while the mind or soul of man cannot be conceived but as indivisible; because we would not know how to conceive of half a soul" (in Flourens, 1851: 56). Chomsky reads this, however, as a "homogeneity principle", that there are "no mechanisms of mind". This must be wrong, since we have already seen Descartes distinguish between gross faculties like intellect, sense and imagination. Descartes has a picture of functionally distinct mental capacities, even if they are not physically divisible or otherwise dissociable. Perhaps they are the independent powers of a single, unitary soul—i.e. he pictures an ontologically unified mind which has functionally distinct faculties, much as the connectionist may picture a homogenous neural structure underlying the distinct cognitive capacities (Farah, 1994).

Such a view would meet the core criterion on modularity, that it posit the independence of mental capacities.

Chomsky's modularity view does have important differences with the bland admission that the mind has distinct capacities. Chomsky (1966) begins to elaborate a pattern of argument for nativism that ties in a modularist view of mind: language is a product of a language *faculty*, an independent subject of study for psychology. Moreover, drawing on Lennenberg's (1964) influential nativist statement, this faculty *grows* in the mind as any other biological subsystem grows. Just as biological subsystems can be identified as organs, language is produced by a *mental organ*. This initial notion persists throughout the development of his views, as here:

> Every complex biological system we know is highly modular in its internal structure. It should not be a terrible surprise to discover that the human mind is just like other complex biological systems: that it is composed of interacting sub-systems with their specific properties and character and with specific modes of interaction among the various parts. I should say that when we look at a particular system, say language, we also find internal modularity….In fact, it seems to me fair to say that wherever we know anything, that is what we discover. (1984: 16)

Parity of reasoning extends the functional independence of principal body organs to the mental organs. Of course, the heart must beat for the kidneys to receive essential nutrients. They are *operationally* dependent on each other. But their functions are distinct and contained within the individual organs. The kidneys perform a series of procedures on incoming blood that are not fundamentally affected by the heart's outputs. External

cues will invoke various procedural options, but no external system can "re-wire" the kidney to revise its behavior beyond the pre-determined menu. The same should go for the language organ, on Chomsky's view, such that the language faculty works according to its own internal rules. However much we learn about vision, for example, we will not know much about linguistics.

Tied up closely in the analogy to bodily organs, Chomsky looks like he is advocating a doctrine of neural localization along with his claim that the mind has functional parts. Each module has a physical locus, where specialized hardware carries out its functions.

A second key element of Chomsky's (1966) argument for nativism is that intentional, knowledge-like states explain linguistic ability. Different faculties are each explained by the existence of theory-like bodies of intentional states. Furthermore, Chomsky's (1984) follows Fodor (1983) in calling them informationally encapsulated. More than being bodies of knowledge, modules operate independently of any non-input information, such as that residing upstream in higher-order systems. Even though he apparently advocates this view and Fodor has repeatedly underscored its importance to his view, Chomsky never makes very much of this point. He does not detail the connection between two rather independent treatments of modularity: the mental level account of modules as informationally encapsulated bodies of knowledge on the one hand, and the analogy from mental faculties to bodily organs on the other hand. Taking on a computational picture of mind shows how these can fit together, but Chomsky does not take up the subject.

Finally, Chomsky does not set any limits to which mental functions may have dedicated mental organs. He often suggests that vision and arithmetic are modular, and he explicitly considers whether the language system itself may be composed of various modules for syntax, semantics, phonology, and perhaps other functions.

4.2 Gall

Gall is the major modularist historical source, associated closely with Fodor (1983) but his influence on Chomsky is apparent. Gall's "organology" identifies 27 distinct mental faculties, each of which resides in an innate, independent mental organ. Flourens, his principal 19th Century critic, summarizes:

> All of Gall's philosophy consists in substituting *multiplicity* for *unity*. For one brain, general and singular, he substitutes many little brains; for one intelligence, general and singular, he substitutes many *individual intelligences*…[or] *faculties*…Each of these (since each is itself an intelligence) has its own perceptive faculty, memory, judgment, imagination, and the rest. (1851: 29)

To a Cartesian critic like Flourens, one chief aspect of Gall's modularism is non-controversial: modules are innate. For him, the controversial feature is the second: independence. Each of the various faculties—like pride, mathematical ability, self-defense or poetic sense—relies only on its own resources.[12] These resources are generic

[12] The full list, from Flourens (1851): the instinct of reproduction, love of offspring, instinct of self-defense, instinct for predation/carnivorous instinct, sentiment of propriety, friendship, cunning, pride, vanity, circumspection, memory of things, memory of words, sense for place, sense for persons, sense for language, sense for color reports, sense for sound reports, sense for number reports, mechanical sense,

functions like memory or judgment, the traditional mental faculties that range across all the mind's activities. Each module must contain within it some resources of memory and perception in order to complete their specific functions.

Gall's nativism professes to espouse a different basic commitment than the traditional doctrine of innate ideas. Sure enough, "dispositions and properties of the soul and the mind are innate and their manifestations depend on their organization" (Gall and Spurzheim, 1811: 3). But innateness in this context does not mean innate *ideas* or principles. They are happy to admit that sensations of things like birds or trees must come from "outside"; they are not "innate sensations". It is the *faculties* that are innate on their view, a distinction meant to highlight that faculties are not simply collections of knowledge on various subjects. They are fundamental mental powers, the same characterization the Cartesians would give to their class of fundamental faculties like sense. As such, this is a much more "mechanistic" than theory-like treatment of psychological explanation, where mechanism appeals to some material or immaterial mental mechanism.[13]

---

comparative wisdom, metaphysical mind, abrasive/caustic mind, poetic talent, good will, mimicry, religious sense, and firmness. Spurzheim adds 10 more.

[13] In Chapter 3, on folk psychology and theories, I argue that the mechanistic approach and the theory-like approach, from the point of view of contemporary computational psychology, are in fact fundamentally identical. They offer two levels of description for a single system, not competing visions.

Fodor (1983) made a key distinction from Gall's independent mental organs. The traditional faculties are "horizontal" on Fodor's view, cutting across a wide range of mental activities. Memory has a role in self-defense, pride or mathematics. But Gall's organs are "vertical" faculties, organized around depth in a particular functional domain. They carry with them their own generic resources. A chess-playing faculty is more Gallean than Cartesian, since it picks out a highly specific domain of function which requires substantial information of rules and strategy as well as generic resources such as memory and imagination.

Fodor explicitly invokes a comparison to the modern notion of *talents*, which seem to include both natural or intuitive knowledge for the activity as well as superior capacity for some of the underlying generic functions. So a talented chess player will know things about her situation, about threats or possible openings, that a less talented player would not see. Perhaps this is superior innate knowledge of board situations and where they lead. Or perhaps it is a more powerful imagination that permits her to see further into the possible future of the game. Either way, it seems to include knowledge or capabilities for performing certain tasks efficiently.

A look at Gall's list of faculties, however, does not reveal a list of activities like chess-playing or diplomacy (see the last note for the list). Instead, there seems to be an enlarged list of sensory powers—color vision, hearing, sense for numbers, language, sense for place, etc.—along with a collection of dispositions or personality traits like cunning or reproduction. They do not look like talents. Cunning looks more like a set of desires than

a set of beliefs. Nor is cunning the sort of mental quality that is highly constrained in its application. Cunning, as an element of temperament, seems like it would have a role in a wide range of mental activities: chess-playing, diplomacy, theorem proving, etc.

Fodor is not alone in attributing something like the horizontal-vertical distinction to the mental organs in Gall, since contemporary critics like Flourens do emphasize the implausibility of a picture of mind where there are 27 different memories, judgments, and sense. It is possible, however, that we can read out the focus on *types* of faculties and instead see a focus on *which* faculties. The Cartesian view has mental faculties follow the landscape of metaphysics: intellect perceives the world of essences, sense perceives the world of sense, imagination projects beyond perception, and judgment applies reason to known particulars. Gall has more faculties, and they seem to follow a psychological assessment of the mind rather than a purely metaphysical one.

Of course, Gall also saw a very close relationship between the physical organs that housed each faculty and the nature of their functioning. The physical development of an organ should correlate to its mental development, on his view, and this would have observable effects on how it operates. The practical recommendations of phrenology followed entirely from a reliance on this premise, though the former elements of Gall's organology have fared far better.

4.3 Fodor

Fodor (1983) gave a highly specific account of modular cognitive architecture, drawing a connection from Gall through Chomsky and speculatively detailing a number of

characteristics of mental modules. Fodor offers a laundry list of features: intentional, computational, innate, domain-specific, informationally encapsulated, mandatory, fast, shallow output, inaccessible, neurally local, and regular in development as well as breakdown. Like Chomsky, he thinks modules are bodies of intentional psychological states, so his picture is of theory-like capacities. They are explicitly computational, so the intentional states are implemented by physical symbol-processing hardware, a step which Chomsky does not specifically take. While Fodor's arguments for nativism have focused on the innateness of concepts (via an Impossibility of Learning Argument) rather than Chomsky's analogy to bodily organs and the Poverty of the Stimulus, he advocates a pretty similar view. He agrees that mental faculties like language or vision have dedicated neural hardware with characteristic developmental pathways (they "grow") and also patterns of breakdown (since they all grow with similar structure).

Modules are independent because they are both informationally encapsulated and inaccessible. Other systems do not control the language system, nor do they have access to all the information or computational results inside it. This is a lot like Chomsky's account, but occurs at a rather different level of description from Gall's. For Gall, the main focus is on operational resources, like disk space or processing time, and Fodor's emphasis is on the information flow itself. A single physical computing machine with a single processor and disk could simultaneously compute the results of two entirely separate calculations (by breaking down the problems into steps, and sequencing them). In so doing, the operating system of the machine could keep the information flows

between these calculations entirely separate—meeting Fodor's constraint—even though the calculations took place on shared physical resources, and so violated Gall's.

A further feature of Fodor's modules is that they are domain-specific, or that they are limited to only a particular set of inputs. Vision only treats visual information, and not any other types. Memory, on the other hand, is usually considered to be domain-general, since it is not constrained by subject matter. Roughly, Cartesian faculties are completely domain-general, while faculties like Gall's or Chomsky's are more domain-specific. This contrast follows Fodor's distinction between horizontal (domain-general) and vertical (domain-specific). This is a tricky distinction to make precise, however. It is very difficult to say exactly what criteria we should measure in judging specificity or generality. Vision is not constrained by *subject* matter, but by range of physical energy. Language is not constrained to any particular physical structure, but to a class of symbolic systems with the underlying grammar of natural languages. Chapter 5 discusses this issue in more detail.

Chomsky leaves his taxonomy of modules rather open-ended. He knows language and vision are modules, but the rest don't matter much. Gall at least gives us a discrete list of 27 specific area, though he adds general intelligence on at the end, as the capacity that deals with all other issues. This is a bit like Fodor's picture, which limits modular architecture strictly to those cognitive systems that deal with the perception of external stimuli, functions of the classical faculty of *understanding*. All else, on his view, *must* be reserved for non-computational implementation in "central cognition", for reasons

concerning the limits of computational explanation. So only some mental functions are modular. The rest are still mysteries and may always be.

Fodor's account of modularity is speculative. He says the mind is modular, then lists a variety of features he expects every module to have. But *to be a module* is not *to meet Fodor's list*. Most controversially, many theorists have disputed the Fodorian limits on modularity, arguing that higher order cognition is very likely to be modular. But further, as I have argued in Chapter 1, Fodor's particular formulation of informational encapsulation may be wrong; or it may be that modules come with various apportionments of features. Some are domain-specific, others are only fast and mandatory. It is folly to follow the list from Fodor (1983) as a set of theoretical ground rules for distinguishing "real" modules from "fake" ones.

4.4 Contemporary Modularities

4.4.1 Karmiloff-Smith and Connectionism

Karmiloff-Smith (1992) outlines a view of modularity that takes issue with two key elements of the canonical Fodorian modularity account. First, modules are diachronic. They change over time as experience and biological development operate. Fodor's picture is a snapshot of a fully-developed module, and does not leave open pathways for developmental change or learning. Yet, on her view, nearly all modules change in this way. Second, modules are not as deeply innate as Fodor suggests. Some modules may not be innate at all, such as deeply routinized skills. Reading, for example, is entirely learned

but exhibits many of the features of modularity: neural locality, mandatory firing, encapsulation, inaccessibility.

A third contention is that modules are not "computational", but may easily be implemented by dedicated neural networks (Elman et al. 1996). This is not a settled debate, but it does appear that Fodor and Pylyshyn (1988) get the better of it, in that even neural networks implement what can be considered a "classical" computational architecture (broadly construed).

4.4.2 Developmental Psychology

Developmental psychologists working on a wide range of psychological competences, such as folkbiology, theory of mind, or psycholinguistics, have adopted the basic modularist approach from Chomsky's initial arguments on language in taking up the "dominant explanatory strategy for cognitive science" (Stich and Nichols, 1992: 10). Competences are explained by bodies of theory-like intentional states. In studying development, this knowledge endowment is typically more or less innate. It is also typically unavailable for introspection, informationally encapsulated, etc.

The emphasis for these researchers is typically on reasoning about given subject domains, not purely perceptual or input-system activities. Theory of mind reasoning, for example, draws on information about other human agents and forms expectations about their behavior in real and hypothetical contexts from stored resources of information. This is a very high order input system. In general, the research methodology seems to rely on a

"module doctrine" (van Gelder, 1994): where there is a discrete mental function, the theorists looks for a discrete, dedicated physical component.

4.4.3 Evolutionary Psychology

The basic approach of evolutionary psychologists is simply to include adaptive considerations in the study of psychological capacities. Apart from this addition, the framework is much like the developmental psychology approach of positing bodies of innate, intentional states to explain mental faculties.

Some critics have suggested that this approach becomes dangerous when the "module doctrine" is invoked on adaptive considerations (Nichols, 1999). That is, based on the evolutionary plausibility of a psychology feature, the evolutionary psychologist posits the existence of a dedicated physical component. While the module doctrine, in other instances, infers from observed psychological behavior to the existence of unobserved physical systems in the brain, the adaptationist version is stronger. It infers merely from a plausible psychological response to an unobserved system.

4.5 Summary

In this section we considered Chomsky's concept of modularity and its connections to a variety of deployments in cognitive science. Chomsky's modules are independent faculties that are *intentional*, *innate*, and *neurally local*. Independence is how we separate one module from another, and ultimately the key difference between a modular or unified architecture. Yet, for Chomsky, Fodor, and Gall, independence and its affiliated concepts

of domain-specificity or encapsulation are the difficult issues. A simple intuitive account does not suffice.

The next chapter looks closely at the nature of intentional modules by considering their role in the debate over folk psychology.

Chapter 3. Folk Psychology and Intentional Modules

Modules are usually understood to come in two varieties: intentional modules and mechanism modules. Intentional modules are bodies of knowledge-like states, like those Chomsky (e.g. 1980) invoked to explain grammar. Mechanism modules are more like Marr's (1982) account of the vision system, a set of physical items like cones or neurons that respond to particular stimuli by producing particular responses. Some theorists have suggested that modules can be classified as one or the other but not both, a mutually exclusive pair of categories (Samuels, 2000). The underlying distinction between intentional states and mechanisms as *kinds* of explanation reaches beyond modules to describe two supposedly independent ways of explaining psychological phenomena.

One debate in cognitive science that has taken this distinction seriously is over the status of folk psychology. The theory-theory camp claims that folk psychological ability is explained by a body of theory, an intentional module. The simulation theory camp claims that it is explained *not by theory* but by a simulation mechanism. There is a lot happening in these debates, but I claim that at least part of the debate depends on pitting *theory qua theory* against *mechanism qua mechanism*. For this part of the debate, we see arguments like "only possession of a theory can explain folk psychological ability" or "theory-possession cannot possibly explain folk psychology—only mechanisms can".

Some participants have suggested that this theory vs. mechanism opposition is untenable (Nichols, Stich, et al. 1996; Davies, 1996; Heal, 1994; Davies and Stone, unpublished). When psychological explanation invokes "theory", they say, it can mean the possession of tacit intentional states. Possession of such "subdoxastic states" can be attributed even where a person does not properly possess the relevant concepts or any explicit knowledge on the subject. When we apply these particular standards of theory-possession, however, we find that the mental structures posited by the simulation theory actually fit under the category of theories. We find that simulation theories are just a special variety of theory-theory. If this is true, according to some critics, then the concept of "theory" is simply trivial.

This chapter looks at this argument in detail. The conclusion is that theory is not a trivial concept, but that simulation theory does count as a species of theory-theory nonetheless. The debate over folk psychology should be reformed away from this red herring opposition. Furthermore, I take folk psychology as a special case for the larger issue: that intentional explanations and mechanism explanations are not fundamentally separate types, nor are intentional and mechanism modules incompatible. Most modules are probably mechanisms at one level of explanation, and intentional at a higher level. In the Appendix, I look at how this issue plays out in related areas, and how it has caused confusion.

1. Background Issues

Folk psychology is already a familiar debate to many in philosophy and psychology in the 1990s (Davies and Stone, 1995a; Carruthers and Smith, 1996). This first section reviews the main questions in this debate in moderate detail. The key question of this paper, about the fundamental bases for the dispute, begins in section 2.

1.1 Folk Psychology

Folk psychology is the common human ability to explain and predict the propositional attitudes and actions of others. These explanations usually rely on an informal belief-desire "theory" of mind. If we think Fred wants a beer, we conclude that Fred intends to take a beer.

The human competence for reasoning in this way is a complex phenomenon in need of explanation. Normal adults appear to possess this ability without exception, though children do not appear to have this ability until approximately age 4 and make certain regular and universal errors as they develop it (Gopnik and Wellman, 1995; Davies, 1996). Autistic persons are hypothesized to lack this ability for "social reasoning" altogether though they are in possession of otherwise normal cognitive abilities (Baron-Cohen, Leslie, and Frith, 1985; Carruthers, 1996b); and it appears that Williams' Syndrome has the opposite effect, impairing fundamental cognitive abilities but not the development of a recognizable theory of mind. These failures and developmental processes are not well-understood, but are thought to stem from the subjects' lacking the prerequisite conceptual frameworks, e.g. an incapacity for belief-attribution to others, or

perhaps from an inability to adopt alternative perspectives (Currie, 1995). Wimmer and Perner's (1983) famous experiment with the "false-belief" task on children, as well as much subsequent work (Carruthers and Smith, 1996; Davies and Stone, 1995), shows that this is a subtle issue. Premack, Povinelli and others, in work with roots predating the present controversy about human folk psychology, have engaged problems of "false-belief" with nonhuman animals including chimpanzees (Premack and Woodruff, 1978).

One way to explain the competence for folk psychology is by attributing possession of a theory. Broadly construed, this amounts to possession of a body of information about the way other minds work. This body of information need not meet philosophical criteria for status as explicit knowledge, and may be tacit, inferentially isolated from other knowledge, fixed (unrevisable), and innate. On this view, "the theory-theory", folk psychological ability is access to what is literally a theory of mind enabling us to understand others.

On the competing view, a person does not possess any such theory at all. Rather, a process is employed, the operation of which delivers predictions and insights about the mental life of other persons. While the theory-theory might pass through inferential reasoning steps including sentences like "Fred wants the beer", there may be no such corresponding states in the operation of an opaque simulation process. The "simulation theory" instead posits that we obtain predictions about others by simulating the operation of their practical reasoning capacities. We take our own practical reasoning ability "off-

line", and feed pretend states into it, simply taking its outputs as the likely attitudes of the subject under consideration.

For example, I can attempt to predict how a drug will affect you in each of these two ways (Davies and Stone, 1995). On the one hand, I might consult documentation on the drug's psychoactive properties, consider how your case might differ from the situations documented, and then draw conclusions on the drug's effects. I would be drawing on a body of theoretical knowledge. On the other hand, I might not consult any body of information about the drug or about you. I might simply take the drug myself, observe its effects, and predict similar results for you from the assumption that we are biologically similar. On this method, there is a step in the procedure where I would myself be in a state that was *isomorphic* to the state you would be in upon taking the drug. The theoretical approach would include no such isomorphic state, instead relying on some body of propositions concerning the drug and your biology.

This is, I think, a non-controversial characterization of the debate. The theory-theory posits the possession of some sort of body of knowledge, while the simulation theory suggests *instead* that no such information is possessed. Rather, an opaque process is triggered which delivers only conclusions.

There is *more* going on in this debate as well. In general, the simulationists posit a single cognitive device; it is the practical reasoning system itself that serves as the simulator. By contrast, the theory-theorists seem to suggest two devices: the usual practical reasoning

system, but also a folk psychology theory.[14] And there are other differences. But the straightforward characterization of what makes a view a theory-theory or a simulation theory relies on the type of psychological entity invoked. Evidence from the debate's arguments themselves supports this reading. We will return to these other dimensions of the debate later.

The essence of the theory-theory is a patently Chomskyan argument for the possession of a folk psychological competence. At the outset, the approach favors attribution of intentional states in the explananda. Then the view takes an evasive position on the epistemological status of the mental representations invoked to explain the competence, appealing to their at-least-tacit character. In so doing, the theory-theorists have confirmed the status of Chomsky's basic strategy from psycholinguistics as a central paradigm in cognitive psychology.

But the opposition to this strategy has not come from behaviorist anti-cognitivists. The simulation theory is a cognitivist challenge that relies on *other* internal mental functions to explain the theory of mind. The simulationists are simply arguing for something like the view that the capacity for practical reasoning has been *exapted*—adapted by evolution for some further function than for which it was originally selected by evolution (Stich and

---

[14] In *principle*, a theory-theorist could claim that the theory is a part of the practical reasoning system itself. That is, that there is *one* device. The simulationist could suggest that there was a practical reasoning system as well as another little model system used for simulating, i.e. that there are two devices. So it's not accurate to say that the essential difference between the two camps is the one-device-or-two issue.

Nichols, 1992; Gould and Vrba, 1982)—to the additional, dependent function of understanding others. For that reason it seems too strong to claim, as do Stich and Nichols (1992: 124), that "if these philosophers are right [the simulationists],…the dominant explanatory strategy in cognitive science, the strategy that appeals to internally represented knowledge structures, will be shown to be mistaken in at least one crucial corner of our mental lives." Rather, the simulationists will have shown that cognitivist models can find themselves embedded in multiple-function structures, as some recent results on the activation of vision system regions during imaginative visualization indicate (Currie, 1995). The intentional module strategy, appealing to internally represented knowledge structures, is by no means the only approach for a modern cognitivist.

Rather than challenging cognitivism outright, the core issue requires distinguishing two varieties of explanatory strategies: explanation via intentional states and explanation via mechanism or process.

1.2 The Theory-Theory

There are many incarnations of the theory-theory. Gopnik and Meltzoff (1997) have developed an account wherein the child is seen as a "little scientist", not unlike the Piagetian model, observing, experimenting and running hypotheses. The child moves from one conceptual framework to another in a manner like scientific theory change, and deploys a mix of tacit and explicit concepts. On another extreme, Fodor (1992) and Leslie (1987, 1994) describe a theory of mind module fashioned in a revised cognitivist model:

a body of innate, tacit knowledge implemented as a computational, information processing mechanism which is also fast, encapsulated, impenetrable, and so on.

Stich and Nichols (1992) suggest a characterization of the theory-theory camp that might neatly divide the debate (Davies and Stone, unpublished[15]). They suggest that the theory-theory deploys the "dominant explanatory strategy" (121) of contemporary cognitive science as developed in psycholinguistics (Chomsky, 1965, 1980), visual object-recognition (Marr 1982), naïve physics (McCloskey 1983), and other capacities. This strategy posits an "internally represented 'knowledge structure'—typically a body of rules or principles or propositions—which serves to guide the execution of the capacity to be explained." (121) This setup is meant to draw the borders of theory widely, including "any body of information or misinformation about psychological matters…whether or not it is organized around psychological laws." (Davies and Stone: 3) Perhaps the borders are slightly too wide, as noted in the previous section. But let us work from this characterization to follow the path of the dialectic.

Davies and Stone have urged that this way of characterizing the theory-theory is both a tactically clever position and a tidy instrument for outlining the debate. Tactically, it weakens the criteria on establishing the possession of a theory. It also tidies the debate by ensuring that any knowledge-structure account falls under the same roof, and that any view *denying* the role of knowledge-structures can be counted as "simulationist". Though a convenient characterization, it is inaccurate to say that simulationists deny *any* role for

[15] This manuscript was presented at NYU on January 30, 2001.

knowledge structures. In fact, simulationists usually posit *some* knowledge structures to supplement the core simulating mechanism. The tidy characterization oversimplifies the polemical situation.

Yet, if we look at what simulationists attack, we see what common vulnerabilities the theory-theory views share (Segal, 1996). They uniformly deploy knowledge structures at the core. Some views emphasize the modular limits of the capacity or its innateness, such as Leslie and Fodor. Others are more hybrid views where elements of implicit knowledge are important, such as Perner (1996) and Harris (1995). While it is wrong to say simulationists reject any role for theory, they do reject any view that is *all* theory. Theory-theorists put theory at the heart of their explanations. A theory-theory view can be *all* theory. The simulationist view, as it has been understood so far, permits only some or none.

The way to describe the common position of the self-described theory-theorists is to say they appeal to internal bodies of knowledge-like states. Theory, here, has a special meaning. It describes a set of psychological states, *unlike* "scientific theory" (which may not be psychological at all). Nor is the term "theory" limited only to explicitly-held sets of beliefs or garden-variety knowledge. These states may be much weaker than belief, with limited accessibility to consciousness or other cognitive systems, and fixed or unrevisable. The "theory" invoked to explain folk psychological behavior can be as strong as ordinary knowledge, but also as weak as Chomskyan tacitly cognized beliefs

about grammar. It is in this broad sense of theory that theory-theorists are appealing to the same explanatory strategy. This is the sense we'll use here.

1.3 The Simulation Theory

The simulation theory originated as a challenge from Heal (1986) and Gordon (1986) to the theory-theory. Evidence from development, such as Heinz Wimmer and Josef Perner's (1983) application to children of Premack's false-belief task for primates, shows that children do not develop a theory of mind until about 4 years old.[16]

The simulationist proposal is that this change in competence is the acquisition of a capacity for "imaginative identification" with another subject, such as with the protagonist puppet of Wimmer and Perner's (1983) experiment. In doing so, the child takes up the perspective of the subject as pretence, imagining that she herself is in this situation. Rather than thinking about the mental states of the subject[17], the child reasons directly about the pretend situation[18]. The child then attributes the resultant attitudes to the subject.

---

[16] In the classic false-belief scenario, children are shown a puppet play with two characters. One places a marble in a box and then leaves the room, then the other moves the marble to a new container. When the first puppet returns, the children are asked, "Where will she look for the marble?" The younger children point to the new container, apparently unable to distinguish their own beliefs from the beliefs held by an agent with less information. The older children, meanwhile, say that the puppet will look in the old box, where the marble had originally been placed but is no longer.

[17] "The puppet believes that he put the marble in the box."

[18] "I put the marble in the box."

The structure of the capacity depends on "some relevant isomorphism" between a part of the child's mental machinery and that of the subject being simulated (Goldman, 1986: 85). The simulation works because the child's practical reasoning process in particular is a good predictor of the subject's. Putting the mutual similarity at the core of the explanation is the essence of the simulation theory. *What* in the child's mind is similar to the subject's, however, is usually stressed to be the *process* or *mechanism* or *faculty* of practical reasoning. How that process is explained doesn't matter, and it's usually not explored.

The simulation process relies on the practical reasoning ability that the child already has before the developmental turning point. The simulation account suggests that an additional ability becomes available around 4 years. The child gains a capacity for *perspective shift*, which feeds into the existing process for practical reasoning, and enables judgments about the mental states of others.

This simulationist alternative has been suggested as superior to the theory approach on the basis of a number of prima facie considerations (Heal, 1986; Gordon, 1986). First is simplicity. Human psychology is extremely complex on any account, and arguments for mental holism imply that the job of identifying individual intentional states is itself highly complex. The idea that anyone possesses a theoretical apparatus sufficient to decipher human behavior is implausible, much less that children possess one and indeed develop it themselves from scratch.

Second is that children don't theorize much. Children are generally poor theorists in other domains, and their improvement in this particular ability happens rather suddenly. The sudden spike doesn't look like the smooth curve of learning. Even after their performance improves, they are no better able to articulate the theory ostensibly in use.

Third is the typical pattern of learning. Children's improvement in theorizing about mind happens very predictably and universally, in a manner inconsistent with independent discovery hypothesized on the theory view. Fourth, and finally, folk psychological information may fail to stitch together in the proper way to qualify as a *scientific* theory, or may fail to describe any genuine psychological laws at all (if the eliminativists are right, e.g. Churchland et al.). If folk psychology is not a theory, then how can the right explanation depend on theory-possession?

These charges apply best to the "explicit theory" version of the theory-theory, such as Gopnik's, and are interestingly similar to the main charges once made against Chomsky's theory of language acquisition (Chapter 2; Cowie, 1999). The natural path for the dialectic there and here is towards appealing to innate tacit knowledge structures, as some theory-theorists have adopted. With innate, tacit theory, there is no question of children learning a theory; their ability relies on the availability of a set of subdoxastic states. This avoids the simulationists' charges. Of course it also makes the two camps look more similar. Tacitness and innateness resemble key aspects of the simulationist view, for

whom the perspective shift of imaginative identification is not learned and is not an explicit, personal-level process. Simulation is not conscious.[19]

Though the initial proposals for the simulation theory attacked features of a "little scientist" or "explicit theory" version of the theory-theory, there are further arguments designed to recommend it over even "tacit theory". A number of positive arguments have been offered in support of the simulation model. Simulation is considered (a) parsimonious, (b) more consistent with impairment evidence (such as autism), (c) confirmed by evidence from introspection, and (d) in line with precedent from the study of analogous capacities (e.g. imaginative visualization). Within the bounds of the dialectic, simulationists have criticized explicit *and* tacit versions of the theory-theory as mere variations on a theme that suffer from the same core flaws (Goldman, 1995).[20]

---

[19] To understand Fred, I do not consciously take up the desire for beer, myself intend to drink beer, and only then cut off the rational procession of my mental states to attribute this intention to Fred. Folk psychological judgments usually spring from a rapid, sub-personally conducted process. My practical reasoning faculty takes the pretend inputs of the initial conditions and provides a recommendation as if it were for me in an "off-line" mode, where it is not "hooked up" to my primary, first-personal reasoning.

[20] Of the positive arguments for simulation over theory, parsimony is immediately suspicious. The model is not that simple. The initial sketches of the simulation theory require both a faculty of practical reasoning, and a faculty for imaginative identification or re-centering (e.g. Heal, 1986; Goldman, 1989). The first faculty is taken as a common cognitive resource. We, including young children and autistics, can all entertain beliefs and desires about the world, and reason about them toward conclusions. The further ability to simulate this process in others requires the ability to determine the relevant differences in a situation, feed these pretend states into the practical reasoning faculty, and take the outputs as indicative of the states

The shape of the debate targets the plausibility of an all-theory explanation. Practical reasoning is at the core of the simulationist model, but simulationists admit there are a number of additional functions involved which may well involve theory structures. The key debate is not whether there are *some* knowledge structures in use, but whether there are *only* knowledge structures in use. It is telling that there has not been an emphasis on how well a simulation structure matches the observed data; instead the focus has been on proving or disproving the plausibility of an all-theory view. In the next section we look at

---

of the subject under consideration. This process is opaque and the steps are invisible to the person actually doing the reasoning.

The appeal to parsimony by simulationists is suspicious for this reason (Stich and Nichols, 1992). Characterizing the simulation model as driven by a psychologically-opaque process exaggerates the distinction with a theory constituted of many interlocking axioms and inference rules. Though we do not consciously experience them, there *are* several internal steps where the simulation might go wrong. We must accurately judge the subjects mental state and different situation, translating them together into the correct pretend inputs into my system. One's own reasoning system should be permitted to run, and itself be relevantly similar to that of the person being modeled. Then, with the outputs in hand, any non-rational influences need to be accounted for—influences such as temporary insanity or excessive boldness or intoxication. After these adjustments to control for dissimilarity between simulating mechanism and that of the subject under consideration, the output attitudes must be integrated into the general scheme of beliefs, overriding any personal or quasi-theoretical knowledge I may have previously entertained ("Fred's a tee-totaler!"). Simulation is a very complex system, so its parsimony over theory-theory is not obvious.

Issues of parsimony meant to favor "simple" simulation over "complex" theory appear to be misleading, as elaboration of simulation architecture yields a rather complicated picture of interactions (Stich and Nichols, 1992, 1996; Gordon 1995).

a notable exception to this pattern, in the proposed test from Stich and Nichols (1992, 1996).[21]

1.4 The Empirical Standstill

Given the structure of the debate, empirical tests should be able to adjudicate some of the key issues. This has not happened. On a string of issues, each camp's model has matched the performance of the other. Both views appeal to accounts of development to explain the peculiarities of 4-year-old performance on false-belief tasks: theory must be learned, while a simulation organ must mature (Gopnik and Wellman, 1995; Goldman, 1995; Harris, 1995). Impairments such as autism can equally be understood to impair either the operation of mechanisms or damage the relevant store of information (Leslie 1987). Error patterns and "cognitive penetrability"[22] issues can be due to gaps in the theory or systematic inadequacies of the simulation mechanisms (Stich and Nichols, 1992, 1996, 1998; Heal 1996; Perner, 1997).[23]

---

[21] There the focus is on testing the simulation concept itself. If simulation is true, they say, there can be no errors of prediction. If there are, we see that there must be at least two devices including a separate (faulty) one for predictive purposes), not only one. Of course, this second device *could* be relying on simulation – though it is a faulty simulation. So this discovery will not settle the issue of whether *simulation* is true or not, even while it weighs in on the one-device-or-two question.

[22] Such as the systematic failure of most subjects to predict the "irrationality phenomena" of Nisbett and Ross or Kahneman and Tversky's cases.

[23] One area where this symmetric pattern of argument has not appeared, drawing on non-experimental sources of evidence such as a priori arguments for simulationism (Heal, 1994) or pure introspection (Goldman, 1986), have proved even less promising.

Neither side has been able to deliver what Heal (1996) calls a "quick empirical knockdown" on any of the issues considered to date. For each new consideration that exploits an intrinsic feature of one view, the opposed view is able to adjust their model to match the new criterion. Let me consider a bit of detail for two such issues, to see precisely how and why this pattern has developed.

Consider one theory-theorist hunt for an "empirical knockdown", where Stich and Nichols (1992, 1996) introduce a consideration meant to decisively favor theory-based approaches from simulation. A simulation approach, they argue, cannot produce false predictions insofar as the approach consists in using cognitively similar mechanisms to those of the subject under consideration. Yet they point to evidence where people consistently fail to predict how others will behave in cases where "irrationality phenomena" produce counterintuitive behavior. For example, test subjects given a choice of lottery tickets assign their tickets higher value than subjects given no choice ("the Langer effect"). Yet interviewed subjects failed to *predict* that experiments would produce this valuation disparity. Stich and Nichols contend that this is evidence that the interviewed subjects are not simulating the test subjects. They argue that only a theory is cognitively penetrable—fallible or incomplete in a way that permits this type of error.

Harris (1995) and Heal (1996) respond that simulation models *also* can be cognitively penetrable, and that a variety of effects might be blamed for the failures of prediction. Interviewed subjects may inaccurately assess the starting conditions of the test subjects,

failing to recognize the choice vs. no choice feature of the experiment and focusing erroneously on some other aspect. The manner in which the situation is explained to the interviewed subjects, written descriptions or videotaped sessions, may obscure inputs necessary for proper simulation. The time permitted for the interviewed subject to consider the situation may not suffice for full simulation of the situation[24]. In addition to these "pre-simulation" or "upstream" errors that may accrue in the construction of pretences for input to the simulation, there may also be errors in interpreting the outputs of the simulation. The "post-simulation" or "downstream" recommendations of the practical reasoning module may conflict with strongly held theoretical views, or be otherwise misused. Finally, the simulation activity itself may be interrupted, disturbed, or modified by intense emotion, lack of time, or other non-rational factors. Nichols and Stich (1998) conclude about the cognitive penetrability criterion: "it was a mistake to claim that pretense-driven-off-line-simulation is not cognitively penetrable." Other arguments for the theory-theory have gone this way, such as the appeal to diachronic development of theory of mind as evidence that it was explicitly learned rather than innately endowed as simulation capability. The simulationist can keep adding extra processing steps, including knowledge structures, to meet the test behavior offered by a critic.

---

[24] In Stich and Nichols 1992 observations, the test subjects picked lottery tickets one week before they were asked to value them, while interview subjects were presented with the situation and asked for an immediate response. In an improved 1995 version, methodological problems and inconclusive results continued to obtain, says Perner 1997.

Consider a second hunt for an empirical knockdown, where empirically founded arguments for simulation have also seen the theory camp evolve its position to better resemble simulation. In their original papers on this topic, Gordon (1986) and Goldman (1986) appealed to evidence from autism to support the idea that folk psychological ability was not merely a matter of theory but of special-purpose mental organs. Autistic children have a deficit in the capacity for pretend play, a fact that some simulationists believe is linked with their universally poor performance on the Wimmer-Perner false-belief task. Autistic children do not develop a capacity for attributing false-beliefs to the puppet, even though their general cognitive abilities are not so radically impaired. Down's syndrome children, with much lower average IQs, suffered no such impairment. The implication was that general theory-forming ability does not play a role in the capacity for theory of mind, which develops quite separately.

In response, theory-theorists have simply contended that the capacity for belief-attribution is underwritten by a specific body of tacit knowledge, not by general intelligence or capacity for explicit theory formation (Leslie 1987; Fodor 1992; Stich and Nichols 1992). Elements of folk psychological ability do require the inter-operation of distinct sub-faculties—such as belief-attribution or desire-attribution—but these abilities are explained by the possession of tacit theory, not simulation. The theory-theory can keep evolving to meet test behaviors offered by critics, by taking the folk psychological processing further away from central, domain-general cognition into specialized, encapsulated, innate devices.

The pattern of argument raises the issue of whether *any* purely empirical consideration can distinguish between simulation as such or theory as such. There are certainly specific formulations in play on both sides that stand to empirical confirmation or rebuttal. For example, the evidence from autism as well as the regular and universal pace of development of folk psychological ability both indicate strongly against the "little scientist" picture. But such piecemeal arguments have not shifted the balance of the overall opposition, since both sides are resilient.[25]

Why should there be an empirical standstill? One explanation could simply be that the situation is temporary. The discussion will eventually be settled by a consideration that is shortly forthcoming. There is no reason to rule this out, but there is so far no hint that such a resolution is on the horizon. This paper considers a different explanation: that there is a mistake underlying the construal of the disagreement as theory vs. mechanism.

2. Threat of Collapse

Concrete experimental data has repeatedly failed to discriminate between the two camps. What one camp explains, the other evolves to match. Observing this pattern, Nichols and Stich (1998) openly worry that the debate has deteriorated into a nominal dispute only.

[25] The empirical contest has pushed each side in predictable directions. The theory camp, which began in Gopnik and Wellman's formulation as a learning-based, explicit knowledge store run through general cognition, has increasingly become isolated into a peripheral module with encapsulated information flows and innate databases. The simulation camp has added chunks of knowledge to get the simulation going, such as rules for constructing pretend beliefs, which are incorporated into extra non-simulating processes steps.

Perhaps there is nothing of real substance at stake, but only the right of one camp to claim victory. Perhaps, they speculate, the theory-theory and the simulation theory do not describe views that are fundamentally different at all.

One way to put this worry is to say that the definition of the theory-theory is too broad. The theory-theory, as defined, includes any view positing an "internally represented knowledge structure". As we have seen from Chomsky, it is pretty easy to posit such a structure. If a person can do something, then they have a knowledge structure to explain their competence. They "know" how to do something, like turn declarative sentences into polar interrogatives. *If that is all it takes*, then simulation theory will surely count as a kind of theory-theory. If there is a simulation process implemented in the mind that explains a person's folk psychological ability, this is just a detailed account of that person's tacit, innate knowledge structure. The simulation theory suggests just one way that a person can have such a tacit theory of folk psychology. As a result, there is no deep distinction between simulation theory and theory-theory at all.[26] The distinction that seemed to motivate many of the arguments in this debate faces a "threat of collapse" (Heal, 1994).

This line of worry is supposed to show that interpreting "theory" too broadly can lead to trouble for the polemic. We should *redraw* the camps to avoid this threat of collapse, this reasoning goes, to faithfully represent *what is really going on* in the folk psychology debate. Some participants in this discussion take this as their explicit aim.

---

[26] The debate would be come, "Yes, the theory-theory is true. But is the simulation theory *also* true."

This is a mistaken line, on my view, since the present situation suggests precisely that something might be wrong with that polemic. The camps have been drawn faithfully. "What is really going on" may be a confusion. The collapse of the distinction may in fact be an insight. So we should look for more general grounds to protest this collapse.

There is a second possible aim for this line of worry, and it goes deeper. This aim looks at the wider consequences for "tacit theory" explanations in cognitive science. In folk psychology and elsewhere, we need a workable definition of what counts as tacit theory. Perhaps the very concept of tacit theory is *trivial*, so easy to apply that any mental or non-mental structure counts as a "theory" by its lights.

This chapter takes this second aim seriously. What is a reasonable account of tacit theory? Once we have such an account, we will know whether the simulation theory counts as just another variety of theory-theory.[27]

---

[27] There are two ways for the distinction to collapse. One is if theory is a trivially broad concept. We will consider that in the body of the chapter. As an aside, it may also be that *simulation* is trivially broad. And therefore no theory-theory can fail to include simulation elements.

Heal tactically construes simulation to include any models exhibiting *any* reliance upon the similar functioning of my mind's to the other's case. Heal's definition of simulation is extremely broad, as Stich and Nichols complain. The result may be that the simulation theory is trivially true; any proposal for folk psychology *must* include elements that Heal would count in favor of her view.

Many mundane prediction tasks will require appeal to certain beliefs about which I cannot possibly have theoretical resources ("information or misinformation") in advance. Consider a somewhat

specialized case, where Nichols must predict what Stich will respond to the question "Who will succeed the president if he dies?" Nichols may never have discussed presidential succession rules with Stich, though it seems reasonable for him to assume that Stich's relevant knowledge is similar to his own. But that presumption of similarity, even if it is itself a theory-constituent rule of belief-attribution, invokes a simulation on Heal's construal. This type of *filling-in* simulation is likely to be promiscuous: does Stich know he has ink on his shirt? Would he condemn terrorism? Will he like spicy tuna maki? None of these cases requires us to *reason* or apply decision rules as Stich does; they only depend on knowing a key attitude of his. Since we don't know his preference about tuna or belief about his shirt, we simply fill in this fact. These cases ignore folk psychological reasoning altogether.

Perner (1996) suggests that whole categories of knowledge about others are entirely dependent on filling-in simulation, such as knowledge about what others will judge grammatical or obvious or funny. To predict what someone else will accept as grammatical, I simulate their grammaticality judgment with my own language system and attribute my outcome to them. This type of argument suggests that simulation must be true, on the construal given by Heal, but not in a way that addresses the central debate. They depend on simulation for filling in perceptual, preferential, or accidental psychological facts about someone. This trivial type of simulation does not get at the question about knowing how others will think about an issue.

One possible reaction to Heal's trivializing characterization is to call simulation a "theoretically uninteresting category" (Nichols and Stich, 1998). Continually weakening the standards under which an account qualifies as a simulation model does not help advance understanding of the underlying mechanisms. Virtually every theory-theory account, accepted without revision, would count as a simulation model since they all need to explain judgments about what is funny or obvious. Surely the more interesting debate requires a stronger picture of simulation. If, by redrawing the borders on what counts as a simulation model, Heal is simply weakening the criteria to declare victory, then perhaps there is no use distinguishing the two camps at all.

The situation has not deteriorated that far. Simulationists have not adopted Heal's weak strategy. Nobody is claiming victory because simulation is required for judgments about specific subject matters.

2.1 What Is a Tacit Theory?

In trying to characterize the contours of the dialectic, Stich and Nichols (1992) provide an account of the theory-theory as deploying the "dominant explanatory strategy" of modern cognitive science by positing "internally represented knowledge structures". There is no epistemic requirement, so the posited structures don't have to be full-blooded knowledge. They may therefore be tacit, unconscious, and inaccessible subdoxastic states. The content of the theory may also be false. It may predict people's behavior using concepts or laws that don't actually exist in their subjects' minds. The knowledge structure is also free from a coherence of subject-matter requirement; it might just be a hodgepodge of various propositions on multiple subject matters. This is a rather liberal account of what counts as theory, yet we saw in section 1.2 that it is faithful to how theory-theorists defend their own view.

We attribute a theory to someone when they possess a set of intentional, belief-like states. It is a *tacit* theory if the person can't explicitly report it. That the states are intentional means they contain content about the world or other psychological states. That they are

---

The simulationists still maintain that we predict Fred's movement toward the fridge for a beer by turning beliefs about Fred into pretences. They are rejecting that these beliefs are simply submitted to some set of judgment rules embodied in a knowledge structure. So it is premature for Nichols and Stich to call simulation a trivial category. Nonetheless, there will be other reasons to doubt that simulation is a theoretically interesting category later on.

belief-like means they have a particular causal-functional role in the mind with respect to other belief-like states and with respect to actions.[28]

A theory is a body of mental states. The states are described by sentences, and have propositions as their content. Theory does *not* mean the same thing as "scientific theory" or ordinary uses. Theory, here, is interchangeable with psychological theory, knowledge structure, set of belief-like states, and other related usages.

The question, then, is whether this account of tacit theory works or whether it is so broad as to be trivial. It will be trivial if too many things count as theory. In the context of the debate over folk psychology, the question will be whether we can distinguish simulation theory from theory-theory given this account.[29]

---

[28] They are "subdoxastic states" à la Stich (1975).

[29] There is a line of concerns that we will not consider in the main thrust of the chapter. Instead let me discuss them here. The worry is that *part* of simulation includes some theoretical machinery. That is, the worry is not that simulation is itself entirely theoretical; rather, it is that simulation relies in some crucial way on bits of theory.

Goldman (1986) considers a first worry about simulation raised by Dennett (1987):

How can [simulation] work without being a kind of theorizing in the end? For the state I put myself in is not belief but make believe belief. If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what 'comes to me' in my make believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases knowledge of the imitated object is needed to drive the make believe 'simulation' and the knowledge must be organized in something rather like a theory. (100)

Dennett describes "theory-driven" simulation of a suspension bridge (Goldman, 1986: 85). A simulation requires the simulating device to operate in a way that "maintains some relevant isomorphism to" the behavior or the system being simulated (Goldman, ibid.). A theory-driven system does this by setting values to the initial-state variables in the system, calculating the changes from successive application of the system's laws, and taking the end states as indications of how the real system might go. This is how computerized climate models or economic models run.

But Goldman suggests that there is another type of simulation as well. Instead of calling upon knowledge of physics and engineering, I might simply build a model suspension bridge (perhaps to 100% scale) and observe what happens when the wind blows, I will be able to make predictions about suspension bridges without applying rules from physical or engineering theory. This example of "process-driven" simulation takes place outside the mind, but it can also take place in the mind. This is what happens when I take a drug to predict how that drug will affect Fred, for example. And of course, this is what Goldman claims is going on when I take make-believe beliefs as input into my practical reasoning system. Pretend beliefs are not an alien kind, like a suspension bridge, but are "relevantly similar to [normal] belief states" (Goldman, 1986: 86), and the simulationists contend that our practical reasoning system simply takes them as ordinary inputs.

Dennett's worry was that the process described by the simulationists is actually beholden to knowledge of a theory. We have no psychological system for dealing with make-believe beliefs, he says, so I must be relying on some theoretical knowledge. But we can see that the simulationists are proposing a different picture, where pretend beliefs act as real beliefs. When the practical reasoning system runs off-line, it is running as normal on what it takes to be ordinary beliefs. This is a process-driven simulation. You don't need to "know" anything about folk psychology for it to work, just like you don't need to "know" anything about suspension bridges to use the model bridge.

A different kind of worry is that the off-line system, though a simulation, cannot work without some theoretical background. To come up with the right inputs, a person has to "re-center" themselves or identify with the subject under consideration. No simulation method has been proposed for this. It must

simply proceed from what the person observes: Fred has been working all day in the sun, has not had anything to drink, and is in the habit of drinking a beer when he comes home. Picking out these facts as relevant to the situation requires some judgment formula, or a *re-centering formula*. Some of the facts need to have their consequences drawn: the implication that a hard day's work in the sun causes thirst needs to be inferred through use of some rule or knowledge about work. If the person has no drinking habit, the pull of Fred's daily beer needs to be available via some theory. All of this must rely on knowledge (perhaps tacit) of psychological principles. After this analysis of the observed information, the person is ready to subject themselves to pretend-thirst, pretend-habit, and pretend-beliefs about beer in the fridge. The simulation proceeds from there to the intention to get a beer.

There are some post-simulation steps also. As Davies and Stone (unpublished) point out, all of the major simulationists concede that a further piece of psychological knowledge is required to take the outputs of the simulation and attribute specific states to the subject under consideration. There are some options here, but the general point is this: the simulation must be taken as justifying an inference about the mental states of the person being simulated. This *attribution principle* will state that the output states of the simulation—e.g. an intention to get a beer— are the states that Fred would have under similar conditions *ceteris paribus.* So, Davies and Stone (unpublished) say, "it has been explicit since the beginning of the debate that attributions based on mental simulation must rely on at least some elements of psychological theory."

This problem is rather like Perner's judgments that require simulation, such as concerning things that are obvious or funny. That case showed certain peripheral judgments to be intrinsically simulation based. To know if Fred will find this funny, I first see if I will find it funny. The present case, however, shows that there must be theoretical knowledge in the process without contesting the status of the central phenomenon. Theoretical knowledge is required to *attribute* output beliefs to others, but how do I get from the input observations to the outputs? The key issue is how we predict and explain transitions between mental states—going from the input desire for beer and the input belief that it's in the fridge to an output intention to get a beer. For that, the process-driven simulation story is still quite different from one based on knowledge of rules or principles of cognitive function.

2.2 Attributing Theory and Computationalism

We now have an account of theory in purely mental terms. But how do we know when a person has such a set of mental states? After all, they may be tacit and inaccessible, so the person cannot report them to us.[30]

One reasonable way to attribute theory is by looking at the computationalist context of our theory of mind. A common commitment of everyone in the debate is to the computational theory of mind, where mental states and activities are identical with physical symbols and their interactions.[31] We solve the problem of materialism and

---

[30] One traditional way is by following Chomsky's appeal to the best explanation for a given cognitive capacity. If positing a body of tacit theory is what best explains a human competence, then we should assume that the existence of these states is indeed the explanation. The simulationist will not find this route very satisfying, since their contention is precisely that their non-theoretical view does a better job of explaining the phenomena. Yet as we have seen, the empirical results leave us at a standstill. So the Chomskyan inference to the best explanation is not helpful for attributing theory in the case of folk psychology.

[31] All this depends on the computational theory of mind, but it is a broad version of computationalism. You do not need a Turing-compatible, "classical" system for this to work. Even in the case where a connectionist neural network implements our practical reasoning module, the causal-explanatory role played by that system will be equivalent to that of *some* theory in the overall system, though not necessarily the *same* theory that the classical computer implements (connectionist network or rules-based sentential system alike are instances of "theory" on the liberal Stich and Nichols, 1992 construal, see p. 135). So there will be a meaningful level on which the two architectures are computationally equivalent (*functionally* equivalent, if not *strongly* equivalent; Pylyshyn, 1984). At a minimum, there is a high level of functional description where both theories make the same predictions even if they do it via different calculating steps.

mentalism by appealing to the idea that the mind is a computer, which gives us the relation between mental structures and their implementing physical structures: "the relation between an implemented computation and an implementing system is one of isomorphism between the formal structure of the former and the causal structure of the latter" (Chalmers, unpublished).[32]

We can use the "mirroring" relation between physical systems and cognitive systems to identify where a physical system implements a theory-having cognitive system (Block, unpublished). If a person has a tacit, inaccessible belief-like mental state M with the content *Q*, there must be a physical state P somewhere whose functional-causal relations to other physical states mirrors M's logical-rational role. As Block puts it, "causal relations among those [physical] symbol-states mirror useful rational relations among the meanings of those symbols".

If it turns out that there is a way to distinguish between a neural network account of practical reasoning and a Turing-computational account, it will not be on the level of inputs and outputs in general. We must consider specific proposals for the successions of states and their constituent parts: one of them uses the concept of "stress gradient" while the other only stipulates "breaking point", for example.

[32] "Isomorphism" is back. We said before that it was the crux of a simulation. Here we see that isomorphism is also key to computation. The *meaning* of isomorphism in both contexts is the same. But nothing follows immediately from this. I hope it does not cause any undue confusion. In fact, it is suggestive of where the argument is heading: that simulation is just a special case of theory.

If a person has a given theory T, they have a particular assortment of mental states M1, M2, M3, etc. with particular logical-rational inter-relations, R1, R2, R3 etc. This array of states and relations forms a logical structure. The computational theory tells us that this person will have an isomorphic array of physical states (brain states, probably), P1, P2, P3, etc. and causal relations C1, C2, C3, etc. To establish that some arbitrary person possesses a theory T, we need only look for this physical-causal structure. To isolate a particular belief-like state such as M1, we locate P1 within that system.

An explicit theory, of course, has some physical implementation in the brain. That set of brain states implements that *particular* theory purely in virtue of its physical structure. The appeal to computationalism shows us that an explicit theory and a Chomskyan tacit theory will look a lot alike from the physical perspective. They will both be sets of physical symbols with intentional contents. Their causal relations will link them to the production of action and the production of belief-like states via rational reasoning processes. The principal difference, from this perspective, is that tacit theory will not be available to conscious review, merely a feature of its information flows.

With the computational theory we now have a physical account of theory-possession. On the view we have so far, the theory-theory explains folk psychological capacity by appealing to sets of intentional states. The intentional state that says “beliefs lead to action” will be implemented by a physical state with that symbolic content and the mirroring rational-causal relations to other intentional states. In effect, the theory-theory posits physical states with particular causal roles.

2.3 Is Computation Vacuous?

Now we can ask whether this account of theory works. Searle (1990a) raises a worry that "computation" itself is a vacuous concept. He advocates the very strong position that every sufficiently complicated physical system implements every computer, so that Searle's wall is a computer running WordStar. Since a computer is a physical system whose dynamic operation mirrors the structure of logical relations between propositions in a theory, Searle's view makes every physical system an implementation of every theory (also Putnam, 1988). A theory is just a set of intentional psychological states. Since psychological states are computational states on the computational theory, Searle can find mental states anywhere he can find a computer.

If computation is a vacuous concept, then any physical system will count as implementing a theory of folk psychology. Thus, Searle will have undermined the concept of theory in psychology, since attributing theory would do no explanatory work. Theory would be everywhere.

In part, the computation of Searle's WordStar consists in a succession of computational states with particular semantic content. By looking at the progression through time of any complex physical system, Searle can find an isomorphic succession of states. But it is a mistake to say that therefore this physical system is computationally equivalent to an actual digital computer running WordStar. While Searle can retrospectively designate the various physical states as corresponding to computational states in the program in a manner that matches their logical structure, Searle's wall will not support the same

counterfactuals that the computational system should (Block, 1995; Chalmers, 1996). If Searle's wall is isomorphic with WordStar printing a capital "N" on the screen, it would not support the *counterfactual* case where the user typed a capital "B".[33] A digital computer that genuinely implements WordStar would be able to enter physical states corresponding with these various inputs, while an arbitrary physical system cannot both maintain isomorphism with the logical system and keep its semantic designations constant (this bit of wall at time t0 = the letter "N" on the screen).[34]

Searle is therefore wrong to worry that computation is vacuous, but we do see that computation is a pretty broad idea. Any physical system with the right internal causal structure (including counterfactuals) and the right symbolic relationships will count as the implementation of a given theory.

2.4 Rule-Fitting vs. Rule-Guiding

Blackburn (1995) raises a worry that tacit theory can be attributed in too wide a range of circumstances if its only criterion is conformity to a hypothesized rule. Any proficiency can be characterized as the "tacit (very tacit)" possession of a theory of the relevant subject matter. After all, any psychological ability *must* be implemented by a physical set of brain states. And if such brain states are the physical correlates of a computational

---

[33] Nor would the rule *project* properly onto future cases. While Searle can describe the first "N", when I tap "N" again in the future he will have trouble. The wall won't respond properly when I hit "N" again.

[34] Evans, (1985), Peacocke (1986) and Davies (1987) develop a usage of isomorphism with a special meaning for matching the logical structure of systems and their component states.

system as we have assumed with the computational theory of mind, then there will be *some* theory such that they are its implementation. For any psychological competence, we immediately have a tacit theory.

Davies and Stone (unpublished) rightly point out the similarity between Blackburn's worry and Quine's (1972) challenge to Chomsky. Quine asks how we should distinguish behavior that simply "fits" a (tacit) rule from behavior that is actually "guided" by that rule. After all, there are innumerable true descriptions of any behavioral pattern. There are many rules I might use to convert a declarative sentence into a polar interrogative, or to interpret the edge of a two-dimensional image. If we simply permit any *description* of behavior to count as a piece of tacit theory, we will hopelessly proliferate such attributions.[35]

The answer to Quine's worry and Blackburn's is similar. They each worry that any true description will become attributable as a tacit mental state. The solution is to give some standards for theory attribution that avoid this trivialization and still keep with our computationalist treatment of tacit theory.

Peacocke (1986) considers this issue with a case from Evans (1985). They translate this question into a problem about the attribution of a meaning theory. When someone knows the meaning of 100 sentences, how can we decide which theory to attribute to them? The

---

[35] Quine's proposal is that the rule be explicit: "guidance requires verbalizable knowledge" (Davies and Stone, unpublished).

person may not know *how* they know these meanings, as indeed in the classic Chomskyan case. Yet if we are to legitimate the attribution of one theory over another, we need individuating conditions to distinguish tacit theory X from tacit theory Y—*even where they explain an identical pattern of behavior*. This case is evidently like our problem in folk psychology, where two different explanatory strategies are unable to produce distinct behavioral or testable predictions.

In Evans' case, a speaker S understands ten predicates and ten names, and therefore understands the 100 sentences that can be thus produced. Another speaker U understands each of the 100 sentences as unstructured. At one level, a description of the meaning theory for these two speakers will be identical. They can take all the same inputs and map them to all the same outputs. They compute extensionally equivalent functions, which for Marr (1982) is a computational equivalence at his Level 1.

But spelling out a detailed algorithm by which speaker S takes a sentence S1 as input and maps it onto a meaning M1 will require appeal to axioms that link names to things and predicates to satisfaction conditions. A meaning algorithm for speaker U will require no such elements. On Peacocke's account, the two speakers do not deploy equivalent algorithms in their computation of the sentence meanings. But how can we establish whether speaker S is indeed deploying a distinct and sophisticated meaning theory or merely the same mapping from sentences to meanings as U?

Peacocke suggests that the speaker's tacit possession of a particular meaning rule requires that "a state which carries the information drawn upon is causally influential in the operation of the algorithm or mechanism; indeed it requires that the algorithm or mechanism produce states with the content they do in part because of the content of the information-carrying state" (1986: 102). What it is to possess a tacit rule is to have a physical state which corresponds in its causal role to the logical role played by the rule; it "is a single state of the subject that figures in a common causal explanation of a battery of transitions [between mental representations] that conform to the rule" (Davies and Stone, unpublished: 23). Possession of a theory is having actual cognitive states that correspond to all the relevant rules of the theory.[36] Exhibiting this causal structure between cognitive states gives the conditions for a finer grain of computational equivalence than merely input-output extensional equivalence. Maintaining the corresponding causal structure is exactly what it is to have isomorphic structure. We can tell when two brains have the same theory by checking to see if they are in physical isomorphism. Similarly, we can tell

---

[36] If I possess a belief "Windswept bridges fall sooner", then I have an *actual* cognitive state that corresponds to that belief. I also have a disposition to have a belief about the Golden Gate Bridge – "The GGB falls sooner." But I don't actually have that belief. And so I don't actually have a cognitive state that corresponds to it until I think that thought. (Davies likes to refer to the rules or lawlike statements in the theory as "theorems" of the formal system; this is misleading to some since it suggests that these are merely deductive consequences of axioms. And so the axioms are what we actually possess, while the theorems are merely possible beliefs. This is not what Davies means. When we have a theory, we actually have axioms, inference rules, and even theorems in our possession. Of course there will also be some theorems that we don't actually have, that are merely possible.)

if a mind possesses a theory—by checking if the mind displays a physical structure correspondent with the logical structure of that theory. [37]

With this apparatus, we can answer Quine's challenge. A particular theory or theory-element actually *guides* someone's behavior if that person possesses a physical state with the appropriate causal relations to mirror the theory. So attributions of tacit theory are not idle, even if they are presently difficult to confirm. There will be some "tacit (very tacit)" theory, but it will not open the door to reckless attribution of *any* theory that truly describes a physical system. Isomorphism places a sufficiently strong restriction on tacit theory attribution, as we saw against Searle.

2.5 Is a Model a Computer?

We rejected the idea that *anything* can be *any* computer, but perhaps there are certain specific physical systems that will inappropriately count as embodying theories. Dennett (1987) raises an example that has been influential in the folk psychology debate. You are asked to predict what will happen to a suspension bridge if the winds gust. How can you give the answer? One option is to appeal to a scientific theory of suspension bridges. If you know the sets of rules and concepts that constitute this science of bridges, you have a *psychological* theory of bridges. You can use this set of mental states to reason about the gust of wind and deliver a conclusion.

---

[37] We might call this Marr's Level 2 of explanation, or follow Peacocke's request that we are here appealing to Level 1.5. In any event, the account of explanation by tacit knowledge that Davies and Peacocke develop from Evans is essentially an explanation by appeal to the computational model of mind.

The second option is to use a model bridge. Knowing nothing about bridges in general, you might simply reason that a model of the target bridge in question is likely to behave in a similar fashion. This physical model bridge has all the same parts with all the same causal relationships as the target bridge. Indeed, using a model bridge in this fashion is deploying a system that is *isomorphic* to the target bridge in the manner suggested by the simulation theory. Using the model is *simulating* the target bridge's behavior; a patently non-theoretical activity, according to simulationists.

Davies and Stone (unpublished) consider a more serious variant of Searle's original worry. On the computational view, any physical system that properly mirrors a set of logical propositions can be said to implement it. The physical structure of the model bridge mirrors the real bridge's perfectly. The computational view seems to commit us to counting this model bridge as a computational system, a body of theory that the you use to reason about the real bridge.

Davies and Stone wonder where this might lead. For *any* physical system, there will exist at least one theory such that the physical system implements that theory. Indeed, this is partly what drove Searle's (1990) concern, that all physical systems implement *some* set of rules. Our reply to Searle was just that this relationship is not *arbitrary*; there are structural restrictions on how promiscuous this is. One wall does not implement *every* computer. But, accepting this limit, we might still object that the wall implements a

theory *at all*. Searle's charge of vacuity was that a) the wall implements a computer, and that b) it implements *every* computer. We have only answered the latter charge.

A consequence of the former charge, however, is that a suspension bridge does implement *some* theory. Call this theory B, a set of physical and engineering propositions that perfectly describe the bridge's microstructure. Say there is a variable in that theory for wind speed with only two settings: *L* (light) or *G* (gusty). Say also that there is a setting of particular variable that draws its value from various other variables in the system, and that it too has only two settings: *C* (collapsed) or *S* (standing). We can make an observation about this theory that when winds are *G*, then the bridge is *C*.

The physical system that implements *G* does so in virtue of its physical structure alone. So there must be some physical fact that corresponds to the settings *L* and *G.* Indeed there is: the physical speed of the wind applied to the bridge. Of the bridge we can also observe that when winds are gusty, the bridge is collapsed. When winds are merely light, the bridge is standing, just as the theory connects variables *L* and *S*. The physical conditions of the model correspond to the variable settings of the theory. It is in this sense that the physical structure *mirrors* or *is isomorphic to* the logical structure of the theory.

In virtue of this mirroring structure, the physical states of the bridge implement computational states with the engineering propositions of B as their content. The bridge implements a *theory* of bridges. The theory takes inputs about wind and produces outputs about the bridge condition. The model takes real wind gusts as its input symbols,

produces collapse- or standing-conditions as its outputs, and does it all by computing through the relevant intermediate states. The result is a set of states that represent predictions about the target bridge.

The model bridge is what Goldman (1986) considers a paradigmatic *process*-driven simulation, the type that is patently not "theory-driven". Yet the scale model that I use to simulate the effects of wind gust will itself be the implementation of a set of rules for how bridges behave. The scale model is a simulation, since it maintains the appropriate similarities of structure with the target bridge. But it clearly meets the criteria for "implementing a theory", in this case B.

Consider the brain of an engineer that explicitly knows theory B. There is a set of psychological states corresponding to B in her mind. When we look at her brain, we will find some set of physical states P that are the implementation of this psychological theory. As computationalists, we say P implements B purely in virtue of its physical structure.

Compare this to what a Chomskyan might argue. Say he discovers that children are able to understand and predict bridges *as if* they have theory B, without having been taught anything. He might posit that they have an innate, *tacit* endowment of B. Unlike the engineer, the children do not have explicit knowledge of those rules. They "cognize" B. What is Chomsky's standard for attributing a tacit theory here? He observes their ability and infers the existence of a set of mental states that cause the children to behave just like

someone who knows B. Just as in the case of tacit theory of grammar, we might suggest that there is a tacit theory of bridges at work.

Just as for the engineer, the child's brain contains some set of physical states P^2 that implement this tacit theory B. And again, it is purely in virtue of their physical structure and interconnections that P^2 is the implementation of B. It is the mirroring relation between the structure of the engineering theory and the structure of the brain states that is key. The requirement is to find a physical system with the right structure.

When we turn to consider the model-user as he thinks about the suspension bridge, we will again find the right structure. Purely in virtue of its physical structure and causal relations, the model bridge perfectly mirrors the structure and role of the engineering theory. Furthermore, a person is using this physical structure to make predictions about suspension bridges. The person *meets our requirement* for possessing a tacit theory of bridges. Just as the child's P^2 implements B, we have no resources to stop the implication that the model bridge also implements the tacit theory B.

Let's clarify a bit what this implication means. The model bridge is not "a person who knows a theory of suspension bridges". The bridge does not *know* anything or possess any theories, since it will be incapable of having an attitude towards its intentional states. However, when a novice person has that model bridge, he has a physical system that is isomorphic with an accurate theory of bridges B. It produces accurate predictions of what the bridge will do. For comparison, an engineer who learns B will have a complex set of

beliefs in her head. That network of beliefs takes inputs about bridges and produces true predictions about the bridge. The novice and the engineer answer many of the same input-output questions in virtue of structurally similar physical structures.[38]

---

[38] Some readers may find the entire bridge case overly simplified. Recall Davies and Stone's drug case. I simulate the effect of a drug by actually taking the drug, while the expert relies on his neurotransmitter suppression theory. Does this work the same way, such that the drug taker has a *theory* of the drug's effects?

Yes. First note the differences: the drug user has a crude theory, while the expert has a sophisticated theory. They are not the *same* theory. For example, the expert's theory might be adjustable to people with different body types or to different environmental conditions. Various concepts in the expert theory can be used in sophisticated ways, and the expert can explain the steps in the process of intoxication. The drug user can do none of this. Indeed this expert theory may go *beyond* simply explaining the behavior of the subject under consideration. It may also fall short of full explaining, by use of shortcuts or heuristic rules in its structure to gloss over particular mysteries in ways that are inconsistent with the real details.

There is something the drug user *can* do however: predict the first-personal experiential outcome of ingesting the drug.

Imagine an expert with a very simple theory which is true but not detailed. It says simply that the drug THC cuts levels of the neurotransmitter X by 50%. And that X levels correlate perfectly with paranoid experience and behavior. This expert has a certain set of physical states that implement this theory. There is some physical state P^{THC} that represents the concept THC, and its role on X. We should see some physical connection between P^{THC} and P^{X}, for example. This set of physical states are probably situated in the broader set of states that make up general cognition.

A novice with a model bridge cannot do many things that the engineer can, however. He cannot describe the lawlike relations that govern various aspects of the bridge. He can't revise his beliefs if someone tells him that some law is incorrect, and come up with different results. He cannot stop his calculation halfway and observe the intermediate results. So he does not have full-blooded knowledge about bridges. However, the novice does meet the conditions for having a set of psychological states that constitute a theory. He can produce all the same true predictions about the bridge's end states by observing his model. And he does it by manipulating a perfectly isomorphic set of physical states. The model's states are not *brain* states. If something has to be a brain state to be mental, then that is the only criterion the novice does not meet. The person plus the model, however, constitute a physical system with the same behavioral characteristics as a person. This is like the difference between an expert linguist and a native speaker: both can produce grammatical speech, both use similar physical structures, but their theories have different epistemic status. The linguist's is full-blooded knowledge; the native speaker's is tacit knowledge.

The drug user also has a physical state that corresponds to THC in his system: the THC molecules themselves. He has the neurotransmitter X to correspond to the concept of X. And the two are physically inter-related in precisely the inhibitory relationship described by the theory and represented by the expert's states P^{THC} and P^{X}. If we look only at the physical structure of these states and their role in producing beliefs about drug use, we will find it identical to the simple experts brain. There will still be differences: the expert knows his theory explicitly in general cognition, while the drug users exists in a more distributed state throughout the brain. But in both cases it is a set of physical structures in the brain which entirely account for the person's ability to answer questions about a drug's impact.

The implication of this line is that a model is indeed a theory. For the folk psychology debate, it means that the simulation theory's appeal to a group of brain mechanisms will in fact qualify as a theory-theory. For any set of brain mechanisms, there exists some theory such that those mechanisms are its implementation. So a simulating mechanism is indeed the implementation of a psychological theory.

2.6 Psychological Theory vs. Theory of Psychology

If a simulating model is a theory, then any simulation theory is a theory-theory. Davies and Stone (unpublished) seek to resist this conclusion on the grounds that a model would not implement *the right kind of theory*. In effect, they concede what computationalism implies about physical mechanisms implementing theories. Instead, they seek to distinguish the kinds of theories posited by the simulation theory and the theory-theory. The theory-theory posits a theory *of* folk psychology, a theory about folk psychology with concepts like belief and desire. On the other hand there is simulation theory, which may posit a model that implements a psychological theory, a set of psychological states. But it is not a theory *about* folk psychology.

Davies and Stone try to distinguish simulation models from theory-theory by arguing that there are important differences in the role of a *theory of suspension bridges* and the role of the *model bridge*. While both permit me to make predictions about the suspension bridge, they operate on different inputs and produce different outputs. In reasoning with a theory, I would begin with thoughts *about* the bridge and the wind gusts:

> The target bridge is in such-and-such condition.

The target bridge's material has such-and-such rigidity.

The wind is gusting at X miles per hour.

The theory would take these inputs as legitimating particular inferences leading to certain conclusions, in a manner like a proof given in a formal system. I would end up with thoughts or representational states about the bridge:

The target bridge will sway Y degrees.

The target bridge's left truss will crack.

The model bridge does not take such representations as its input; it is itself *in* such-and-such initial condition and subjected *to* a wind gust. The model ends up in some state, swaying Y degrees or with a cracked truss. But the model itself will not exhibit physical states that *represent* the subject bridge. So while the model might implement a theory of some sort, it is not a theory *about* suspension bridges.

The case of folk psychological prediction is meant to be analogous. A theory of mind would take beliefs about Fred's beliefs as inputs, and give beliefs about Fred's beliefs as outputs[39]. But the simulation proposal does not take beliefs about Fred as inputs; it takes beliefs about the world itself. So the simulation mechanism is not itself a psychological theory couched in *third-personal vocabulary* about how practical reasoning proceeds. It is a practical reasoning device that implements a theory of decision making[40]. The mechanism produces folk psychological decisions from particular inputs about the world. It is a folk psychology computer implementing a "folk psychology program"—a bunch of

---

[39] The inputs and outputs could also be sub-personal intentional states, not necessarily beliefs.

[40] Drawing from propositional attitudes about the world as its inputs.

instruction for deciding how to behave in various contexts (“If you want beer, go to the refrigerator”).

Is this answer enough to settle the problem of models implementing theories? Not sufficiently. Davies and Stone are right about one thing: the model bridge is not the physical implementation of a theory *about* bridges. It does not implement the right kind of theory given their considerations.

But they are wrong to look only at the model bridge. We are not considering the bridge alone. We want to know whether a *person* running a process-driven simulation on a model bridge can be said to possess a tacit theory of suspension bridges. That is the relevant case, because that is the case analogous to folk psychology. When a *person* has a simulating mechanism for folk psychological judgments, we want to know if that *person* has a theory. This is a subtle point; the simulating mechanism is *part* of what the person has at her disposal. But more is happening than the running of the simulation itself. There is relevant activity both before and after the simulation is run. Consider the entire model bridge simulation process: The person *begins* with beliefs about the subject bridge, judges a model to be relevantly similar, applies a wind gust to the bridge, *then lets the model bridge run its course,* observes the results, infers that they are applicable to the subject bridge, then predicts these states of the subject bridge. In the overall simulation, the input and output conditions are indeed *about* the subject bridge in the relevant ways.

The model bridge alone does not match the role of a theory about bridges. But the model bridge is wrapped in input-adjusting and output-interpreting activities that make the simulation process work. This overall process of simulation, partly in the brain and partly in the bridge, takes exactly the right types of inputs.

We have established so far that a) a simulating model counts as implementing a theory, and b) when the model is considered together with relevant adjusting activities, it plays the same role we would expect of a theory about the relevant matter.

2.7 Psychological Inputs vs. Non-Psychological Inputs

Davies (1994) considers a particular simulationist scenario. Let's say the simulation theory is true. The practical reasoning system is the simulating mechanism at the heart of the system. When I reason about Fred's situation, it is the similarity between our practical reasoning systems that makes this simulation work. Typically, we assume the inputs to my practical reasoning system are beliefs like "The beer is cold" or "The beer is in the fridge." In this case, imagine the practical reasoning system works differently. Say instead that it takes input states *about* intentional states (i.e. inputs couched in third-personal vocabulary). So the inputs look like "I believe that the beer is cold" and "I believe that the beer is in the fridge." Of course, the simulation would still work because my system is similar to Fred's.

However, as we saw in the previous section, this simulation system is the implementation of some theory which describes it. So the system implements a theory. But because of the types of inputs taken by the practical reasoning system, it implements a theory explicitly

about folk psychology. It takes thoughts *about* mental states as inputs. On a case like this, we have a simulation model which uses a tacit theory at the core of its explanation.

Trying to preserve the distinction between simulation and theory-theory, Davies (1994) proposes conceding this example, but instead reserving a narrower domain for simulation. An alternative simulation model would give different inputs to the reasoning mechanism. Rather than pretending that "I believe the beer is cold", I might simply pretend that "The beer is cold"—a propositional content without any intentional terms. Davies is revising the definition of simulation here. Only systems accepting inputs in a first-person viewpoint count as simulations. On this model, Davies suggests that the simulating mechanism will no longer be a theory *about* mental states. On this version, the simulation model may still implement a theory, but not a folk psychological theory about decision-making in others. Thus it tries to preserve the distinction between theory-theory and simulation by specifically excluding the peculiar simulation model imagined to start this section.

The difficulty of the previous section again obtains here (Heal, 1994). Davies's strategy is to remove "psychological notions from the beginning and endpoint of a simulation exercise" to prevent the simulation from falling into the same functional role as theory. But this is not possible if a *person* is using the simulation mechanism precisely to predict and understand someone else's practical reasoning. As Heal points out, the whole exercise begins with my forming certain attitudes about the subject's mental states, as I attempt to re-center my perspective and form inputs for the simulation. On the other end,

my practical reasoning module produces results which I then assimilate back into predictions about the subject's behavior.

In the case of the suspension bridge, we considered whether the model bridge might embody a theory of bridges. While Davies and Stone objected in the previous section that the model did not itself mediate transitions between representational states about bridges, this seemed to ignore the role of the person *using* the model bridge. The relevant case is of an exercise in simulation, after all, where a person is using the bridge as an instrument of simulation. For this total system, the person plus the bridge, the input and output states do turn out to be representational states about suspension bridges. As Davies and Stone (unpublished) admit about Heal's argument, "if a mechanism is used to simulate the operation of mechanisms of the same type so as to permit predictions about them then the mechanisms embodies tacit knowledge of theoretical principles about how mechanisms of that type operate" (29).

2.8 Causal Structure

Dissatisfied with this situation, Davies and Stone attempt to draw a finer requirement on the role a piece of theory should play. To see this, let us consider in more detail the procession of states involved in predicting a bridge's behavior. A simulation follows this sequence (scheme S):

S1 – Thought about the subject bridge beginning-state

→ Sa – Re-centering formula

S2 – Model bridge in beginning-state

→ Sb – Physical laws of bridges operate

S3 – Model bridge in end-state

→ Sc – Attribution principle

S4 – Thought about the subject bridge end-state.

S1-S4 are states; Sa-Sc are the drivers of state transitions. S1 and S4 are thoughts about the bridge. From S1, I must make use of information about the subject bridge to properly set up the model, "re-centering" the model's conditions upon the situation of the subject bridge. In this case, S1 might be the thought that "The wind gusts at 100mph on the bridge." I must apply some relevant theory to adapt this input to the model at hand, as I do in Sa. If the model is 50% scale, perhaps I should reduce the intensity of my model wind. If it is 100% scale, then I know that I should try a 100mph wind rather than some other strength. So Sa is a piece of theory, an inference rule I use for modifying the subject bridge's conditions S1 into the format or scale required for my simulation's beginning state S2. If part of S1 is "The bridge is 100 feet tall", Sa might instruct me to multiply by 0.5. S2 would then include "The bridge is 50 feet tall".

The transition from S2 to S3 is a product of the natural laws governing the model bridge itself. Presumably this is a complex of actual natural laws of physics and engineering, which we can capture under a single heading with Sb. Sb is the common causal-explanatory factor in the transition of any bridge from states like S2 to S3. Sb is not the application of theoretical rules; we are watching the concrete model to see what it actually does. From the end-state of the model, I must again deploy a piece of theory to connect the simulation back to the subject, a step governed by some principles of attribution, Sc. Both re-centering and attribution were discussed as ineliminably

theoretical elements of any simulation model above, a fact that simulationists and theory-theorists alike have accepted. Finally, I end up with a thought about the subject bridge, S4.

In contrast to the simulation scheme S, Davies and Stone have the theory-driven version of this process run more simply (scheme T):

S1 – Thought about the subject bridge beginning-state

→ Ta – Psychological theory of bridges

S4 – Thought about the subject bridge end-state.

Ta might be a set of thoughts about principles or equations, but it could also be a single rule or principle. Perhaps there is an explicit rule about suspension bridge construction that "No bridge can withstand 90mph winds." S1 will fall clearly under this rule and give the results that "The subject bridge will collapse." S1 and S4 in this scheme are the same as they were in scheme S.

It is clear from scheme T that the role of Ta is different from the role of the model bridge simulation Sb. Ta mediates the transition directly between S1 and S4, while Sb is clearly not sufficient for that on its own. Sb is sufficient only to move between S2 and S3, leaving a gap on either end. For this reason, Davies and Stone conclude "that the putative state of tacit knowledge…does not play the right causal-explanatory role" (30). This is a finer grained version of the previous objections. The point is not what type of inputs theory takes—what the theory is *about,* or whether there are psychological inputs or non-

psychological inputs. Instead, the point is that the causal structure of a theory-theory looks different than a simulation theory.

Does this objection work? We already admitted that the model bridge on its own cannot count for us a tacit theory *about* suspension bridges. The model does not take the right kinds of inputs. It takes only physical states (like a wind gust) as inputs. Supplementing this talk of tacit theory with a concrete account of what such theories consist in, we find the problem reformulated in terms of causal-explanatory states. The model bridge Sb plays a different role from a theory Ta, because it does not connect thoughts about bridges like S1 and S4. It only connects bridge-states S2 and S3.

Surely Heal's (1994) response will again obtain. The simulating system is composed of three elements:

a re-centering formula (Sa),

a simulation mechanism (Sb), and

an attribution principle (Sc).

These three operate together to take S1 as an input and provide S4 as an output. We know already that this system employs simulation: the causal structure of Sb is isomorphic with the subject bridge.[41] Sb is also isomorphic with the deductive structure of *some* theory SbT which describes its causal structure, and which Sb can be said to compute.

[41] Remember, Sb is the actual laws of nature causing the model bridge to behave however it does.

How can Sb implement a theory of bridge behavior? A simple one that it implements is "If winds blow over 100mph, the bridge falls down." If you apply a 100+mph wind to the model, it falls. It implements this rule the same way an AND-gate implements the logical AND-operator. The deductive structure of the AND-operator is to output TRUE if given two TRUE inputs. The physical AND-gate stands in isomorphism to this logical function when it returns a symbol meaning TRUE only when it has received two TRUE symbols. Sb implements a theory of bridges by responding to inputs in ways that correspond to the rules of the theory. When the bridge falls down, it corresponds to the *logical consequence* of "If winds blow over 100mph, the bridge falls down."[42]

But now we have three inter-linked chunks of theory—Sa, SbT, and Sc—which take S1 as an input and yield S4. The causal role of this clump is no different from Ta's, and yet scheme S implements a simulation and scheme T does not.

---

[42] It's the *logical structure* that matters here. Peter Godfrey-Smith points out a useful distinction from Giere. The sentences of the theory only describe an abstract model. The abstract model, like an actual physical model, is isomorphic to the actual bridge. For Giere, this is an imaginary bridge – it is abstract as opposed to real. It seems more natural to say that this imaginary bridge "resembles" or "is similar to" the real bridge. In my discussion, I take the sentences of the theory to describe *propositions*, not an imaginary bridge. These propositions could be a system of axioms, inferences rules and theorems, or they could be a system of mathematical equations and variables. It is this abstract system of equations—the logical structure—that matters. The sentences themselves, the representations that describe an abstract model, cannot be isomorphic with the bridge. In the same sense, the sentences don't have logical structure. It's the logical propositions that the sentences represent that have logical structure.

Could scheme T implement a simulation? The requirement for doing so is rather liberal, since either Ta itself or some component part of Ta must be in isomorphism with the subject bridge. As long as Ta is true of the bridge, its states will correspond to real states of the bridge. The theory of bridges might include a rigorously bottom-up model of the physics of the bridge's design and materials. The conceptual elements of this theory correspond to the parts of the bridge, and inter-relate by means of mathematical formulas of load-distribution, etc. The wind would be represented as a force of such-and-such strength against particular elements, causing certain energy distributions through the bridge, and resulting in some final state of the system. A detailed theory of this nature may indeed serve as a theoretical simulation of the actual bridge. If such an account is at the heart of Ta, surely it will also require elements analogous to the re-centering formula and the attribution principle (if only at the point where concrete numbers are plugged into variables in the formulas).

On the other hand, Ta could be a single high-level law like "No bridge can survive 90mph winds", and fail to be isomorphic with the bridge's causal structure at a detailed level. But it *will* be isomorphic at a very crude level: when winds are over 90mph, the bridge is collapsed, otherwise not. The theory does correspond to the facts at that level. So a detailed Ta or a simple Ta will both match the structure of the subject bridge at varying levels.

Ta might also *fail* to correspond at some detailed level because the theory is false in its details even as its predictions are correct. A Newtonian theory of motion will match the

structure of the physical system only at a high level of description, even as it gets all the microstructure wrong. It is similar at a high level and wrong at a more detailed level. So the Newtonian Ta will not be isomorphic to the system it describes at a very detailed level.

It is possible to have true theories that fail to function as simulations at various levels. This is weaker than the converse observation we have been making that all simulations implement some theory.

The case we have been considering in detail involves an external, physical simulation. I have been arguing that scheme S represents an explanation of bridge-understanding that attributes a tacit theory. A 100% scale model bridge might seem like an odd constituent of a tacit theory. The bridge is not in the head. The case of the theory of mind capacity does not have this odd feature however, since the simulation mechanism is itself inside the head. (Neither does the drug case, discussed in the notes to section 2.5.) When folk psychology's simulation theory appeals to a simulating mechanism, it is a set of brain states. In that case, the brain states which constitute the simulating mechanism linked into various cognitive systems will count as implementing some psychological theory. As such, any simulation theory will be a theory-theory.

2.9 Collapse of the Distinction

In his original paper on this topic, Davies (1994) begins worrying about the threat of collapse with an observation about what it is to attribute a tacit theory to explain a cognitive capacity such as folk psychology. Attributions of bodies of mental

representations must further involve commitments to particular physical cognitive states, whether we posit a cognitively real linguistic grammar (Chomsky 1965; Higginbotham 1987), a vision system (Marr 1982; Peacocke 1986), or a folk psychology. Where our successful theories for such capacities involve the attribution of systematic or compositional states, we have good reason to think that they are computationally implemented. Any intentional states account of a cognitive ability must ultimately be cashed out in terms of physical mechanism—whether or not they exhibit isomorphic structure.

But the theory-simulation opposition to date seems to be conceived as a dialectic about the *nature* of the cognitive system: is it a body of knowledge or a physical mechanism? The empirical arguments fielded to date are organized around this opposition. Yet, this opposition fails on *a priori* consideration alone, since we know that any body of mental representations must be explained by some physical mechanism. By the computational theory we know that any physical mechanism implements *some* theory. Further, in the case of folk psychological simulation, the simulating mechanism implements a theory with precisely the right characteristics to count as a theory about folk psychology. The theory-simulation opposition collapses.

This does not mean there is no real debate over folk psychology. Davies (1994) considers, as do Stich and Nichols, that the right way to carve the debate is not against what is theory and what is not, but rather against what is simulation and what is not:

> So as we construe the controversy, it pits those who think that prediction, explanation and interpretation are subserved by a tacit theory *stored somewhere other than in the practical reasoning system*. (Stich and Nichols 1992: 135, n7)

The critical issue as they conceive it is simply that simulationists predict we make judgments about others using the same machinery we use to make our own decisions, and the theory-theorists deny it. Simulationists say we have one device: the one we use for making decisions, on which we run some simulations. Theorists say we have two: one for making decisions, another for understanding other people. This puts the spotlight on a specific architectural proposal for the flow of information as the simulationist proposal. This shifts the debate to very different questions than those that have been the main points of contest. Hopefully, the new debate will lead to more decidable results.

Chapter 3 Appendix. Implications for a Modular Psychology

If we accept the collapse of the theory-mechanism distinction, the debate that has occurred over folk psychology looks very odd. Everywhere in the literature, we see arguments based on the assumption that theory is one type of explanation, and mechanistic explanations are quite another. As just one example, consider Spelke's argument from the folk psychology's malleability in children. Since their ability is changing as they age, she says, it must be that they are revising a theory. *Only* theory has such-and-such features, so simulation is false. But arguments like this based on the difference between "theoretical" explanations and "mechanistic" explanations are doomed to fail. They are interchangeable, and the apparent empirical standstill in the debate is proof.

This mistake is actually more general than folk psychology. In fact, the theory-mechanism distinction is a mistake that has undermined debates very widely in cognitive science. Modular psychology has carried this distinction with it since Chomsky started positing knowledge structures to explain linguistic ability. The result has been the mistaken idea that there are *two* types of modules: intentional and mechanical. This section looks at the pervasive character of this problem.

1.1 Intentional Modules vs. Mechanisms

A module is an independent cognitive subsystem which explains a competence over a well-demarcated domain. Segal (1996) classifies two types of modularity that follow in the vein of Fodor's (1983) characterization: intentional or Chomskyan modules, and computational modules. Samuels (2000) thinks there are three main types: intentional modules, computational modules, and neural modules. Intentional modules are bodies of tacit theory, like Chomsky's grammar or a folk psychological theory, which explain a cognitive capacity. Samuels calls them "systems of mental representations" (2001: 16). A computational module is a Turing-compatible machine that implements an intentional module, of which type a Fodor module is a special case. Some intentional modules will be computational, but they could also be otherwise: connectionist or dynamical systems or some other thing. Neural modules are discrete regions of the brain whose operation performs the relevant computational capacity. Of course this neural module could be the site of an intentional or computational module, and as a distinction it is orthogonal to the former two. Being a neural module has no conceptual implication for a capacity's status as a computational or intentional module.[43]

Intentional and computational modules are typically associated with different research programs in cognitive science, and different psychological domains of inquiry. Since

[43] The computational and intentional approaches do not typically connect directly to the evidence from neuroscience about various types of cognitive impairments due to physical insult or damage to the developmental program. This latter type of evidence yields information about the boundaries of regions in the brain where such functions reside, or hypothesizing about neural modules.

Chomsky, a dominant mode of explaining complex cognitive abilities at apparently higher-levels of abstractions has been by appeal to innately specified, tacitly held, subdoxastic states as the causal-explanatory mental states at issue. This has been the approach adopted in theorizing about folk physics, naïve sociology, folkbiology, mathematical ability, and a number of other domains. Theorists explain a class of behavior by appealing to what the agent knew or "cognized" about a certain domain.

Alternatively, some researchers have adopted what Marr (1982) considered an engineering approach, which has attempted to breakdown discrete tasks into detailed information flows. This latter approach treats the system like a series of mechanisms receiving narrowly characterized inputs and producing certain outputs. A red-detector or an edge-detector in the vision system is an exemplar of such a mechanism. These systems are cognitive and operate under highly precise circumstances, but they are not typically associated with "knowledge". In many cases, the complex details of how an object is perceived, from color and feature perception to its identification as a particular object, are completely unknown to the person. This approach has been adopted in theorizing about various sensory systems. One theory of rudimentary counting ability relies on an "agglomerative mechanism", which estimates quantities when presented with small numbers of items (McCloskey, 1992; Farah, 1994), and there are other such views. Another sub-discipline where mechanist accounts frequently appear is in evolutionary psychology, and "when evolutionary psychologists speak of modules, they are usually concerned with…a computational module" (Samuels 2001: 18).

A basic dynamic of the debate over folk psychology is that theory-theorists appeal to knowledge structures typical of intentional modules, and simulationists focus on a mechanism that is not supposed to be a knowledge structure. Rather, they describe discrete steps of an overall cognitive procedure from a first-person perspective, stipulating that no element of the procedure is conscious. The character of this explanation is more action-oriented, describing the strategies or tactics used to treat a piece of input. And it is emphasized that neither the outcome nor the precise nature of the process is known beforehand. This is Goldman's "process-driven" simulation. A bridge is not "knowledge", so how could it be knowledge if you had a bridge in your head?

As I have argued above, distinguishing intentional modules from mechanisms does not track a fundamental difference in the nature of the cognitive structures described. Different lines of research indeed demonstrate distinct usages of the module concept—Chomsky's usage relies on a different characterization than Marr's. But these usages have not traditionally been *contrasting* types of cognitive structures. Chomsky never tries to rule out a computational mechanism; he just emphasizes that knowledge-states must play a role. Indeed, Fodor (1983) says that most intentional modules, such as language, are likely to be implemented as computational systems. As such, they will be complex networks of physical processors, not unlike Marr's (1982) characterization of vision. Segal (1996) makes a similar point: an intentional modules can be a computational module.

A computational system or mechanism, such as a system of AND-gates or color-detectors, is a possible implementation of an intentional system. If we attribute knowledge of syntax or folk psychology to children, this is entirely consistent with saying this intentional module is implemented as a complex physical system. In the present context, it is worth strengthening this point. As Fodor has pointed out, there is no alternative to the broadly computational theory of mind on offer at all. Even connectionists fall into the broad category of information processing psychology, where mental states like knowledge are identical with states of a physical system. The debate in folk psychology is not, and in fact could not, be about the truth of the computational theory of mind. As such, it can only be the case that the theory-theorists' proposals will be implemented as mechanisms, and that the simulationist's systems will embody a logical, informational system. So we should reject the module distinction that both Segal (1996) and Samuels (2000) make between intentional and computational types. At best it talks about which level of explanation a theorist is emphasizing, not about fundamental differences in the nature of the psychological capacity.

1.2 Intentional Mechanisms

. We should consider some further objections to collapsing the distinction between mechanism and theory. The first attempts to show a reductio ad absurdum. For example, a pin prick on a person's finger travels through the nervous system in a predictable way. In fact, the signal produces an experience of pain in a manner that is entirely governed by the laws of biology and psychology. As such, that piece of the nervous system physically embodies the logical structure of the biological theory. Yet, because I feel pain on my toe does not mean I "have a theory of pain sensation"; and it especially does not mean that I

have this theory in my leg where the nerves are. Shouldn't this show that mechanism must be different from theory?

This first objection turns on ignoring the condition that only mental things can be knowledge structures. The criteria do not give a way for testing whether a bridge or a foot-nerve are mental or not. But if we know that legs and bridges do not contain anything mental, then we need not look at the physical structure they embody. We already know that they are not mental. As such, the only physical structures that will turn out to embody theories will be "in the head".

Now, some theorists may take a liberal approach in characterizing what is mental. This may be justified: it is difficult to say what is intrinsically mental without simply appealing to the same mirroring criteria already discussed. If cognition is computation, and computation is simply the physical implementation of certain logical forms, then it may turn out that silicon wafers can "have theories" or "know rules". By parity of reasoning, one might build such a knowledge system into a bridge. But again, it all depends on what we admit to be mental.

A second objection focuses on the invocation of "knowledge" in contexts where the agent seems not to know very much at all. The simulation theorists, for example, insist that the agent has no awareness of the "re-centering" their theory invokes. The agent neither knows that he is re-centering, nor does he know anything *about* how he might re-center if it was required. And once it is done, the agent does not know anything about psychology:

he only knows what came to mind, that this other person is thinking such-and-such. Similarly, theorists protest that having a vision system does not entail knowing how to discriminate edges of objects. You just *do* it, without knowing how anyone would do it or build a robot to do it. And of course, Chomsky faced numerous objections of the same variety against claims that anyone "knows" syntax.

Chomsky's response is simply that there are many ways to have knowledge-like mental states. Garden-variety knowledge is the most familiar to epistemologists. It has particular features that distinguish it: it is known consciously or can be recalled easily, it is verbalizable (as Quine 1972 required) or can be assented to, it can be reasoned with, it is subject to revision, and it is perhaps true and justified as well. Of course, if we strip off truth and justification, we end up with simple beliefs, which are still knowledge-like, intentional states. A number of theorists have argued that intentional states can be degraded in other ways, namely by making them less explicit or more isolated (Fodor, 1971; Stich, 1975). Tacit knowledge too has many grades. It might be something known but not available to explicit recall. It could be partially isolated, so that only some cognitive subsystems have access to it. It could also be implicit, so it is only implemented by a system but not explicitly represented in symbol form.

Given these options for intentional states, it is clear that the simulationist objections against "theory" are not about theory *qua* intentional states, but only against certain properties of these states. Namely, that the proposed simulation mechanisms are neither articulable by the agent, nor explicit representations of propositions. Nonetheless, this

counts as a type of intentional state: it plays a role very much like other intentional states in mental life. They have similar causal roles in producing other beliefs and producing actions. The only reason *not* to count these knowledge-like states as forms of (impoverished) knowledge, would be if they unreasonably trivialized the concept. But as we have already seen in some detail, the computational model provides a robust strategy for handling this risk.

There is simply no principled basis for denying the identity of intentional modules and cognitive mechanisms. They are different concepts, to be sure. One can imagine that there is one without there being the other. But given a computational theory of mind, where cognition is computation by physical systems, there is no such possibility. To persist is to cling to a mysterious, anti-scientific notion of the mind.

1.3 Knowledge-how

The problem of tacit knowledge has come up before. Ryle (1949) argued against "intellectualism" on the grounds that it fudged the distinction between "knowing how" and "knowing that". The view he calls intellectualism is the view that behaviors are based on procedural knowledge of the type invoked by the theory-theory. Knowledge of how acts are undertaken by agents cannot be articulated or stated in sentence-form. As such, it is not "knowledge that", or procedural, garden-variety knowledge. Instead, he argued, it is "knowledge how".

Fodor's (1971) critique of this position is essentially that Ryle's argument turns entirely on the inaccessibility of the mental states in question. Since we cannot explain how we do

many things, it must mean that "knowing how" is simply a different mode of knowing than "knowing that". But this does not mean that there is *no* explanation of how they are done. Indeed, it is surely possible that someone, somewhere can work out the detailed explanation of steps and procedures required to do things like riding a bicycle or identifying faces. Once this is known, surely knowledge is the only thing we can attribute to give explanations of structured behavior. Knowledge and related intentional states are the only type of mental stuff we have to explain structured, complex routines of behavior. It will not be emotions, feelings, or moods that explain how we ride bicycles. On Fodor's view, if the knowledge is tacit, implicit, inferentially isolated, or otherwise, it will simply be another type of knowledge (though probably *sub*-personal level knowledge).

A parallel argument applies to distinguishing mechanisms from intentional states. If a red-detector implements a sophisticated light-wavelength rule for determining precisely which waves should be called red, and if you have such a red-detector in your vision system, this does not imply that you can explain how to detect red. But there is something you know in virtue of knowing how to detect red. If you cannot articulate it, then it is tacit knowledge. But if we want to insist that it is not even tacit knowledge, then the problem remains of saying *what* causes the belief that "this is red" or *how* a person in fact detects red. And it will ultimately be a type of mental state that looks much like an intentional state.

This is the challenge for any cognitive ability, whatever the explanation. While Ryle was interested in things we "know how to do", the field is somewhat broader. We might not

say that one "knows how" to feel pain, or perhaps knows how to see red. One *can* see red or feel pain. Nonetheless, this distinction does nothing to lift the burden of explaining *how* this complex cognitive activity is carried out. At the physical level, we would describe the interaction of light-waves and chemical signals. But at this physical level, we would give the same type of explanation for deliberation, language production, and other thoroughly mental tasks. To explain this activity at the mental level, the causal structure of the input-processing relies on the same mental states we usually rely on for explaining regular decision-making routines. Something like a belief plays the role of concluding from an input to a conclusion. Of course, the state is far below full-blooded knowledge in its cognitive accessibility, epistemic condition, and so on. But it is a causally efficacious mental state with worldly content. Detecting red is like bicycle-riding. Ryle is wrong to insist on leaving it as merely as "knowing how"'; detecting red is at least partly a matter of "knowing that" certain wavelengths are "red".

1.4 Software

Another way to draw the distinction between intentional modules and mechanisms is often to appeal to a distinction between software and hardware. Some parts of the mind are hardware, like input devices or storage devices. The keyboard and the disk drive of a computer do not do any "computation", one says. They simply transmit signals to the main processor. Other parts of the mind are software, like operating systems or word processors. The software programs are complex sets of instructions and procedures, expressed in symbol form and explicitly carried out. If acts or behaviors are based on knowledge, then they can only be based on knowledge that resembles software. Gerrans

(199?) appeals to the difference between software and hardware explicitly, attempting to associate software with intentional modules and hardware with mechanisms.

Two of the key features of software make it an attractive analogy. First, one typically thinks of software as dynamic and changeable, whereas hardware is physically set out. You cannot change the size of your monitor or the positions of the keys on a keyboard by giving instructions. You can, however, modify arbitrary features of software in this way. A second feature is that software is expressed in explicit, symbolic form. A software program is a list of readable, discrete instructions. Nothing is implicit. Third software is multiply realizable, where the individual hardware is just a determinant. It asks for inputs in only informational terms. The hardware does not. Software is an abstract characteristic of a machine.

As Pylyshyn (1986) and Newell have often pointed out, the software metaphor is mistaken. At the very least, it is mistaken because the software-hardware distinction tracks no sharp divide in computer science either. A "computer" is simply an abstract machine. A "Turing machine" is a physical system which is a universal machine because it can implement the computations of any other computer. In so doing, it is identical with that other computer. At the information processing level, which is the only level that is germane to computer science, there is no way to distinguish a "native" machine and a machine that is "emulating" it. The point, simply, is that running software of particular type simply means "implementing" a machine. Personal computers are capable of implementing many different types of machines. A pocket calculator is only capable of

implementing one machine. There is no sharp division between hardware and software, only between machines.

Indeed, any software is surely a set of physical states: the software itself is implemented in memory as a set of magnetic charges. The physical states are tokens, just as a particular computer disk drive is a token of a disk-storage machine design. Some systems implement the sets of instructions required to operate the system in non-revisable, "firm" form. In such cases, the software is built into the physical device, and is not dynamic or changeable. On the other hand, the general processor at the heart of most computers is a highly dynamic and changeable piece of hardware. It can be set to implement any arbitrary "computer". Nor is software necessarily expressed in discrete symbols.

Software is equally capable of obliquely implementing desired functions. A piece of software designed to alphabetize its inputs may carry a set of instructions that are barely recognizable as a procedure for accomplishing this goal. The fact that software is typically composed of discrete elements does not mean they will correspond to the expected referents. Indeed, hardware can exhibit much of the same discrete, symbolic structure: a remote control has fixed code signals for various television channels which it transmits.

1.5 Further Resolutions

The bottom-line should be this: the distinction between intentional modules and mechanisms exists only as a matter of levels-of-description; in point of practice, mechanisms will usually represent intentional states of some form, and intentional states

will usually be implemented as mechanisms. "Usually" appears only to leave the possibility that there does exist some radically non-informational element to cognition, as yet unknown. But as far as the computational view is concerned, there is no substance to the knowledge-mechanism distinction, and there is certainly no way to make heavy weather of the distinction for debates like folk psychology.

This chapter has focused intensely on folk psychology, but there are a number of other debates where this theme appears and causes confusion.

1.5.1 Chomsky

Chomsky's account of knowledge of syntax is often considered the paradigmatic theoretical usage of intentional modules. Fodor, in particular, has made heavy weather of this usage (1998, 2000). He argues repeatedly that Chomsky's account of knowledge necessarily implies that agents possess structured, propositional mental states. He has insisted that the similarity between garden-variety knowledge and tacit knowledge is very strong indeed.

Cowie (1999) follows a different interpretation of Chomsky's nativism about language. She interprets him as postulating domain-specific mechanisms, namely, a domain-specific learning device. On her view (2000b), this is importantly distinct from there being domain-specific propositional attitudes (*knowledge* about language). For example, Chomsky appeals to the "innate human *faculté de langage*"(1965: 37, 57), the "language acquisition system" (ibid.: 53, 54), the "language-acquisition device" (ibid.: 55, 56). She also cites Chomsky (1986) to make the point that he is not wedded to intentional states so

much as he is making a case for mechanisms or devices responsible for the observed behavior: “We should...think of knowledge of language as a certain state of...some distinguishable faculty of the mind – the language faculty – with its specific properties, structure and organization, one ‘module’ of the mind.” (Chomsky, 1986: 12-13).

Cowie (2000b) intends this evidence to refute a criticism by Fodor (2000) that “what Chomsky proposes is a nativism of domain-specific propositional attitudes, not a nativism of domain-specific devices”. For Fodor’s purposes, a very particular and explicit kind of propositional attitude is required. She says, “Chomsky’s later writings make it amply clear that his is a nativism of mechanisms, and not (or not primarily) of attitudes” (2000b). Cowie may or may not show that Chomsky’s view is really about devices. However, Cowie is nowhere near showing the distinction of interest to us here: Chomsky certainly is not rejecting an intentional or knowledge-like account of the language capacity in favor of a purely mechanical, anti-intentional account. On the contrary, he is systematically equivocating on its true nature. Chomsky characterizes the language faculty as operating in accord with a certain body of “knowledge of language” which is simply implemented by some distinctive cognitive “device”.

The grounds for this dispute entirely vanish if we take Chomsky as reserving judgment on the question of whether a mechanism is an intentional module. Indeed, he seems simply to be saying “knowledge of language” is identical with a state of a physical, biological “device”. It is Cowie who is in the grips of claiming that talk of mechanisms could not possibly imply knowledge; and perhaps it is Fodor who does not think he can

get structured propositional attitudes from a language device. [44] In this case, both are mistaken in their readings of Chomsky.[45]

1.5.2 Domain-Specificity

The concept of domain-specificity is typically considered to be one of the constitutive features of a modular competence, yet it is very difficult to give an account of how to draw the borders on a domain, in particular for the hypothesized modules of evolutionary psychology. Theorists have proposed wide ranges of distinct cognitive modules that each operate on highly restricted domains (Cosmides and Tooby, 1992). Some of these modules are explicitly intentional modules, such as language (Pinker and Bloom, 1990). Others, however, are "mechanisms". If mechanisms such as counting devices or motion-detectors are considered to be "brute" mechanisms, a problem will arise for characterizing how precisely they are domain-specific.

One attractive account of domain-specificity relies on the informational properties of the domain and cognitive capacity involved. The cognitive capacity embodies a certain set of intentional states or propositions. The domain can also be characterized as a subject matter or set of related information. A capacity can be shown to be specialized to a

---

[44] In all probability, Cowie is mis-reading Fodor. Fodor actually seems to be arguing only that the *relevant* level of description for Chomsky's view is that of knowledge. How it is implemented, nobody knows. But it is at least true that it embodies a bunch of propositions of the sort Fodor requires, and those propositions are suitable to subtend his arguments about concepts.

[45] Some of these issues about interpreting Chomsky are dealt with in Chapter 2 of this dissertation.

domain in virtue of certain relevance properties of the capacity for the domain. This is only possible with intentional modules however. If we insist that mechanisms are not identical with or embodiments of intentional modules, we are faced with a much more difficult task of characterizing "domain-specificity".[46]

1.5.3 Modules

Modules are informally characterized as the independent bases for various cognitive capacities or faculties. This risks a number of trivializations and proliferations of the module concept. It is not clear, for example, whether the simulationist's account of folk psychology yields a separate "simulation" module or whether that is simply part of the practical reasoning module. Each separate function is not necessarily a module, since many complex activities might be self-contained modules deploying many sub-elements.

More formal attempts to give precise criteria for identifying a module require informational features. One promising approach is to focus exclusively on the functional characterization of the capacity as a constant transformation of a set of inputs into a particular set of outputs. Here, modularity is simply informational isolation, the inflexibility of the implemented function despite the informational states of any other cognitive capacity. But if we make this characterization available, then all capacities that are modules will be identical with tacit rules for linking certain inputs to certain outputs. Essentially, the definition of modules will rely on treating all modules as bodies of tacit knowledge, and any account that does not treat them as such will face deep difficulties.

---

[46] This issue is considered in depth in Chapter 5 of this dissertation.

Chapters 1 and 4 discuss the concept of modularity itself and its dependence on this issue in greater detail.

1.5.4 Truth-Evaluable

As Samuels (2001) points out, theories are at least in principle truth-evaluable. A theory of engineering can be evaluated for its success in accurately predicting the behavior of suspension bridges. A similar standard can be applied to intentional modules, judging their accuracy in predicting the subject matter to which they pertain or to which they are domain-specific. Some, of course, will be false theories, as is folkbiology. Mechanisms, of course, would not permit of such a treatment.

1.5.5 Domain-generality

Samuels (2001) specifically denies that an intentional, Chomskyan module entails a computational module. (This distinction is introduced in Chapter 1 and discussed in section 1.1 above.) He gives the example of a domain-specific body of knowledge being deployed on a general-purpose (domain-general) computational mechanism. The analogy to a domain-general computer is relatively straightforward: your computer can run many programs, but the code for any one program is just a domain-specific body of knowledge. This analogy is a bit too sketchy, in fact. The program on its own doesn't really do anything. It is an incomplete computer. A full-blown theory will describe the full information processing mechanism. Insofar as some theories merely hypothesize partial rules or heuristics requiring deployment on a more sophisticated system (e.g. Fodor, 1992 for folk psychology), Samuels characterization is correct. But a full-blown theory will not reduce in this way. It is itself a full computer in the way a Turing machine is a computer.

The underlying problem is discussed in detail in Chapter 5, on domain-specificity. When Samuels' is hypothesizing is not a domain-general processor "loading" various databases of domain-specific knowledge, he is simply picturing specific facts being deployed on a processor to create *one* domain-specific "machine". Beforehand, there is a category mistake in calling the body of information domain-specific, insofar as it does not contain any rules for implementation. For example, the fact "Napoleon is dead" does not merit the cognitive trait "domain-specific". Indeed, it is about only one subject matter, a triviality. But the term for cognitive science needs to be more adequately characterized, in particular to apply only to full-blown capacities. Similarly, it is a mistake to call a processor "domain-general" if it is in fact not processing anything. For example, a blank sheet of paper is not specialized for any subject matter, but that should not let us apply "domain-general" to it. A rock will be a domain-general computer too, in this case. Domain-general needs to mean that it is in fact capable of effectively processing input on several domains.[47]

Samuels, in brief, uses a confusion of domain-specificity and domain-generality to claim that computational modules are not intentional modules. Indeed, intentional modules can be either domain-general or domain-specific, and the same is true for computational modules. We should reject the metaphor of an intentional module as a piece of software running on a computational modules as a piece of hardware.

---

[47] More on this in Chapter 5.

Chapter 4. How Modularity and Innateness Connect

1. Twin Concepts: Innateness and Modularity

Cognitive science since Chomsky (1959) and Lennenberg (1964) has given a major role to the doctrine of nativism, to the point that "most cognitive scientists no longer think of nativism as a broad theoretical commitment that requires empirical justification" (Matthews, 2001: 215). This is manifest in the frequency with which theories of psychological capacities make nativist claims, and in the broad range of capacities to which this strategy is applied. But while appealing to innate cognitive endowments, theorists very often draw upon a second doctrine: the modularity of cognitive architecture. Indeed, claims for nativism and modularity occur together so often that the presence of one concept in a theory is a reliable indicator of the other's presence. Some theorists even assert that the two are intrinsically or conceptually linked (Khalidi, 2001).

1.1 Psycholinguistics

Chomsky himself can be credited with a revival of interest in both of nativism and modularity in his critique of behaviorism (1959) and following elaboration of his psycholinguistic theory (1966, 1975). His famous "Poverty of the Stimulus" arguments that children must be born with a significant endowment of tacit linguistic knowledge are at the bedrock of modern nativist thinking. Equally he has always advocated the view that this linguistic endowment appears in the shape of an independent "language organ" which "grows" in the mind (Chomsky 1980) the way any bodily organ develops from an

innately present plan. He has specifically espoused the distinctness of a linguistic module from the operation of "general cognition", but also has suggested a modular composition for linguistic ability itself: a collection of abilities implemented by syntax, phonology, and morphology modules (Chomsky 1980, 1984).

This fundamental posture broadly dominates psycholinguistics, informing the overall research program of the discipline (e.g. Pinker, 1994). Nativist, modularist hypotheses are serious contenders (or dominant views) in nearly every branch of research in the domain (Whitney, 1998). Sophisticated innate constraints have been proposed to explain a range of linguistic phenomena, where the invoked principle is typically specialized for a particular domain of application. For example, Pinker (1990) describes a "uniqueness principle" for morphology which requires that there can only be one past tense form of a verb. This innate principle aids in the acquisition of irregular forms, which require that a rule-derived construction (such as "goed") be deleted and replaced by observed irregulars (such as "went"). But the principle is restricted not only to linguistic acquisition but specifically to the acquisition of certain word-types, since objects are often named non-uniquely (e.g. "Dad" and "Mr. Sarva"). This narrowly-applicable principle has no broader function in general cognition; the unique problem it solves is very particular and so the principle is inside an innate, functional module.

Modularity and nativism support a framework of assumptions that is absolutely pervasive in psycholinguistic research: phoneme distinction very early in childhood (Gerken, 1994); Chomsky's Universal Grammar as a theory of syntax and the claim of its

independence from semantics; sequence constraints and pronunciation variations for phonemes which distinguish the space of phonologically valid constructions; early ability for speech segmentation and word identification (Juscyzk and Aslin, 1995); rules for sentence processing that disambiguate garden-path constructions without appeal to semantic meanings (Frazier, 1987); a procedure for choosing between multiple meanings for words in various contexts (Duffy et al. 1988); evidence from study of speech production and errors (Bock and Levelt, 1994); taxonomic assumptions for word learning in the face of Quinean "gavagai" problems (Fodor, 1981; Markman, 1990); and others.

Even connectionist approaches which seek to emphasize the role of learning and cross-system interaction have so far achieved success by developing system-by-system models with specially designed learning pathways from the outset (Elman et al. 1996), an approach that seems to presuppose "some kind of global modularity"; without this assumption "a free-standing face-recognition model [for example] is surely not possible" (Chater, 1994:66). Innate cognitive mechanisms operating independently of nearby processes function centrally in a broad range of explanatory projects; of course, the approach also has many critics and progress has been made on approaches that resist these dual assumptions.

1.2 Wider Cognitive Science

Theories in the wider domain of cognitive psychology also frequently deploy modularity and nativism together. Marr's (1982) theory of early vision characterizes an independent computational system which deploys a set of vision-specific rules to perform such tasks as edge-detection or depth perception (Kitcher, 1988). The approach has been

characterized as a classic modularist strategy (Garfield, 1987), but it is also nativist. The most fundamental rules for discriminating basic visual phenomena—such as horizontal lines, motion, color gradient—and for assimilating this information (e.g. constructing a 2-½-D sketch) are implemented by dedicated biological units which are considered innate *tout court*. Broadly similar approaches have been applied to face recognition, object recognition, feature extraction, as well as auditory processing, motor control and other processes (Fodor, 1983; Arbib, 1987). Developmental and evolutionary theorists have suggested a wide range of modular, innate capacities for dealing with various subject matter domains—mathematics (McCloskey, 1992; Campbell, 1994), social interactions (Baron-Cohen, 1994), theory of mind (Davies and Stone, 1995a; Carruthers and Smith, 1996), logic reasoning (Sperber, 1997), deontic reasoning (Cosmides and Tooby, 1994), human "kind" or naïve sociology (Hirschfeld, 1992; Dupré, 1983b), folks physics (Spelke, 1991; Carey and Spelke, 1994), folkbiology (Atran, 1994), religion (Pascal Boyer, 1994; Wilson, 2002), among many others. Theories in these areas often appeal to independent cognitive systems, roughly modular in nature and part of a universal human cognitive endowment.

Beyond merely invoking these concepts together, some researchers in cognitive science have suggested that these concepts *must* occur together. It has rarely been argued that the two concepts entail each other. Indeed they are completely distinct and some theorists have exploited this fact to argue precisely that modularity is true while nativism is false (Karmiloff-Smith, 1992), or that nativism is true while modularity is false (at least for theory of mind, Gopnik and Meltzoff, 1997). Nonetheless, it is very infrequently the case

that theorists will introduce models such as this, which are nativist but expressly non-modular, or vice versa. More often, theorists will explicitly state an assumption about an empirical, contingently true relation between the two concepts. Gopnik and Meltzoff (1997) take there to be a one-way relation: "while modules are innate, not all innate structures are modular" (51); in doing so they seem to be following Fodor's stipulative definition of module, where innateness is a requirement but no further claim is made about the innateness of non-modular systems (Fodor, 1983; Elman et al., 1994: 37). Samuels (2000) observes that one type of modularity requires innateness—that of "theory" modules such as a grammar module or a folk psychology module—while other types may not. Cosmides and Tooby (1994) argue that a common explanation—the brain's adaptive past—guarantees that many of the mind's capacities are both modular and innate. Botterill and Carruthers (1999) point out that while the concepts are distinct, they are "mutually supportive" (56). Khalidi (2001) discusses modularity under the label of "domain-specificity"[48]: "there is a widespread assumption in the cognitive sciences that there is an intrinsic link between the phenomena of innateness and domain specificity" (105).

---

[48] Khalidi gives a idiosyncratic account of domain-specificity, characterizing the restricted applicability of a the putatively domain-specific mechanism to a particular range of inputs, but then also requiring that the mechanism is "psychological real". In doing so, he claims that the mechanism must be "not generalizable", a concept quite similar to Fodor's "informational encapsulation" requirement. Nonetheless, Khalidi distinguishes his proposal for domain-specific mechanisms as weaker than "module", which would involve meeting all Fodor's (1983) conditions: "domain-specificity is one of the chief characteristics of modularity: all modules are domain-specific, though not all domain-specific structures are modular" (PAGE NUMBER 108?). More discussion of this in my "Chapter 1: What is Modularity?".

The pattern in cognitive science of the regular conjunction of these two recently revived, still frequently challenged, and conceptually independent concepts is a phenomenon in need of explanation. Of course, it also true that other high-level approaches—such as cognitivism, experimentalism, computationalism—also appear widely in the literature. This can be attributed to an ordinary pattern of deployment of a body of doctrine as its various pieces win acceptance. But this is not quite the case with the subjects under consideration here. They have been conjoined from the start of the modern revival (Chomsky 1975), and have historically been closely linked, as Fodor (1983) has argued about the tradition of "faculty psychology" and philosophical Rationalism. Gary Hatfield (1999) characterizes the projects of many Early Modern philosophers as aiming to characterize the cognitive mind's organization and functions. This project, in Descartes as much as in Locke, relied on a characterization of the mind as composed of independent faculties (such as the intellect, sense, and imagination in the case of Descartes, or abstraction or similarity identification in the case of Locke). It is characteristic of these philosophers that their views were nativist to varying degrees about innate knowledge or innate mechanisms of cognition (Cowie, 1999; Stich, 1975). And so, even in the most speculative stages in the development of psychological ideas, the two doctrines were confederated. The association between modularity and nativism is long-running and deeply-embedded, yet not easily explained by simply looking at the coincidental historical convergence of two distinctive lines of research.

The particular natures of the modularity and nativist concepts invoked in these diverse research areas are subject to substantial variance. Modularity ranges from an informal notion of independent function to a strongly orthodox set of properties. Nativism varies also, from a traditional Rationalist-style view about inborn knowledge to a more delicate psychological thesis about means of acquisition for fundamental mental operations. But it is my claim that this wide range of theories follows two basically constant concepts, fundamentally related throughout in virtue of certain features I will characterize in the next section. Furthermore, as I will argue following, these features have not appeared next to each other in such a range of literature and across time by mere accident; there are in fact basic reasons to believe that the truth of one implies the other.

2. What They Are

Modularity and nativism are each complicated concepts, partly because they have found themselves employed in widely varying contexts in contemporary theory-construction. For my purposes in this paper, it is necessary to give characterizations of these concepts; both to point out what aspects are continuous throughout the varied contexts and to structure their proposed interrelation. In doing so, I propose to give *minimal* accounts of each concept, not necessarily reflecting the majority or best view. It will not be possible to resolve issues that are deep disputes about the nature of the innateness concept, or issues that I have argued elsewhere are important equivocations about modularity.[49] A minimal account will amount to a characterization of the basic set of shared features that

[49] See Chapter 1 and 2 of this dissertation.

the relevant concept invokes, with some suggestion of the further (optional) elements that are frequently employed.

2.1 Nativism

Nativism is the claim that humans are born with a substantial cognitive endowment, which contains, in some form, at least some of the mental capacities we eventually demonstrate. In the contemporary debates in cognitive science, this almost always includes insisting that the capacity is *not learned*, in a strong sense, while often requiring that the latent capacity be *triggered* by normal environmental cues[50]. In Plato's *Meno*, Socrates argues that the right types of questions can reveal the knowledge already possessed by a slave boy; contemporary psycholinguistics claims that a modicum of linguistic input can activate a robust grammar-generating system (Bickerton, 1983).

As a substantive claim, contemporary nativists provide an alternative hypothesis for the acquisition of the relevant capacity: a typically biological story including evolution, genes, and other sub-psychological, biological phenomena. The essential features of this hypothesis are that the innate cognitive capacity derives from *inner*, developmentally *prior* states of the organism (Godfrey-Smith, 1994). The precise details of the account that provides this function do not generally interest the nativist theorist. Chomsky has

---

[50] Samuels (2002) introduces these two ideas as constraints on giving a theory of *what nativism is*. His Fundamental Conceptual Constraint is that something that is innate is *not learned*. His Negative Conceptual Constraint is that this does not rule out environmental triggering. The account he goes on to develop, it seems to me, fails to explain what the positive claim of the nativist is.

notoriously suggested that it might in fact be quantum mechanics rather than genetics that ultimately explains the provenance of language (Chomsky, 1980: 99-100); the point being simply that he is neither prepared nor required to commit to a particular story as long as there will be one.

Samuels (2002) has pointed out that the nativist hypothesis functions in cognitive science to claim that a cognitive capacity is *psychologically primitive*, or, not explained by other psychological operations. That is an element of the nativist's alternative hypothesis. More than this claim that the existence of an innate capacity is taken as *logically prior* to any psychological explanation, however, the dialectic requires that the existence of the capacity is indeed *temporally prior* to interaction with any properly psychological capacity. The child is born with an endowment (e.g. "cognizing" a grammar, Chomsky, 1984: Dewey Lectures again; Pinker, 1994); it is not left open that some later non-psychological process delivers it (e.g. brain lesion, Fodor's "Latin pill").[51] Indeed, it is not left open that all of the structure of the capacity come from "outside" at all (Godfrey-Smith, 1994; Stich, 1975; Khalidi, 2002). It is constitutive of nativist views in cognitive science that they appeal to the inner origin and management of the relevant properties.

---

[51] Samuels (2002) discusses this type of problem, but attempts to address is with a claim about "normal" conditions. This seems likely to lead to the traditional problems faced by "canalization" accounts. Nonetheless, I think it is important to specifically point out that this approach misses the aspect of *virtually every* nativist claim which emphasizes the antecedent, temporally prior character of the capacity as well as the inner nature of the capacities. An innate capacity is meant to be literally "inborn"—though subject to development like any other biological or other inborn condition.

Whereas historical nativism focused on innate *ideas* or knowledge, modern nativism treats both mental content—like ideas, concepts, propositions, representations, beliefs, knowledge—and cognitive mechanisms or structure as wanting explanation. A nativist can deny that there are any knowledge-like states innately, and remain a nativist by claiming there are robust mechanisms, devices, or structure. Some varieties of connectionism take this form, claiming the antecedent existence of complex network-connections and non-arbitrary start conditions, but denying that there are any knowledge-like states whatever (Rumelhart and McClelland's past tense learning model as reviewed in Whitney, 1998, or Elman et al., 1996). The simulation theory of folk psychology essentially relies on an innate configuration of reasoning systems to explain the ability to understand other minds, a proposal designed to contrast sharply with accounts based on knowledge (Carruthers & Smith, 1996).

In construing nativism to include both ideas and mechanisms, one might worry that historical Empiricists start looking a bit like nativists (Cowie, 1999); but big differences remain about just how *many* types of capacities are innate, what their range is, and how informationally rich each capacity is[52]. The weakest nativist will typically claim that

---

[52] "Informationally rich" is a bit unclear, but see Khalidi (2002). The idea is just that we wouldn't want to let a non-nativist smuggle lots of content in under a single, very general "capacity". If we draw wider borders around a capacity—saying we have an innate "linguistic capacity" rather than a grammar, a lexicon, and a phonological system—then we are trading off the count of capacities against the "richness" of the particular capacities. So an acceptable criteria for informational richness would just be one that

there are *few types* of innate structure (though perhaps there are many tokens of a type, e.g. many AND-gates are innately available for the computational system); *broad ranges* for those structures (not at all specialized to particular domains); and *informational poverty* of those structures (having simple internal structure that carries little information). This is true of the Humean associationist and of the contemporary connectionist (Fodor, 2000 . The strongest nativist, in contrast, argues that there are *many types* of innate structures with *narrow specializations* on domains where they apply *informationally rich* structure or content (relative to their environmental subject matter). Fodor's nativism about concepts is patently of this variety, as is the nativism of evolutionary psychology. One can imagine intermediate positions—a nativist who posits extremely rich, but perfectly general, innate knowledge.

So nativism is the claim that *at least some* cognitive capacities—either bodies of knowledge or mechanisms—are not learned and are internally present in some form temporally prior to interaction with the world. Stronger forms of nativism will claim that more capacities are innate, and that the innate portion itself of these capacities is more substantive. In general, this is the property being invoked by psycholinguists, evolutionary psychologists, developmental psychologists, and others who make claims about the origins of particular capacities. The present characterization clearly contrasts the nativist view with classically empiricist approaches such as behaviorism or some connectionist positions.

---

prevented a theorist with a single very complex "capacity" from looking less nativist than a theorist with many distinct capacities.

2.2 Modularity

The modularity of mind is fundamentally a claim that some of the human *mind's* distinct capacities are independent of each other. This is distinct from the neuropsychological claim that some of the *brain's* distinct regions are independent of each other, as evidenced by selective lesioning and dissociations. Rather, it is a claim about the capacities of the cognitive mind. Insofar as we think of mental operations as functionally individuated, modules will be functional units. Insofar as we further adopt a computational theory of mind, as is the broad current of most present research, modules will be computational systems. However, if instead we step back to an Early Modern view of mental powers, a module resembles what Descartes or Kant would have called "faculties" such as reason, sense, or judgment as opposed to the unified operation of a single mental instrument. So the fundamental concept is mental *independence*, which finds different criteria depending on what we think the mind is.

Present cognitive science generally considers the cognitive mind to be an information-processing device of a broadly computational nature, where mental operations are implemented as syntactically-driven calculations on physical symbols.[53] The identity

---

[53] "Classical" and connectionist architectures both implements systems with these features. The dispute is over what the physical systems represent. Classical systems involve computations over representations of the rules of the cognitive model itself, e.g. "move the N-bar phrase to the extreme left". Connectionist architectures only implicitly embody these rules, performing their explicit computations across the relative weightings of associations and inhibitions. This is a common difficulty in the debate, where theorists assume that the implications of connectionist models are stronger than they are in fact (Pylyshyn, 1984).

conditions on a computational system can be given in a range of strengths. The highest level is roughly what Marr (1982) called the *computational* level of analysis, though it is quite similar to the mode in which Chomsky (1965) began developing an account of psychological *competence*. At this level the computational system is conceived as a functional mapping of inputs onto outputs. Any two systems which produce the same mapping are identical in this sense.

The claim of complete independence is a claim of *informational isolation*; at this level it is the claim that a cognitive capacity, identified as a specific function on a range, produces the same mappings no matter how the other functions (computed in other modules) are changed. Color-name mappings can have no effect on multiplication, for example. No matter what input is given to other capacities, no matter which of its typical outputs it produces, and no matter what counterfactual input it *could* be revised to produce, this other capacity has no effect on the isolated module. Notice that this is a conceptual claim about the logical functioning of an information processing system, not a claim that a given capacity will continue to operate *come what may* (e.g. energy or nourishment is cut off to the underlying physical system).

The concept of independence permits of degrees. Modularity as informational isolation preserves this characteristic. It is a matter of "more or less", like the varying strengths of the innateness doctrine. If a function is perfectly insensitive to the external environment, it is perfectly modular. Perhaps an idealized random number generator could work like this. The system asks its random-number module for a number. The module simply spits

one out, absolutely regardless of any informational condition elsewhere in the system or environment. On the other hand, many cognitive systems will be slightly modular. An acquired response like "avoiding open flames" can be pretty insensitive to revision. But the typical person can usually override this behavior rule. This is just one freestanding rule, and it is not very fixed. As a "module", it's far less interesting than a large complex of instructions that are totally insensitive to revision; but it shares the quality of being at least somewhat informationally isolated.

A stronger requirement for equivalence between two machines than the equivalence of functional mappings is at Marr's (1982) *algorithmic* level, the level at which Pylyshyn (1984) claims we can establish *strong equivalence* between two computational systems. At this level, the actual procedure implemented by the two systems for establishing their mappings is also identical. For example, two adding machines that implement the same addition-function might deploy different algorithms and therefore fail to be strongly equivalent: one machine might rely exclusively on iterated application of the successor function, while the other machine might rely on a large table of mappings ("9+4" → "13") for some large set of inputs and only invoke the successor function for extra-tabular inputs. These procedures involve different sequences of intermediate representational states for at least some inputs.

The claim of independence at this level requires that a cognitive capacity operates by computing the same algorithms regardless of how the other functions are changed[54]. A cognitive module not only continues to provide the same outputs, but it continues to provide them in the same *way* regardless of the function of other capacities. By implication, the module accepts no information from outside itself after having accepted the inputs themselves. Given the same input, a regular sequence of intermediate states should unfold on every occasion.[55]

For example, Frazier (1987) suggests a strategy for disambiguating garden-path sentences ("While Anna dressed the baby left the room") specifically designed to isolate syntactic computations from semantic or lexical information. Specific language impairments like Broca's aphasia have been shown to disable certain syntactic functioning; yet impaired subjects continue to perform well on these types of tasks. This success is attributed to compensatory cues from semantic meanings (e.g. "The mouse was chased by the boy",

---

[54] Changing the function always changes the algorithm. If we revised addition to map 1+1 onto 3, then *whatever* algorithm we have devised will have to be changed to afford this output, even if the change is only restricted to a single cell of a lookup table. But there can be indefinitely many different algorithms that implement a single function, as in the case of the two different addition machines described above. So modularity at the algorithmic level surely requires that the algorithm be unaffected by change in other functions; but what about cases where the function stays the same and the algorithm alone is changed. Should modularity require that the module be unaffected by this type of change also? As far as I can tell, yes. The capacity is just independent of the other system, period. So what's the point of introducing all these grains of explanation for it?

[55] Suitable construed to permit probabilistic internal phenomena.

where word meaning rather than order indicates the agent). Cases of the reverse phenomenon are also available, where syntactic clues fill in for semantic knowledge. While the same answers are being given, they are being given in a different way—a key issue about the modularity of the process. The comprehension algorithms are different for sufferers of Broca's aphasia and related impairments (Whitney, 1998).

Modularity is typically a claim about a *psychologically real* feature of a cognitive capacity. Perhaps there are domains of knowledge that could be *logically unique*, such that they consist of interrelated concepts and propositions which are not related to *any other* concepts or propositions outside the logically unique set (this is a "special" body of knowledge, as discussed later). If one were to come to possess some logically unique knowledge, the logical properties of the knowledge ought not imply all on its own the modularity of that faculty. Of course, it could be the case that the physical implementation of cognitive systems mirrors the logical structure of those systems, in which case the logical uniqueness is an important clue. Nonetheless, neither logical facts nor accidents of history that may serve to isolate a mental structure are *psychological facts* (i.e. information pathways that isolate a mental structure); only when the latter obtain would we call the module psychologically real.

This characterization of modularity gives chief importance to the notion of independence, which in the context of the computational model of mind implies the *informational isolation* of the modular system. This is very similar to what Pylyshyn (1984) and Fodor (1983) identify as a primary characteristic of modular cognitive architecture:

informational encapsulation. The discussion here requires only that the module itself be unaffected by conditions outside of it, a condition which permits of degrees. The system may be perfectly isolated from whatever other systems exist, or it may be isolated only from certain other systems.

Most discussions of modularity involve a longer list of features than described here (for example, see Bates's, 1994, lengthy and near-verbatim listing of Fodor's, 1983, list). I have discussed why these features appear and their more proper place in an earlier chapter of this dissertation.[56] The basic constitutive nature of a cognitive module is its independence from other cognitive capacities, in particular the independence of its functional extension and algorithmic procedure from informational states of other capacities. The key idea is *informational isolation*, and it is in terms of this concept that we will consider modularity's relation to nativism.

2.3 They Are Distinct

One simple reason to expect modularity and nativism to co-occur would be if they were intrinsically linked in a manner such that the truth of one implied the other. The simplest way to establish this would be to argue that the concept of one contained the other. But as I have presented them, they clearly do not seem to entail each other. It could be that we

---

[56] Chapter 1 and 2 of this dissertation. An important and controversial exclusion here is *domain-specificity*. I claim that a module need not be domain-specific. Nor need it be the exclusive module that functions on this set of inputs.

have a substantial innate endowment which underwrites a single, general-purpose ability or perhaps a complex of several deeply interconnected abilities. Plato or Descartes might have had such a picture of the "intellectual", or cognitive, powers: a body of innate ideas and a natural mental instrument which is itself essentially unstructured and *a fortiori* not modular. The opposite view is also possible. Some empiricist views suggest that all task specific skill sets are learned, but that the resultant bodies of expertise are likely to be quite independent: chess mastery and piano virtuosity would involve sets of non-overlapping conditioned responses. Karmiloff-Smith (1992) denies that the modular structure of adult cognition is due simply to its innateness; rather, the structure of independent functional units develops through complex interaction with the environment.

One might try to further develop a purely conceptual connection between these faculties despite the apparent counter-examples to the necessity of their conjunction. For example, the basic nativist position is that at least some of our capacities exist in some form at birth. This endowment may be either *fully-formed* or *incomplete* in a way requiring "activation" before it is expressed. If fully-formed, then the knowledge or mechanism is ready to perform as a mature capacity, taking the ordinary range of inputs and producing the relevant outputs. For example, assume that the ability to discriminate phonemes is one such cognitive capacity, as demonstrated by the observation of this ability in very young children. Purely physical information is received by sensory transducers; certain discriminating criteria are applied; and outputs are produced that categorize the perceived sounds into phonemic units. Since this ability is already fully-formed, no type of input that it receives will be treated in any fashion other than as an auditory input; i.e. there are

no special "triggering" inputs for which alternative procedures are invoked. There is no plasticity built into the function which gets triggered by particular types of inputs. Since this depends on what responses the function maps onto given inputs, its fixity can be read off a description of the function itself. As such, the capacity's sequence of procedures is fixed from the outset. And indeed, this seems to imply precisely that the capacity is also modular. Since the adult capacity to discriminate phonemes is wholly available from birth, it implements the same set of mappings from inputs to outputs.

But this is not a very good argument for the modularity of the "fully-formed" capacity. Though by stipulation the capacity does not undergo development—since it exists in essentially its adult state from the outset—this does not mean that it *cannot* be trained or otherwise influenced by experience or learning during development. And if it can be so influenced, then surely it is not modular in the informationally isolated sense we have assumed. Folkbiology, for example, is a proposal for an innate, naïve theory of natural organisms. But even if this body of theory were wholly innate, nothing should prevent the adult from incorporating new information as it became available and altering the way future judgments were made. In this manner, the extensional mapping provided by the particular capacity would be altered. While one can conceive of innate endowments that are modular, it is not entailed by a capacity's innateness that it therefore be isolated from change.[57]

---

[57] Keep in mind that modular means informationally isolated. A common usage in psychology literature for modular is to mean "insulated from top down influences", in the way the vision system falls for optical illusions regardless of high-level awareness of the illusion. To be insulated from top down information

Perhaps, however, an argument for fully-formed capacities can be suggested that takes incomplete capacities as the starting point. An incomplete capacity requires some environmental cue or even learning before it can operate properly, but also necessarily requires the contribution of some already-present, unlearned component that is innate in the sense presented.[58] One proposal for such a capacity is Fodor's theory of concepts as unique, inert and innate items which are triggered into referring by a brute-causal process not unlike imprinting (Fodor, 1998; Cowie, 1999). Until the concept refers to something, however, it has no function. As such, it is simply unavailable to other mental processes—learning, other types of psychological acquisition, or any other mental process. Therefore, by being innate, this capacity is *fixed*. And this may be the sense in which even fully-formed capacities are modular in virtue of their innateness. While learned knowledge may alter the judgments produced by one's "theory of biology", it might be that the innate elements simply remain, trumped by new and contradictory information. Therefore, the innate capacity of folkbiology may simply be an incomplete biological faculty. On this characterization, a capacity is innate when it is *fixed* against the operation of any psychological operations—present, temporally prior, and so independent of their

---

flows *just is* to be informationally isolated in a particular way. But it is possible to be *even more* isolated than that, such that even bottom up information will not re-train the module to behave differently. See Chapters 1 and 2 for more complete discussions.

[58] In a very weak sense, every cognitive capacity requires *some* innate contribution to obtain, even capacities that are almost entirely learned. Here, however, consider some of the stronger conditions suggested earlier: that the innate though incomplete component is informationally rich (relative to the experiential component) and specialized.

operation. If this is so, then that is *ipso facto* a variety of the information isolation characteristic of modularity.

This argument also fails to go through, however, and for an interesting reason. While some proposals for nativism also require that this endowment is *fixed*, this particular aspect could easily be false while the endowment continued to be temporally prior. The Ur-concepts of Fodor's model could simply exist in "junk" form until activation, therefore susceptible to deletion or manipulation by general cognitive processes. Similarly, folk biological knowledge might develop by the deletion and replacement of the innately endowed concepts (like "essentialism") with scientific ones ("population thinking").

Neither Atran's (1994) folkbiology proposal, work on early phoneme detection, nor Fodor's (1998) concepts proposal are in fact expressly proposed as "revisable" bodies of innate endowment, so they do each imply a kind of modularity. Indeed, the overall aim of this paper is to explain why it should be that innateness claims link into modularist claims. Here we see one aspect of a connection I propose to elaborate later: nativist theories sometimes require the fixity of that endowment. But for now I think it is sufficient to highlight that innate endowments need not be fixed, and so need not be at all modular.

We just considered some ways that nativism might imply modularity; but now let us consider the inverse relation. Does modularity imply nativism? Modularity is the claim

that a cognitive capacity is informationally isolated, yet it does not imply that there is absolutely no exchange of information between a particular capacity and others. Rather, the claim discriminates between two classes of information. On the one hand, there are the inputs and outputs, which are two definite sets of informational cues and responses. On the other hand are the procedures or proprietary data structures which interact with the inputs to produce outputs, but do not themselves change. These are the algorithms that product mappings of inputs to outputs. The algorithms themselves are *fixed*, even as the system moves from state to state during its computations. It is in virtue of this fixity that the definition of modularity obtains: that the function instantiated is unchanged regardless of the informational states of any external cognitive capacity. But if no other psychological capacity can change the core function of the module, then it cannot be operated on by learning mechanisms. Nor in fact could any state obtain such that the module "changed itself", because the immediately proximate cause of that state would then be implicated in causing the module to change its core function. So to be a module is to be unchangeable by any psychological capacity.

Let's consider one particular version of nativism. On Samuels's (2002) account of nativism, the definition of innateness is psychological primitivism (such as that of a module unchangeable by any psychological mechanism) and nativists believe that at least some psychological mechanisms are innate. On this view, to be innate is to be unchangeable by psychological means, and therefore to meet our definition of modular. But that is an idiosyncratic view of nativism.

Let's consider whether modularity implies nativism on my advocated treatment of nativism from section 2.1 (i.e. not learned and antecedently present). On first appearance, it looks like modularity does imply nativism: if learning cannot alter the capacity, then surely it cannot have been acquired through learning (or other psychological modes of acquisition). For if it had been acquired that way, there would have had to be a stage at which the incomplete module was susceptible to outside information input to its core capacities. But if the module were *ever* open to inputs, then it could only have been closed off by either a non-psychological event or a psychological event. If the closing-off event is non-psychological, then that event at least is innately programmed and therefore modularity implies some sort of nativism. Or else if the closing-off event is psychological, i.e. some information input, then there must already be some non-psychological mechanism for implementing this psychological "command" and barring the effect of future inputs. Otherwise, a future input could simply "override" the closing-off event. And if there is some non-psychological implementation mechanism, then again the modularity of the system implies the existence of some innate non-psychological element.

This argument from modularity to nativism seems a bit suspicious, unfortunately. In particular, it is probably false that *only* a non-psychological mechanism can implement an irreversible command to "become modular", i.e. ignore further instructions to change the core function. It seems plausible that an open system can operate as isolated if it follows a rule like "Ignore all new rules."

Indeed, relying on modularity's implication of psychological primitivism invites the other problem suffered by Samuels's (2002) account of nativism. There are ways to be psychologically primitive that do not involve innateness. Brain lesions or other trauma produce conditions that are primitive yet non-innate[59]. So do the fundamental perceptual operations. The very first stages of perception are non-psychological: light energy hitting transducers, for example. At some step after that, the processing becomes psychological. But it will always be that the immediately previous step is a psychological primitive—there is no psychological explanation for the origin of that particular stimulus signal, the explanation is physical or biological. So primitivism is not identical with nativism. And so we should expect that linking modularity to primitivism will not on its own suffice to imply nativism.

2.4 How They Connect

So what is the connection? It is not that the concepts as used in cognitive science have any conceptual implication for each other. Rather, I will argue that the same evidentiary bases upon which theorists typically build their arguments for nativism also imply the existence of modularity, and vice versa. When we have reason to believe that a capacity is innate, the same reason often suggests modularity. And the hallmarks of modularity also function as arguments for nativism. In the following sections I will consider the various types of modularist and nativist arguments; I will suggest that they are each

---

[59] He treats the problem of lesions by invoking a concept of normal conditions. I suspect this will lead, when pressed, to something not unlike the canalization account. This would then make the primitivism at the core redundant.

stronger than required, implying more than the single concept they are meant to support. Modularity arguments imply the truth of nativism, and vice versa.

If this is right, there is an epistemic or evidentiary link between these two features of cognitive capacities. The way we come to know that a capacity has one property also suggests the other property, even though the two properties are in principle distinct. So while there is a vaguely transcendental feel to the structure of the argument, it does not imply nor rely on any intrinsic connection between the two ideas. Nor does it have any implication for the empirical truth of hypotheses employing the two concepts. It could be that the two properties are linked for important empirical reasons, or it could be empirically false that the two ideas are in fact linked (many, as yet undiscovered, modular but non-innate capacities, for example).

The conclusion of this argument will be asymmetric. Nativist lines of argument are *suggestive* of modularity. They require a premise that *implies* the truth of *modularity-so-far* (not full blown modularity). Modularity-so-far is the idea that a capacity has so far behaved as if it were modular, though it may not in fact be restricted from non-modular behavior. Modularity-so-far is itself only suggestive of full modularity for the particular capacity under consideration.

In the other direction, modularity arguments do *imply* nativism, but only a very weak variety. For there to be multiple modules, there have to be multiple "executive" authorities, or separated pools of cognitive resources, present in the mind innately. This is

very far from the robust nativism that is at stake in the debates. But it is a very weak type of endogenously specified cognitive structure—and innate endowment. The weak nativism isn't at all suggestive of the stronger variety. So just because there is some very generic innate structure that we begin with—a *tabula rasa*, say—it does not deflate the genuine debate between empiricists and nativists. That debate is about just *how much* is innate, and *how specific* it is.

The claim that there is an epistemic link suffices to explain the phenomenon identified at the outset: that in the literature thus far there appears to be a strong correlation between modularity and nativism. Theorists often employ both doctrines because of they way they come to each idea, not because they discover some connection.

3. Anti-Empirical Disclaimer

The empirical truth of the arguments to be considered is much in dispute. But this paper is not concerned with their truth, so much as their extra-empirical connections. It would be one project to show that nativism and modularity are *in fact* true simultaneously and only together of such-and-such mental capacities. That is not the project here. I have argued against another project: of showing that the two concepts are intertwined logically. Rather, the project here is to demonstrate that the grounds for arguing one claims typically link into the grounds for another—whether or not those lines of argument are ultimately vindicated.

4. Arguments for Nativism

There are three major types of arguments for nativism: Poverty of the Stimulus, Development of Fixed Capacities, and Impossibility of Learning.[60] Most of the concrete arguments presented in the present and historical literature fall into one of these categories. In the following sections, I propose to present the outlines of these arguments and suggest that they imply not only nativism but also modularity.

In general, the strategy of nativist arguments is indirect. There is not yet any direct evidence that knowledge or structures exists innately: we are not able to read anything directly off an infant's brain. Instead, nativist's rely on what must be the case given the observed facts. Poverty of the Stimulus arguments point out that a particular piece of *knowledge* that can be found in the adult cannot be found in the learning environment. Arguments from the Development of Fixed Capacities emphasize that some *feature* of the adult capacity is not present in the learning environment, a related but different claim. Finally, arguments that claim that particular learning or all learning is impossible focus not on the environment but on the means of acquisition.

4.1 Poverty of the Stimulus

The general structure of the Poverty of the Stimulus argument is to deny that the environment contains something that the cognitive capacity demonstrably contains. This

---

[60] No doubt there are others. In Chapter 2 I argue that these three represent the three major lines of argument that have been offered since antiquity for nativist hypotheses, from Plato to Descartes to Chomsky.

can be achieved by claiming that the world itself lacks some such property, as Descartes did for the property of "perfection". This is an argument for the poverty of the *environment*. The same end can be achieved by arguing that the proximal environment of the subject has not in fact contained this property (though the world in principle could contain it). Plato, in the *Meno*, uses such a poverty of *instruction* argument which localizes his claim to the specific experience of the slave boy. Even if knowledge of geometry exists and is widely discussed, the fact that the slave boy has not been exposed to these stimuli is the relevant fact.

Chomsky's claim for the innateness of grammar rests partly on a claim of the latter type. Natural languages have a particular grammar. Children come to speak languages with this grammar even though they are not exposed to many elements of its elements. The most dramatic cases of this type involve creolizations or spontaneous sign-language construction (Bickerton, 1983). These speakers are not even exposed to grammatical languages, yet they begin to use languages with the fundamental grammar of natural language. But since they could not have learned the grammar without having been exposed to it, it must be innate.

This argument leaves open the possibility that the knowledge is neither learned nor innate but *deduced*, a case such as one might argue applies to a fact like "15+23=38". If the fact is evident to reason, for example, then one might argue that it is the product of rational thinking but not contained within reason itself. Nativists typically argue either that something cannot be the product without having been contained (in the case of historical

nativists), or that the knowledge (like grammar) is so arbitrary so as not to be the unique product of reason—so must be either learned or known innately. This latter branch is an argument sometimes summarized as "language is special" (Whitney, 1998), and often a suppressed premise in this type of argument. After ruling out that language could be learned, arguing that it is "special" rules out that it could be *deduced*.

In practice, Poverty of the Stimulus arguments always conclude that some particular knowledge or mental content is innate, rather than that some mechanism is innate. Chomsky posits innate knowledge of a grammar, Plato posits knowledge of geometry, and Descartes posits the idea of God. One reason for this is that Poverty of the Stimulus arguments are usually deployed on a subject matter that is plausibly learnable, a piece of knowledge or mental content for which the empiricist has often suggested a learning-based account. Rather than refuting the learnability of the knowledge, the nativist seeks to show that there was no opportunity to learn it.[61] This is the dialectical context that nativists have historically found themselves facing. The empiricists rarely argue that some "skill", like color detection or pain sensation, is learnable; the contest is always starts over a piece of knowledge.

---

[61] How precisely to discriminate a case where the theory posits a piece of knowledge (such as knowledge of a grammar) vs. where it posits an ability or mechanism (such as the ability to discriminate between phonemes)? I think this is in fact a problem. See Chapter 2. It may not be a tenable distinction. Also see the end of the next section, re: Fixed Capacities.

4.2 Development of Fixed Capacities

The argument from the Development of Fixed Capacities observes some rigid characteristic of a mental capacity and claims this very rigidity as evidence for innateness. There are a number of different specific argument that can be categorized under this rubric. Lennenberg (1964) argues from two types of universality for the innateness of language. On the one hand, the universal possession of a piece of knowledge by all members of a species at a particular time implies innateness. On the other hand, the possession of a piece of knowledge throughout the (cultural) history of any particular group also implies innateness. Both are varieties of *species-typicality*, a feature invoked by Chomsky (sometimes simultaneously with the Poverty of the Stimulus argument)[62] and also by evolutionary psychology (Pinker, 1994). In each case, the argument appeals to a *feature* of the knowledge—its universality across a certain dimension.

The rigidity of the capacity is the first part of the argument. The second part is usually implicit. It is simply assumed that this feature could not have come from the

---

[62] Chomsky does exactly that here, running together the Poverty of the Stimulus about knowledge and the Development of Fixed Capacities arguments into one, "It is clear that the language each person acquires is a rich and complex construction hopelessly underdetermined by the fragmentary evidence available. Nevertheless, individuals in a speech community have developed essentially the same language. This fact can be explained only on the assumption that these individuals employ highly restrictive principles that guide the construction of grammar" (Chomsky 1975: 10-11).

environment, that this feature is "special". In fact, this element corresponds to the role of the "language is special" premise above. So: we observe that all humans have this capacity and at all times, and because we know that human experience has varied dramatically across these dimensions, we conclude that the uniformity is unlikely *unless* the feature is innate. The debate over past-tense formation for English verbs uses essentially this argument (Pinker, 2000; Whitney, 1998). Without this suppressed premise, the argument would not go through. It is because adults' language users never vary their past-tense irregular verb forms that the rigidly-timed stages evident in children's development call out for explanation. By contrast, we cannot conclude that because all humans do and have believed that "water is wet" therefore "water is wet" is an innate belief. Rather, the uniformity of the belief exactly matches the uniformity of the *veracity* of the fact that water is wet.

Universality of a particular capacity across a range is one argument of this type. A second major type draws on the regular schedule of development of particular capacities. If a capacity develops on a regular timeline—as has been observed for language, arithmetic, theory of mind and others—and the environment does not contain any sequence of stimuli following this particular timeline, then we can conclude that the development is "driven" by some sort of internal "bioprogram" (Bickerton, 1981) or "unfolding". The timeline need not be species-typical in this case—it must only fail to correlate with any external driver.

A third major type of Fixed Capacities argument relies on dissociations. Various types of damage or disorders can impair particular capacities in isolation from others. This suggests an articulated (or, indeed, modular) structure to the cognitive bases for the ability. Yet the world and experience do not have this type of structure—linguistic, social, and arithmetical facts are undifferentiated from each other. So it must be that the individual elements or at least their architecture are innate. This type of argument is already clearly an argument for the innateness of modularity.

The class of Fixed Capacities arguments is very similar to Poverty of the Stimulus arguments. Rather than identifying a piece of knowledge or information that the subject has and the environment lacks, these arguments identify a non-contentful feature of the capacity: universality, developmental timeline, or architectural isolation. Indeed, it would not be unreasonable to characterize the entire group as Poverty of the Stimulus arguments. Yet, while the former focuses on content, the latter is almost always taken to imply a mechanism or structure. The rigid developmental timeline is not taken as an indication that the mind "knows" something; rather it implies a program or developmental mechanism. So the former class establishes poverty of knowledge or content, and concludes that such content is innate. The latter identifies a feature or property of that knowledge or capacity, and concludes that a mechanism or structure innately implements this feature or property.

4.3 Impossibility of Learning

Impossibility arguments attack the learnability of a specific or general type of knowledge by a learner. Leibniz, in the *New Essay*, argues that the materiality of sensible experience

cannot possibly interact with the immaterial mind. Therefore, learning anything at all from experience is not possible. Any knowledge we have must exist from the outset. It is also possible to admit that some learning is possible but rule out that knowledge of a particular domain can be acquired by learning. Gold (1967) is taken to show that a person cannot learn a language from an arbitrary finite sample of the language without either constraints on the possible languages or negative feedback. If there is indeed no negative feedback provided to children, then it must be that they start with at least some constraints. Evolutionary psychologists have argued not that learning of certain problem-solving strategies is impossible, but that it is not practical given the survival risk posed by certain types of evolutionary problems; therefore a general purpose learner would be a poor adaptation to certain evolutionary problems (Cosmides and Tooby, 1994; Pinker, 1994).

An obvious feature of this type of argument is that it must characterize what exactly is unlearnable. Leibniz's metaphysical argument separates knowledge into a category of immaterial stuff. But the more specific arguments must also give criteria to circumscribe what precisely they believe is unlearnable, enumerating evolutionary problems about: how to recognize your friends, avoid predators and so on. Furthermore, this knowledge must be characterized with respect to the conditions under which it is unlearnable. Presumably one could learn the idea of God from God or from the direct operation of another mind; Leibniz is ruling out learning from the basis of sensible experience regardless of your mental condition. Gold rules out learning from linguistic data as long as you have *no linguistic knowledge* about the target language, regardless of other general

knowledge you may have. And finally, evolutionary psychologists argue, for example, that predator avoidance cannot be learned during the first predation encounter. So Impossibility arguments import assumptions about the condition of the learner, and argue from there that he cannot have learned.

Impossibility arguments are generally addressed to knowledge-like states exhibited by a cognitive capacity, like the ones just considered. One can also construct arguments that particular features of mental capacities could not be acquired, and must therefore be innate. For example, all infants display preferential attention towards certain stimuli. While the preferential attention itself suggests innateness, because of Poverty of the Stimulus, the uniformity of the preference as observed across many children also suggests a common innate basis.

4.4 Implications for Modularity

Several features of nativist arguments in general and of some in particular reveal a close relationship to claims about modularity. In general, if you have a nativist argument of the type discussed so far, you also have the beginnings of evidence for a module. The first step is to look at the proximate consequences of the arguments presented.

4.4.1 The Minimum Hypothesis

Cowie (1999) argues that the Poverty of the Stimulus argument is intended to open a "gap" between what is known and what can be learned. This same gap is a consequence of the argument from Fixed Capacities as well as the Impossibility argument. In each case, for a particular class of knowledge, the nativist shows that the person has *more*

knowledge on some subject matter than can be explained through learning, where learning is any psychological means of acquisition from the environment.

With the exception of Leibniz's metaphysical argument[63], each nativist argument is structured to demonstrate a gap of this type for a *particular* domain of knowledge. Chomsky and others show various gaps in linguistic knowledge by scouring the environment for a particular body of information demonstrably present in the language user. Lennenberg's arguments extrapolate from rigid characteristics of linguistic knowledge as it is found in humans through time and across cultures. Gold's theorem concerns the learnability of a language grammar from a finite set of linguistic information. As Cowie points out, nearly every nativist will concede that "the empiricist succeeds in explaining some acquisition phenomena" (1999: 39); the challenge is not to show that all knowledge is innate. Rather, the nativist invariably argues that one of the three core arguments obtains *with respect to a certain domain*.

The conclusion of each nativist argument is, therefore, a specific gap between what is known and what could have been learned. From this position, then, the nativist must propose an internal condition to close the gap. The *minimum* proposal is typically an endowment that specifically addresses the missing information, what Cowie calls the

---

[63] Strange though it is, Leibniz's argument has no correlate in the modern milieu. Not even Jerry Fodor, who has occasionally expressed skepticism that learning can be at all explained, believes what Leibniz was arguing for. Everyone today will agree that some things are learned. As such, the partialist nativism I am describing is *a propos*.

"special faculties hypothesis". So if a linguist can show that children develop languages with phrasal structure even when they are not exposed to any languages with phrasal structure, then we should conclude that there is a body of special-purpose knowledge to subserve phrasal constructions. We have, in fact, an argument that among the mind's many capacities, one in particular is inborn and regular, no matter how the rest turn out to be.

But perhaps the minimum hypothesis is not so parsimonious as it hopes. By making the minimum claim, the nativist might in fact be cutting off another possibility: that there is in fact an innate *general* capacity which explains the appearance of this specific type of mental capacity. Cowie (1999) concerns herself to emphasize the possibility that this is the case. While it might be true that Chomsky's Poverty arguments show that *something* innate underwrites the acquisition of language, why should it be *knowledge of language*? It seems plausible, at least *a priori*, that a Poverty argument might allow some domain-general knowledge or mechanism—like a preference for simpler hypotheses, or rationality—to explain the appearance of the relevant phenomenon. Indeed, one might add to Cowie's position that the nativist is diminishing the explanatory breadth of his explanation in the name of parsimony, only to increase the strength of the conjecture by claiming a specialization of function for the resultant faculty. Perhaps it is more parsimonious to claim a domain-general innate principle, or at least to leave both options open by making a disjunctive claim.

This challenge focuses attention on less prominent features of the nativist argument. The Poverty argument, for example, begins by showing that experience is not alone sufficient to acquire all the features of language. But couldn't language be like arithmetic, like "2+2=4"? If humans have reason and reason has some laws which compel one to believe "2+2=4", one need not claim that this fact is either innate *or* learned from experience. It comes to be known from the operation of an innate, but non-linguistic or non-domain-specific, faculty. This is nearly Cowie's suggestion: we might have a principle like "Prefer simpler hypotheses" which is a domain-general bias or belief that *also* suffices to underwrite language acquisition. But for the Poverty argument to get started at all, it must already establish that the subject matter under consideration is "special". Special means precisely that the domain is unique and does not share basic principles with other domains. Therefore knowledge of other subjects does not suffice to learn about this one.

By arguing that the environment lacks the appropriate stimuli, the theorist establishes that there are certain specific things that need to be known. But furthermore, these things could not simply have been deduced from reason or some other general faculty—these facts are special. In practice, this means that they are arbitrary and historical in nature, not structured upon more fundamental or general principles of a sort that might be part of a general capacity. So there is no rhyme or reason to our particular linguistic grammar as compared to other possible linguistic grammars. If anything, as evolutionary psychologists have argued, we are likely to find only *adaptive* explanations for innate states, but not explanations that are rational from the individual perspective. By contrast, it is difficult to argue that the fact "13+32=45" is known innately without denying that

there is a central plenary capacity for reason, i.e. by specifically impoverishing our picture of the general capacities to make room for a "special" subsidiary capacity such as arithmetic (see McCloskey).

But does parsimony itself similarly militate in favor of the minimum hypothesis? When there is evidence for Poverty of the Stimulus with respect to a particular domain, the minimum knowledge enrichment to close the gap (between what is known and what could have been learned) need not be domain-specific. If indeed "prefer simpler hypotheses" could push a learner toward the right grammar, surely this would be a more minimal endowment than positing innate knowledge of, say, Chomsky's government and binding theory—a large and sophisticated body of propositions. One criticism of "prefer simpler hypotheses", no doubt, is that it only pretends to simplicity. The formulation is hopelessly vague, and any spelling out of a methodology would surely be highly complex (Matthews, 2001). But let us leave aside that objection to consider another aspect.

The type of approach under consideration keeps score at the wrong level. Counting the propositions to be pre-packaged with the innate endowment is unlikely to yield any sensible accounting: is one long conjunction simpler than three separate propositions? Should the consequences of a proposition be counted as part of the endowment? Parsimony on that dimension—of simplicity or minimality of cognitive endowment—is neither practical nor relevant. The relevant minimality, I would suggest, is of the functional strength of the capacity. That, after all, is the level at which we are conducting our functionalist theory of mind. The parsimony of the endowment should be measured

by the range of capacities explained. And in this respect, the minimal hypothesis is more parsimonious. The results of psycholinguistic research are taken to apply first to linguistics; and insofar as the researcher is ignorant about the projects of vision or social reasoning, the researcher's conclusion ought not affect them. The minimum hypothesis is a conjecture that explains linguistic capacity and nothing more. The "prefer simpler" hypothesis is, on the other hand, very strong. We could use it to help deduce all kinds of results in a wide range of domains; the minimum hypothesis is not nearly as handy. It is more parsimonious to posit a very narrow faculty, then a general principal of reasoning.

Indeed, recall that the Poverty argument typically begins after already conceding to the empiricist that some substantial portion of adult cognitive capacity is the product of experience and learning. Just how much is learned is unclear, but learning seems to happen for a very wide range of domains. If the nativist's claim is to remain consistent with that concession, whatever the nativist proposes ought not encroach on that range of knowledge on which the empiricist's learning mechanisms operate. A domain-specific innate endowment is best suited to avoid such conflict: the child knows about grammar, say, but nothing in that endowment impinges on the fact that "much else" is learned. If, on the other hand, the nativist were to propose a general knowledge principle, like transitivity or Occam's Razor, the range of learning domains affected would be immense.

So the Poverty argument has special features that resist Cowie's objections about parsimony and domain-generality. The other argument that is insulated against this attack is Impossibility. If it can be shown that some body of knowledge is unlearnable unless the

subject already has that very knowledge, then the inference to a domain-specific store of knowledge obtains. Gold's (1967) theorem is taken to have precisely this result (Matthews, 2001). Cosmides and Tooby (1994) make an argument of this structure as well. They argue that *any* domain-general learning on the present models requires trial-and-error patterns. Yet many adaptive problems tolerate zero errors, such as the problem of predator avoidance.[64] Therefore, those tasks must have innate, domain-specific solutions. Of course, one simple conclusion of unlearnability is that some body of knowledge is innate. Given the possibility that specific *or* general learning procedures could explain the acquisition of this body, one should opt for specific faculties simply for reasons of parsimony (by parity of argument to the case of Poverty).

For the final type of argument, Fixed Capacities, the inference to domain-specificity need not obtain. Fixed Capacities does not rely on the relevant capacity, such as language, being special with respect to other mental capacities. Lennenberg's observations of species-typicality could just as easily be explained by the possession of a single, domain-general facility such as reason. So the Fixed Capacities argument should not be taken to imply the minimum hypothesis.

---

[64] Is the criterion too strong? For example, if I "know" beforehand something unknowable, i.e. tomorrow's lottery numbers, then should we conclude I knew it innately? This seems the wrong way out of a Gettier puzzle, even after attenuating the sense of "know". If I have some correct belief about the unpredictable future, surely it's just a guess. So perhaps unlearnability arguments leave open two conclusions: either you know it already, or it's a guess. Though of course, the linguist goes on to supply a plausible alternative route by which the knowledge could have been obtained without guessing (evolution, genes, etc.).

4.4.2 The Minimum Hypothesis is Modular

Now that we have considered the route to the minimum hypothesis, we can finally observe that asserting the minimum hypothesis for a given cognitive capacity is also to assert that it is modular. That is, the psychological structure hypothesized to be innate is an independent, informationally isolated capacity. If it were *not* this type of capacity, then it could not have been the subject of a successful Poverty or Impossibility argument.

Consider a toy example: chess-playing ability. No doubt any hypothesized chess-playing capacity would draw heavily on *input* information from many sources: vision for seeing the board, but perhaps also empirical information about historical situations and strategies. But if chess-playing ability is to be innate, then there needs to be some underlying capacity that exists independently of such patently empirical inputs. Let us say that this is some set of decision-making rules such as R, "Always Protect the Queen," and their underlying concepts, which together make for a cognitive endowment C. We could characterize C as a set of instructions for computing recommended moves based on game situations, or a function from situations to moves. Now also assume that C is in fact non-modular, and therefore subject to change in its fundamental rule set. This is not implausible, as there might be some tactic that requires one to "Risk the Queen". In this case, when the player learns "Risk the Queen", we have fundamentally modified the basic innate endowment.

Can one run a Poverty argument on a capacity like this one? Grant for a moment that experience in fact lacks the richness of information from which one can actually learn all

the elements of chess; perhaps one can learn the rules of piece-movement from training, but one cannot learn "tactical shrewdness", for example. And tactical shrewdness is what R codes for and makes up the core of C. An important part of establishing the innateness of a *chess* playing ability on this basis, however, is establishing that chess is *special*. While chess skill does not arise from mere experience alone, it *also* does not arise from experience *in conjunction with* some general capacity like reason. The Poverty argument is meant to show that chess-playing skill in particular is based on a store of innate knowledge, and not on general principles—that chess is special. After all, if chess skill is a direct consequence of general powers everyone already agrees we have, then there is no challenge for the nativist to tackle[65]. But if chess skill arises from a general power that nobody yet agrees we have, then it is unparsimonious to conjecture something stronger than required. It is unparsimonious to conjecture something smart enough to be good at chess without the relevant training but domain-general, i.e. a faculty that's innately good at *everything*.[66]

---

[65] Just because the nativist has no challenge to tackle, it does not mean that the capacity is therefore learned. In fact, if the chess playing capacity derives from more general cognitive faculties—such as reason—then it *is in fact innate*. But there will not be any nativist argument for chess-playing. The argument will only be for reason itself being innate. *Methodologically*, we should only expect to see arguments that cover faculties that are specialized to some complete domain.

[66] It certainly seems odd to posit a chess-playing faculty. It seems to specialized. But the contrast should be to a general purpose faculty that is *good at everything*. When we think of a general faculty, we more often think of one that is general but weak. It needs lots of experience to learn things and deduce expertise. But the nativist has presented evidence that there *isn't enough experience to learn how to play chess*. So a weak general faculty could not do the job. Only a general faculty that already knew how to do *everything* or *lots*

Consider what happens when we allow the core rule R to be revised by a tactic like "Risk the Queen" or even an alternative strategy like A, "Play Fast and Loose with the Queen". R is an element of the unlearnable and un-deducible body of knowledge C. But A clearly is either learnable or deducible from some external source; it was not part of the original innate store. When it replaces R, the resultant C is no long entirely *special*—at least part of it is deduced or learned. In the limit case, it could be that all elements are replaced with derived elements like A.

Since R is part of the innate endowment C, there is some time T^1 when the subject has R, later replaced at T^2. Now for the problem: when the theorist studies this subject, how can she show that R is innate? At time T^2, the subject has either learned or deduced some contrary fact such as A and replaced R. More generally, for C which underwrites the ability to play chess, it is only at T^1 that an argument can be mounted for its innateness. At T^2, only C-R remains to be argued for. In the limit case, C could be completely vanished at T^2, and there is no basis for a Poverty argument left. For, after external information has intervened, it is no longer possible to show that the knowledge is neither learned nor derived.

This does not imply, however, that C is in fact informationally isolated. It only implies that if we can make a Poverty argument for it, it has not *as yet* been modified by external

---

of things without learning them first would be plausible. Yet *this* is surely less parsimonious than a narrower module.

information, a *de facto* modularity. So the capacity may *behave* as if it were modular without the modularity being psychologically real – it's malleable but not changed yet. The successful application of a Poverty argument shows that some mental structure exists innately and that this structure has not been externally modified during development. Indeed, it is a consequence that the relevant mental structure has *exhibited* modular functioning to the extent that the capacity is unchanged. So it can only be *modular* or *modular-so-far* mental structures to which we can apply Poverty of the Stimulus arguments.

A closely related argument establishes the same situation from Impossibility arguments. There is no basis for an Impossibility argument at T^2, when some new information has replaced the innate endowment. So if some knowledge is not learnable, it has been observed at a stage like T^1.[67]

Nativists that take a path through the Minimum Hypothesis with Poverty of the Stimulus or Impossibility arguments, as many do, are very likely to end up with an implication favoring a hypothesis of modularity, psychologically real or *modular-so-far*. But

---

[67] How about a complex system, like a language module, instead of the simpler ones we are looking at? Some part of it is innate at the outset and used to build up a rich set of rules. At the end of development, can you make a Poverty argument? Yes, you can argue that the core bits that were not changed during development are innate. That bit is "the module". The new stuff that was learned is not part of the modular, fixed element. At least, we have not argument to demonstrate that it is modular. It's all new stuff. And the old stuff that we started with innately was replaced, so that clearly wasn't modular. But we do have a *core modular part* of the language capacity.

modular-so-far is a somewhat strong observation. It should not be unusual for capacities to exhibit merely *temporary* modularity. After developing a chess playing skill, I may simply play unreflectively for some amount of time without revising my body of chess-playing rules at all; then I run into a hard problem and revise some core rule. In this case, the chess playing skill will look like a module for a little while. That is not so interesting.

But modular-so-far is stronger than a brief episode of isolation. It implies that a cognitive capacity is unchanged since birth and through all the stages of development up until the capacity is observed. For example, to show that infants have expectations about physical object constancy is to show that these expectations are innate but perhaps not interestingly modular. But to show that a cognitive capacity continues to be expressed as such at later stages of development is highly significant. It means that this set of information is unaffected by a substantial amount of experience, experience which has substantial and wide-ranging effects on all aspects of the cognitive mind. This is precisely what has been observed about fundamental aspects of linguistic ability, vision, theory of mind, folkbiology, naïve sociology, and many other areas where nativist hypotheses are offered. Adults exhibit the core innate aspects of these capacities in modular-so-far form. In these cases, and very many cases are like this, it is highly reasonable to infer that the modular-so-far structure is such in virtue of being modular in the full, psychologically real sense.[68]

---

[68] Khalidi (2001) suggests a possible objection via his criticism of Cowie (1999). His argument focuses on whether we can call the results of a Poverty argument "domain-specific". He holds a strong criterion for domain-specificity, requiring that the domain-specific knowledge comprise more than a proper subset of

Poverty arguments only demonstrate that a capacity is modular-so-far. This is not *real* modularity, where we know a capacity is *isolated* from input. So while we should acknowledge this reservation, there is some implication from a Poverty argument for a type of modularity.

4.4.3 Sources of Evidence

A distinct reason to expect nativist arguments to imply modularity turns on the nature of evidence available for researchers who make claims about innateness. Khalidi (2001) argues that the Poverty arguments require demonstration of the relative unavailability of information about a particular domain of knowledge. But showing that information is relatively unavailable is difficult for domains where the subject is awash in rich volumes of domain-related inputs. The Poverty argument is strongest where the subject can be shown to receive little or no information on a subject matter, though she develops a full and rich cognitive capacity anyway. This is possible with language, as in the cases of creolization and "wild" children. Spelke and others have tried to observe children as early as possible in life, in order to limit the total possible experience available on which to have learned physical principles. To search for a domain of knowledge subject to this type of impoverishment *just is* to search for *special* capacities with domain-specific

---

the domain's information: it must contain information about the entire domain! This is too strong, I think. A genetics textbook is domain-specific to biology even if it does not cover all the areas of biology. But furthermore, this criticism is not relevant to the modularity concept as I have offered it here—a concept constructed primarily around informational isolation, not domain-specificity.

functions. For these types of capacities, it is most likely that experiments can be conducted around radically impoverished information, and clear results obtained.

This suggests a methodological constraint on which types of cognitive capacities will be found to be innate through the use of Poverty arguments: only those which treat very specific domains of knowledge. Khalidi's deployment of "domain-specific" is idiosyncratic, however. A psychological capacity can only be meaningfully domain-specific if "the skills and abilities in this domain are not easily *generalizable* to other domains" (Khalidi, 2001:194). But this just is a characterization of what it is for a domain to be *special* in the sense invoked in Poverty arguments, a property I have already argued is linked with modularity[69]. We already established that a Poverty argument is not sound unless is establishes that the proposed capacity is special. But Khalidi's point adds that it is difficult to collect evidence for any innate capacity that is not *special*. This added methodological feature reinforces the likelihood that a nativist will find herself with good reason to suggest modularity of the capacity being studied.

4.4.4 Rigidity and Independence

The one classic nativist argument that has not been carried forward by the Minimum Hypothesis and the methodological considerations is the argument for the Development of Fixed Capacities. It is not a feature of this argument that the underlying capacity needs

---

[69] They are linked: a logically unique set of information in the sense discussed in section 3 is "special". It does not relate to anything else. And that is a kind of guarantor of modularity. But, as I note there, it is not *psychologically real* modularity, since no psychological facts are barring the flow of information.

be special or eccentric in the way required by Poverty and Impossibility. There is, however, a feature of the Fixed Capacities argument which independently implies the modular character of the underlying capacity.

An abstract feature of modules that has been relevant to diagnosing their existence is their *fixity* or *insensitivity* with respect to the informational states of other mental capacities or of the environment. To be a module is to be a fixed or constant implementation of a function with respect to variance in the conditions of other capacities or the environment. The argument from Fixed Capacities is also concerned with a certain kind of *fixity*: the fixity of a certain feature of a cognitive capacity across some dimension. This dimension might be that of ontogenesis or development, e.g. the rigid timeline of development of a subject's use of the English irregular past-tense of verbs in Pinker (1994). The dimension could also be cultural history, as with Lennenberg's criterion of the universality of a cognitive trait across temporal cultural change (Lennenberg, 1964). Equally the cultural range could be cross-cultural, as in the case of Lennenberg's other universality criterion. In each case, the observed fixity is compared to the environment. If a feature is rigid in a way that does not correlate with any feature of the environment, then the nativist observes that an *internal* driver must be responsible for this feature.

Each of these three types of rigidity—ontogenetic, historical, and cross-cultural—are variable parameters. This is clearest with ontogenetic features. One development timeline might follow a particular progression with characteristics accumulating at 2 months, 18 months, 3 years, and 5 years, after which the fully adult capacity is expressed. Another

timeline might unfold with its key landmarks falling at 2 years, 4 years, and 10 years. The two timelines are different settings for the ontogenetic parameter, obviously. The exercise of the Fixed Capacities argument is to compare these parameters to events in the environment and look for correlation. Insofar as this is lacking, the argument demonstrates an internal rigidity that cannot be explained by correlation to an external pattern.

The structure applies to the historical and cross-cultural parameters. One setting for each of these parameters is merely "universal" and therefore "species-typical". But it is at least in principle possible that there would be other settings. One physical characteristic that has a non-universal character is the cluster of racial traits. Brown skin is observed very widely but not universally among humans—only a proper subset have this trait. But this does not correlate satisfactorily with any environmental facts. So the Fixed Capacities argument implies a biological or internalist basis for this trait. By parity of argument we would find an implication for temporally or cross-culturally rigid, but not universal, cognitive traits.

So for all three parameters of rigidity, a wide range of settings is possible for any given trait brought under scrutiny. For some capacity T, we might observe it to have a rigidity pattern V^1. When we search the environment and discover that there is nothing to correspond to V^1, we can conclude that V is an internally or innately set parameter.

An interesting feature of nativist investigation in this paradigm is the accumulation of various rigidity settings V^i. For example, linguistic ability, theory of mind, folkbiology, folk physics, naïve sociology, and very many other proposed domains of innate ability unfold each along distinct timelines (Whitney; Gazzaniga; some other Development omnibus sources). *Pace* the Piagetian effort to cluster these capacities together into the development of inter-related features, these timelines are different from each other. $V^{language}$ is different from $V^{folkbiology}$ and so on. But now the argument from Fixed Capacities demonstrates the independence of language from folkbiology just as it originally demonstrated the independence of language from the environment. To the extent that the capacities appear to develop on rigid timelines that are independent from both experience *and each other*, there is good reason to think that they are independent from each other. Essentially the same is possible for variably-widespread cognitive features (if there are any). While this type of argument shows nothing for two traits with the same setting—both "universal" traits, e.g.—it at least has a meaningful result for cases where the settings differ.

Finally, of course, note that dissociations caused by brain damage or other disorders are strong indicators of modularity. If one capacity can completely cease functioning or become radically impaired, while another capacity functions normally, we seem to have good evidence that one capacity does not simply depend on the other. There may of course be a highly complex relationship underlying the two capacities that makes the non-dependence true in a way that actually requires a level of implementational

integration (e.g. a connectionist network); but at the level of the information processing system, the independence is the same result.

4.4.5 Adaptive Reasons

Evolutionary psychology provides an entirely separate line of argument implicating nativism and modularity (Cosmides and Tooby, 1994; Segal, 1996; Samuels 2000). On their argument, modular structure is more practical, and more evolvable. Modular structures are more practical because they are less dependent on the operation of *all* an agents other capacities; disabling one part of a highly modular collection of capacities will not disable the rest. They are also more evolvable, since each structure is more simple than a single all-encompassing structure would be, they are closer in the reach of evolution. They are also more similar to the way nature actually works, developing a autonomously functioning unit to solve each custom problem as it arises. Just as theorists are parsimonious, nature is generally parsimonious in developing solutions only strong enough for the problems presented. If modularity is a feature of evolution, then it is surely innate. Evolution only works by programming innate structures through phylogenesis.

The adaptive argument is quite apart from the others I have presented. It does not link into a methodological explanation for the concurrence of nativism and modularity. Instead, it suggests that the two are likely to co-occur *in fact* without commenting on what is more or less *discoverable*. Furthermore, it is likely also the most tendentious, as the familiar criticism of "adaptationist" argument needs to be suitably rebutted before inferring from the *convenience* of a proposed adaptive trait to the *actuality* of that trait.

4.5 Recap

In the preceding sections I have argued that the typical arguments for innateness give theorists good reason to think that the capacity under study is in fact modular. It is the structure of the arguments themselves that wraps modularity into the conclusion, not an intrinsic feature of the property of being modular or of being innate. In particular, the Poverty of the Stimulus argument relies on the "special" nature of the capacity under study, a key feature of drawing the Minimum Hypothesis from evidence of Poverty. But any capacity for which the Minimum Hypothesis can be asserted will have been demonstrably special up to the point of observation, a fact which guarantees that the capacity will have functioned in an informationally isolated way. This informational isolation in fact implies that the capacity under study is modular. I also argued for two other routes to modularity: that the best evidence for Poverty of the Stimulus is usually related to special and therefore modular capacities; and that Fixed Capacities arguments provide evidence for the developmental and functional independence of cognitive capacities.

In the next section, I turn the argument in the other direction. Reviewing the core arguments for modularity, I suggest that they imply the innateness of some key part of the modular capacity.

5. Arguments for Modularity

There are a number of common arguments for the modularity of a capacity:

- informational encapsulation;
- performance independence;

- damage or disorder instigated dissociations;
- developmental independence;
- adaptationist evolutionary argument;
- engineering considerations; and
- the uniqueness of the capacity's domain.

There is a diverse range of argument for modularity, and this list only represents some of the possible types. In the following section I will consider the typical structure of these arguments as they are used to suggest modularity, and suggest how they link to nativism.

We can classify the first four into a category of roughly similar arguments. Each observes a type of independence and infers from it the existence of informational isolation. Developmental independence is just the appearance of distinct function in the *normal* course, whereas dissociations are the appearance of the same phenomenon under *abnormal* conditions. The last three arguments are more idiosyncratic, and will be considered duly.

5.1 Informational Encapsulation

The modularity of a cognitive capacity is its informational isolation. The most direct demonstration of this feature relies on testing a particular capacity's sensitivity to information states in "nearby" capacities. Frazier (1987) proposes a model of sentence parsing in which syntactic information is processed in a distinct syntax module, prior to the sentence information receiving treatment for semantic content. For this model, a key distinction rests on establishing the informational isolation of the syntax capacity's operation *as against* the semantic module. To test this, Frazier and others have investigated the impact of differently worded sentences on the way syntactically-ambiguous "garden path" sentences are parsed. Do word meanings disambiguate purely

grammatical ambiguities? This type of research is an example of how the informational isolation of a particular capacity can be investigated, by observing its operation under one condition and testing for influence from related capacities.

Pylyshyn (1984) argues that a general-information standard is the primary diagnostic procedure. If a subject knows some piece of information at the level of a higher faculty, it can be used domain generally. That fact is just an observation about how humans reason. But the key test pertains to specific sub-faculties: does the domain-general information have any effect on the operation of the sub-faculty? If a sub-faculty is isolated from top-down information in this way, it is *cognitively impenetrable*. Pylyshyn's test does not establish *lateral*-isolation, in the way Frazier's method seeks to, but neither can alone establish full informational isolation.

In practice, it is very difficult to guarantee that no information is being shared at the core of the process. Even cases that appear to be "interactive" could equally be cases of rapid back-and-forth operation: the first module produces a result which is instantly vetoed by a downstream system, so the first recomputes a new result, and so on (Whitney, 1998). Such structures may be perfectly modular, though apparently making use of information from the processing rules of other modules.

The structure of this argument, then, is to begin by characterizing the operation of a particular capacity. This capacity can be taken to have a particular set of inputs and outputs. Then, by varying some information outside the set of inputs relevant to the

capacity, it can be determined whether any other capacities cause changes in the original input-output mapping. For example, consider the syntax capacity. In principle, any non-syntactic inputs that cause the syntax capacity to change should be counted as evidence against its modularity: in a narrow instance, if a different word causes better sentence disambiguation then the syntax module is relying on non-syntactic facts. Its modularity is dubious. But in a broader case, we see a non-modular behavior where some cue causes a capacity to reconfigure itself or change its original contents permanently. So external information should neither interpose on individual processing tasks *nor* should it modify the module's "core" itself.

5.2 Performance Independence

Aside from the information-level analysis of informational encapsulation, there are other measures of a capacity's functioning. Roger Shepard's pioneering work in methodologically cognitivist psychology focused on the speed of task performance as an index of computational complexity. This fundamental methodological paradigm has been applied widely. The length of time taken to perform a given procedure is a useful signature for that procedure. If similar tasks take widely varying lengths of time, we can infer that distinct functions are being invoked. Researchers also frequently test the interaction between multiple simultaneously performed tasks—such as the task of counting backwards by sevens and observing the colors of several objects (Bloom and Keil, 2002). Insofar as inhibitory or excitatory effects are observed, it can be determined that independent or shared resources are in use. Various other types of *performance*-related measures exist. In general, different observed performance characteristics

between capacities can be enlisted as differentiating signatures—implying modularity for those capacities.

A general feature of this type of evidence is that results are typically resistant to training. Performance times or interference affects can be diminished, but they are demonstrable qualities of the fundamental procedure itself. As such, they are not merely characteristic of optional aspects of the task performance. This is an indication that the relationship between the observed capacities is highly rigid, not modifiable by new information or training. While some performance variance comes purely from experience, other variance links to intrinsic features of the implementing process.

Informational penetrability tests for the stability of the capacities mapping between input and output. But it is also possible that the inputs and outputs may go unchanged, even while the actual processing has changed. Two computationally equivalent machines might implement co-extensive addition functions through different algorithms. Informational penetrability would not detect a switch from using one machine to another. But the performance criterion might. Since the algorithms are physically implemented, it is likely that the two processes will take different times to compute, require different amounts of computational resources, and so on.

The underlying strategy is to use *performance effects* to study the nature of the *competence*. The function implemented by the capacity is identical to the competence, on the picture we have been assuming (also see Chomsky, 1965). The performance is non-

identical. So there are serious limitations on what we can learn about a competence from its performance. However, establishing that different capacities have different performance signatures—run-time, resource demands, error-rates, etc.—permits us to implicate distinct capacities in a single task, as in the case of remembering object colors while counting backwards. And showing that there is no such phenomenon is evidence for modular function.

5.3 Damage or Disorder Instigated Dissociations

A variety of well-known developmental disorders have been enlisted to show how traditionally unitary capacities are in fact divided into smaller modules. Developmental disorders such as Williams syndrome and Specific Language Impairments show that general intelligence can be crippled without material effect on language ability, and vice versa (Pinker, 1994; Cowie, 1999). This shows that the capacities are dissociated from each other, and therefore likely do not rely on each other for proper operation. Insofar as Williams syndrome subjects exhibit command of complex grammar, even while their speech does not contain coherent content, we have evidence that the grammar system operates in isolation from other, general cognitive capacities.

This type of developmental evidence is clearly related to nativism's argument from Fixed Capacities. The difference, however, is that the developmental programs in this instance show *independence* from each other. In the case of Fixed Capacities, they show rigidity with respect to their *own* unfolding. The relation between these features will be considered shortly.

A second related source of dissociations is from damage to various brain areas, the most famous of which we discovered by the 19th Century neurologists: Broca's area and Wernicke's area. The possibility that these functions can be separately disabled with largely independent effects on linguistic ability is an argument for the modularity of language from other systems as well as for the independence of underlying systems. A different conclusion from the apparent neural localization of these functions has been that they are therefore innate (Elman et al. 1996 resist this), but this argument is not necessary for establishing modularity. It does not matter which neural systems actually implement the functions, or even if that implementation is highly plastic, as long as the separation between functions is preserved. This position is quite distinct from the main thrust of connectionist-style critiques of nativism and neural modularity—what is in fact, neural "localization" (Farah, 1994; Fodor and Pylyshyn, 1988).

5.4 Developmental Independence

A distinct feature of development is that it is diachronic. So capacities will achieve their mature states on different schedules. There will be points in time where a particular capacity is functioning normally without recourse to certain others. In these cases, we essentially observe a dissociation under *normal* conditions. One capacity, which in the adult invariably appears alongside another capacity, appears alone and functions normally. For any capacity that follows, it still can be argued that it depends or "scaffolds" in some way on the precedent functions. So double dissociations are difficult to demonstrate purely from diachronic effects.

An important aspect of a capacity is the function it implements, i.e. the informational properties of the capacity before and after some other capacity comes online. It is also useful to compare the informational properties at various stages of development. Does the capacity pass through the same stages, regardless of how other capacities are developing? But it is also important to consider non-informational properties of the capacity such as the performance characteristics at various stages, and the timing and nature of the process of development itself. For example, if children learn language more easily or quickly when they have strong general intelligence, this should count against the simple division between language capacity and all other capacities.

5.5 Adaptationist Evolutionary Argument

Evolutionary psychologists argue that natural selection is more likely to have produced modular structures (Cosmides and Tooby, 1992, 1994; Pinker, 1994). It is far simpler to develop a mechanism that addresses a single problem than it is to develop one tool designed for a general class of problems. And because natural selection typically responds to particular environmental pressures, it is most likely that the expedient approach would have been followed. The result would have been a collection of distinct problem solving modules with self-contained resources.

A second related line of argument turns on the preferability of a modular structure for the fully developed system. Modular structures are more robust in the face of selective impairments. Again, because it is likely that evolution would have developed the more robust system, some evolutionary psychologists conclude that the mind is therefore modular. Of course both species of this argument are heavily contested by theorists who

are skeptical that what is actual is what is optimal (e.g. Gould and Lewontin, 1978). While this is not a widely accepted line of argument for modularity, it deserves mention in this catalog of different argument types.

5.6 Engineering Considerations

Marr (1982) takes the supposition that cognition functions through modular sub-systems as a methodological starting point (Kitcher, 1988). The argument runs from the success of approaches assuming modular function back to the actuality of modular structure in the mind.

5.7 Uniqueness of a Capacity's Domain

It is possible to argue from very strong eccentricity of a domain to the implication that any capacity dealing with it must be modular. Assume that there is some body of knowledge which consists of inter-related propositions, but which is itself not related to any other bodies of information or propositions. It can only be the case, then, that this logically isolated body of knowledge has a kind of modularity. No other facts in the world stand in rational relation to it; they are irrelevant. If the mind were ordered rationally, then this body of information would be completely modular in the mind.

One can imagine a related type of mental capacity: a perfectly isolated mechanism. It is possible that there are mechanisms in the mind that are take inputs from such a specialized domain that they draw no external, informational inputs at all. It may be that they once drew on real inputs—a dinosaur recognizer, for example—but no longer get activated by any stimuli in modern life. Or they may draw inputs from the sub-

psychological level, such as nourishment or even chemical stimuli that correspond in some way to what the mechanisms output. But at the psychological level these mechanisms are primitive producers of outputs. Clearly, these systems are modular.

If we were able to establish the existence of a knowledge module with an ultra-eccentric domain or a mechanism with ultra-eccentric inputs, as described, they would be modular *a fortiori*. This is an interesting feature of the informational isolation picture of modularity, though it has not been a strategy of argument widely employed (but see Fodor, 1983 on "eccentricity").

5.8 Implications for Nativism

A number of arguments connect modularist positions to nativism. The most general argument stems from a position about the possible ways for a system to become modular. Others rely on the inversion of the classic nativist positions on the basis of similar results from research on modularity.

5.8.1 Executive Control and Modularity

One reason modularity does not entail innateness is because modular structure can be acquired. One possibility discussed in Section 2.3 was that a non-modular program could be instructed from without to "Ignore all further instructions". In implementing this procedure, the program would be thenceforth modular, completely isolated from any core procedure modifications.

This is an extreme version of modularity, when something is completely cut off. As an analytical tool, the extreme version is useful. It is important to keep in mind that modularity is more-or-less. Because some information is entering a capacity, we know only that it is somewhat less modular that it could be. Many capacities are interestingly modular even though information can revise them. Most skills are like this, capacities like recognizing your mother or reading can in principle be forgotten or de-learned.

Consider the following categorization of types of modularity, which has been only informally used so far in the discussion:

- *logical modularity* – where a domain of knowledge is logically unique, and completely unrelated to any or some others;
- *information flow modularity* – where a mental structure is physically incapable of receiving information from outside sources because there are no connections or because the formats are unreadable;
- *de facto modularity* – where there are no barriers to information sharing, but accidental features of experience are such that situations that would call for non-modular operation have not occurred;
- *rule-instructed modularity* – where some instruction in a capacity specifically prohibits the transmission or reception of information.

If you think modules are innate structures, you are likely to think information flow modularity is the explanation for modular capacities. Indeed, a number of theorists seem to hold this view (Fodor, 1983; Marr, 1982). Logical modularity is not a thesis in popular use; and so far, de facto modularity has not been considered a likely explanation for most phenomena. But for theorists who have sought to explain modular function in a way that is consistent with a less nativist, developmentally plastic picture of mind, rule-instructed modularity has been appealing (Karmiloff-Smith, 1992; Elman et al., 1996).

On this view, a paradigmatic module can be a skill, such as chess-playing, or color-naming. These are learned capacities, no doubt. But it is argued that they develop into rigid and inflexible routines of cognition. The Stroop task, for example, asks subjects to read color words, themselves written in various colors. Subjects find it difficult to interrupt the "color-naming" procedure which brings the names of visually perceived colors to mind and interferes with the accurate reading of the words. Clearly reading and color-naming are learned, but they appear to behave at least partially as cognitively impenetrable (Stillings, 1987).

The aim of this type of argument, typically, is to make it possible for domain-general learning mechanism to develop modular functions. Karmiloff-Smith (1992), for example, is explicitly working from a Piagetian paradigm where general learning and problem solving procedures construct more specialized ones through the process of development. Part of denying that modularity is innate requires the presumption that the initial condition is non-modular. A condition where information is globally shared is meant to become fragmentary and more isolated, until the condition is modular.

One difficulty with the paradigmatic non-innate modules such as color-naming, is that they are very plausibly modifiable. While it may be quite difficult for subjects to interpose into the color-naming routine during a task performance session, there is nothing to suggest that this could not be un-trained over time. Indeed, many oft-cited skills are precisely domains of behavior where training is essential not only to acquisition but also to on-going modification: musical ability, athletic ability, and other complex task

performance skills. It may indeed be the case that the airline pilot can land a plane virtually as an automaton, but it is not the case that this ability is completely isolated from training or other intentional intervention. On the other hand, you cannot be untrained from experiencing the visual sensations of color, a deliverance of the vision system itself and not a learned capacity.

While the empirical cases do not present convincing examples of learned modules, it is at least logically possible that there may be a non-modular capacity that learns a rule like "ignore further instructions" and thenceforth behaves modularly. This is, however, logically possible on a particular scenario. And I will argue that this scenario is quite apart from the one hoped for by the typical advocates of rule-instructed modularity.

If a cognitive capacity is to be learned, there has to be some sort of structure already available at the outset. Typically this initial store is called a learning mechanism. Exposed to experience it takes shape in some way such that it develops (perhaps only implicitly represented) rules for doing certain operations. Even a connectionist system, which implicitly builds in the rules but does not represent them, can be *described* as a the classical, explicitly rule-based system to which it is equivalent. So for purposes of exploring the possibilities, we can talk as if we are dealing simply with classical computers.

The minimum that a learning mechanism can start with is some basic collection of "housekeeping" rules that instruct it when to write a new rule, how to process that rule,

and how to manipulate existing rules. The classical Turing machine simplifies this to a set of instructions for how to read, interpret, write, and delete symbols on an infinite tape. Call this minimum set of rules the "executive". The executive does not "make decisions" per say, but it is the highest in the hierarchy of mental functions to be developed. All operations of the machine consist in recombining these basic executive operations in various ways.

Now for an example to consider. The mind could be composed of one or many such basic machines. Say there is only one executive. It is domain-general and capable of learning from the experience presented how to perform various capacities. Early in this process, there will be external inputs (from experience) that interact *directly* with the foundational instruction set, feeding it rules to copy into its instructions directly. So there must exist some method of feeding information directly to that core instruction set. Later, as rules accumulate, some of these rules will help assimilate less explicit inputs without explicitly calling a single executive function. Instead, assemblies of executive functions are invoked.

Assume we feed this machine the information to develop a strong chess-playing skill. If the set of rules makes no reference to non-chess capacities, the "program" for chess playing will essentially be a module. We can even give the program an additional command, "Ignore any inputs from other programs". If we do this, it should be sealed off from interaction with other capacities. Of course, it is part of the construction of this case that it is not sealed off from the executive. Indeed, the program's very operation requires

the reading and writing of symbols on a tape, for example. However, if the executive can still interoperate with the program, then the executive can still modify the program. And if it can do that, then special instructions given to the executive can delete the program entirely or change it arbitrarily. So the "module" is not quite perfectly informationally isolated.

So if learned skills were simply learned by some machines with basic operations, it should always be the case that appropriate subsequent training can simply nullify or alter the capacity to large degrees.

However, there might be several executives. Perhaps early training can instruct one executive, the one which is directly exposed to experience, to pass on instructions to an interior executive. The interior system could then develop a program for chess, and finally receive an instruction to "ignore further directives". The interior system could then continue to carry out operations on its stored rules using its executive functions without accepting future modifications.

This architecture however, it should be noted, is weakly modular *from the start*. From the start, there are at least two independent sets of executive rules. And because neither is subordinate to the other, that core set of executive functions is informationally isolated from modifications. If they were not, we would simply be back in the situation where a single "master" executive and the same results would follow. Therefore, we can take a

weak result: if it is possible for a learned skill to be perfectly modular, there must be some unlearned basis such as an executive.

This is a weak result, but it is slightly stronger than it first appears. For every capacity that is independent from any other capacity, the same result should obtain. So each distinct capacity correlates to a distinct basis in something like an executive or processor. Arguments for modularity, then, establish the innate existence of a constellation of processors commensurate with the universe of modules. This is a non-trivially nativist result, strong enough to contrast sharply against the typical anti-nativist positions.

Modularity itself therefore conceptually implies a weak nativism. In general, though, this weak nativism is rather weaker than that espoused by theorists like Chomsky who advocate relatively strong forms of nativism along with modularity. The subsequent sections show how the typical evidence for modularity makes an even stronger case for nativist arguments.

5.8.2 Rigidity and Independence

For the Fixed Capacities style of nativist argument, the typical result is a "signature" that distinguishes one capacity from another. This signature may be a typical resource need, processing time, scope of universality across a species population or across time, and so on. We saw for this type of argument that the accumulation of distinct capacity signatures, V^i, as demonstrated to vary against the patterns observed in nature, was in itself the collection of evidence for modularity. Where V^i differed from V^{i+1}, there was evidence therefore for the distinctness of the two capacities in their fundamental bases.

The inverse argument runs from both Performance Independence and Developmental Independence back to nativism. For both these modularity arguments, a particular parameter is observed to be distinct from that same parameter for other capacities. Performance measures include processing time, interference effects and resource demands. Developmental features include relative position in the developmental sequence, timetable for the capacities own development, the capacity's experience requirements for activation of the mature function, and others. As various capacities are shown to be independent from each other on the basis of distinct settings for their parameters, that data itself becomes the basis for running a Fixed Capacities argument. Insofar as the various capacities have idiosyncratic and independent profiles, the chance that this corresponds regularly to a pattern in nature diminishes. The more independent capacities are from each other—and therefore the less "uniform" or Piagetian the steps of development—then the less likely it is that the environment is the controlling variable.

5.8.3 Common Evidence

Modularity as a fact about cognitive architecture is quite different from a claim about cognitive content. Whereas Chomsky's cognitivist hypothesis is the classic claim that a particular phenomenon is explained only by appeal to a particular state of "knowledge", modularity need not involve implication of knowledge-like states. But it is nonetheless a feature of the mind that wants of explanation, in particular of origins or provenance, and can therefore be subject to nativist or empiricist accounts. Whereas much of the argument considered in the review of nativist strategies cast the subject matter as "knowledge", it is in every case legitimate to abstract those arguments to mental *structures* more generally.

Modularity itself is something that can be innate, and it can be demonstrated by familiar lines of argument.

Botterill and Carruthers (2000) suggest that modularity and nativism are "mutually supportive" positions (p. 53). In particular, if we observe a certain modular structure to be true of all humans, then we should be strongly inclined to think it is an *innate*, modular structure (also argued by Khalidi, 2001). In essence, this is a version of the argument from Fixed Capacities for the species-typicality of the modular inter-relation of various mental capacities.

In general, modularity arguments apply to a single individual. If that individual's capacity to perform X is independent from the capacity for Y, modularity is established. Indeed, this is also true for arguments from the Poverty of the Stimulus for nativism. In some instances, arguments are indeed made from this basis, as with "wild" children for nativism or Smith and Tsimpli's subject "Christopher" (Smith and Tsimpli, 1995).[70] In practice, however, modularity claims are generally developed on the observation of a class of individuals. Indeed, it is an assumption of much cognitive science that it pursues the common cognitive features of all human minds via this methodology. The result, of course, is that demonstrations of modularity typically include the claim that this

---

[70] There are other cases too. The evidence for Specific Language Impairment innateness comes from genetic studies of just one family clan. The 19th century neurologists of course began with individual or small groups of cases such as the famous Phinneas Gage, and this is still the case in contemporary lesion studies of humans.

modularity is true *generally* for a large population or for all the human species. Claims of this nature resemble the Fixed Capacities argument closely. If modularity is universal, that is reason already to think that it is innate.

A second thread in this argument pertains to what Griffiths (1999) calls "complex inner structure", another hallmark of nativist theorizing. Modularity is a kind of complex inner structure, in the way the creationist William Paley's famous 'pocket watch found in the forest' is a kind of complex structure. The existence of multifarious internal modules, interactive in some ways and isolated in other ways, specialized to their various tasks while also integrated in an overall way, is a surprising result. Indeed this neatly compartmentalized structure is surprising relative to the surrounding environment, which appears a "blooming, buzzing confusion" of mixed and overlapping inputs. The very complexity of the structure on one hand, and the sharp contrast of this organized, purposive complexity as against its environment on the other hand, suggest that the modularity is in fact there by *design*.

One way to interpret this thread is simply to say that the richness of structure implies its innateness, since it makes the empiricist acquisition hypothesis implausible for this particular case. That is essentially another invocation of the Fixed Capacities argument. The mind has a particular complex structure with no analogue in nature, therefore that structure must come from within.

Another way to read this thread is simply to take its parallelism with arguments for evolutionary design seriously. If complex modular structure does in fact scream out for design, and then perhaps we should attribute it to the author of other natural design: evolution. Since evolution transmits its designs by inheritance, we could conclude that this must be an innate endowment. It is not unreasonable, however, to object that this is simply a case of *artificial* design—where modular structure is the operation of human intelligence. Keep in mind that this objection is different from the more typical explanations given for the appearance of modularity by empiricists. Unlike a bridge, which is designed and labored on publicly by conscious human designers, no one has suggested that modular is the *conscious* and *explicit* construct of human reasoning. Indeed, the best efforts of an entire community of researchers have not yet produced a design capable of imitating the operation of the natural mind.

5.8.4 Adaptive Reasons, Engineering and Uniqueness

Evolutionary psychologists have offered their own distinct connections between modularity and nativism. We have already considered their reasons for thinking that many environmental problems require *adapted* solutions and, *a fortiori*, innate ones. Insofar as their arguments succeed for modularity being a likely product of evolution, then it is necessarily the case that the modular structure itself is innately endowed.

Let me just mention here that Engineering Considerations need not imply anything about innateness. There is no reason why a methodological commitment to cognitive modularity precludes a commitment to developmental empiricism for those functions. This is a common enough view, and many connectionists seem to hold it (Shallice, 1988;

Farah, 1994). Also, a logical argument for the Uniqueness of a domain is not yet an argument for its innateness. Depending on how learning works, it might be possible to acquire knowledge about a subject with no rational connections to anything already believed. A "copying" model of learning might permit this, where physical transfer of symbols is effected without rational engagement of the meanings.

5.9 Recap

In this section I have presented a constellation of distinct arguments for modularity. The main thread that binds the principal, cognitive-scientific arguments is the demonstration of *independence* in one domain as a proxy for direct evidence of the capacity's informational independence (isolation). A general result of these arguments is an implication for nativism. The structure they imply is non-revisable, and therefore primitive. Since only innate or non-psychologically acquired factors can produce primitive constraints, at least some modularity phenomena are likely to be innate.

Chapter 5. The Concept of Domain-Specificity

1. The Role for Domain-Specificity

The concept of domain-specificity has begun to assume an increasingly critical role in cognitive science (Hirschfeld and Gelman, 1994). Though the concept has very long roots tied up with the earliest discussions of nativist or modular theories of mind, the earlier usage of mental "specialization" has only lately developed into a more detailed notion of domain-specificity. A number of progressive research programs rely on this concept centrally, including the basic doctrine at the center of the discipline.

Domain-specificity's importance has grown with the dominance of three major lines of research in cognitive science: computationalism, nativism, and modularity. The computational theory of mind is the core doctrine of contemporary cognitive science. It is the basic framework that animates a great deal of the research, and around which important critiques or alternative programs are organized (Elman et al. 1996; Farah 1994). This approach begins by understanding the mind to be a computational system implemented by the brain. Construed liberally, this includes connectionism and related approaches (Sterelny, 1991; Fodor and Pylyshyn, 1988). This system is widely seen to be composed of partially isolated but interacting modules, many of which have important innate characteristics. These various modules are responsible for particular cognitive

capacities, such as language, vision, and so on. (This approach and its confederated ideas are explored in Chapter 2.)

A key characteristic of this research program in the theory of mind is that modules are domain-specific. Modules are considered domain-specific, roughly, when they are specialized for operating on a particular subject matter (such as language) or a particular type or modality of input (such as auditory stimulation). The sheer ubiquity of this doctrine is striking; if a theorist is working in the modular-nativist paradigm set out in Chomsky (1980), Marr (1982), and Fodor (1983), then the theorist thinks modules are domain-specific. This observation alone is enough to justify a closer examination of the concept of domain-specificity.

Domain-specificity is usually introduced in a sketchy way, and little effort is made to articulate a precise criterion for it attribution (Elman et al., 1996). Sometimes domain-specific specialization is cast in terms of the particular resources of the module itself (Leslie, 1995), such as the tacit knowledge or mechanisms it deploys (Carey and Spelke, 1995; Hirschfeld and Gelman, 1995; Chomsky, 1980); and in other contexts the specialization is cast in terms of the subject matter or set of inputs treated by the capacity (Fodor, 1983). A number of different sketches for making the concept precise are available; the slightest analysis, though, suggests these views are seriously flawed. But even without any settled view to draw on, nearly every theorist gives some role to domain-specificity in characterizing modularity and cognitive capacities. The aim of this

chapter is to suggest a path forward for domain-specificity, prune off some bad options and advocate a promising framework.

Bringing attention to bear on domain-specificity itself as an independent notion is important because of the explanatory weight that a number of moves in recent theorizing have shifted onto it. Some theorists make domain-specificity a lynchpin concept in their models. Coltheart (1999) defines modularity itself precisely as the domain-specificity of various cognitive capacities. To be a module is to be domain-specific in a particular way. Elman et al. (1996) and Karmiloff-Smith (1992) treat domain-specificity as the "crucial" feature of any module. Farah (1994), in an important criticism of neurobiological modularity, nonetheless concedes that the "domain-specific" characterization of modularity is independent of the flawed "locality assumption", and perhaps the significant valid plank of modularity claims. Fodor (2000) gives domain-specificity a different but important role as well, claiming it is one of the ways a module can be "informationally encapsulated", the single feature he takes to be constitutive of modularity. So as with Coltheart, if a capacity is domain-specific, then it is modular (though it can be modular in other ways too). On both Fodor's and Coltheart's types of views, the basic modularity thesis depends in large part on the cognitive capacities being domain-specific. This whole first group make domain-specificity *the key* idea for understanding modularity.

A second variety of modular views give domain-specificity a less rigid, but still very important role. Fodor's (1983) most well-known treatment of modularity attributes a

"diagnostic" role to domain-specificity as a reliable indicator that a system will be a module. Connectionists design their systems to function as domain-specific systems, even though they do not typically accept that the specialization exists *ab initio* (Rumelhart and McClelland, 1986; Whitney, 1998) . Nearly every candidate for status as a cognitive module explicitly bears the characterization of being domain-specific. Whether or not it is a constitutive feature of modules, it is widely recognized to be a key empirical attribute.

In a third vein, other theorists link domain-specificity into a constitutive role for nativism, another important element of the dominant cognitive psychology. Cowie (1999, 2001), for instance, plumbs the historical debate between innate and acquired knowledge to find that nativist arguments typically come down to claims for the domain-specificity of a cognitive capacity. To claim that language is innate, she argues, is precisely to claim that the mind has domain-specific knowledge of language. It simply isn't a point of debate whether there is *some* knowledge present at birth; even classical empiricists like Locke would agree that there is. The question is whether it is domain-specific, according to Cowie's view. Even commentators that do not agree with the reduction of historical nativism into a type of domain-specificity claim do agree that domain-specificity is a major feature of nativist claims: the nativist usually posits an innate endowment of domain-specific knowledge (Chapter 2 and 4; Khalidi 2001; Samuels, 2000; Botterill and Carruthers, 2000; Keil 1999). The broad paradigm of modular-nativist theorizing about the mind relies conceptually on domain-specificity at various points.

More specific sub-areas within cognitive science, focusing on understanding particular capacities, give domain-specificity a special role as well. There is a wide-ranging program of inquiry into the developmental psychology of various cognitive abilities, such as mathematics (McCloskey, 1992; Campbell, 1994), folk psychology (Davies and Stone, 1995a; Carruthers and Smith, 1996), folk physics (Spelke, 1990, 1991), and others. Researchers in these fields study the origin and development of many of our higher order cognitive processes pertaining to particular subject matters. Partly on the strength of observed subject-matter effects—where development rates and functional independence suggest that the cognitive capacities responsible for various subjects are distinct—these researchers have increasingly posited domain-specific capacities to explain large classes of psychological performance. Spelke and Carey (1995) for example argue that we have a domain-specific faculty for reasoning about physical objects, and another one for reasoning about social situations. For this class of researchers, the domain-specificity claim plays an important role in distinguishing their areas of research. It also plays a role in identifying which behavior and stimuli are relevant areas of study.

A second research program in cognitive science that puts weight on domain-specificity is evolutionary psychology (Barkow et al. 1992). These researchers adopt some of the basic tenets of the computational, modular and nativist approach to the cognitive mind, but add an emphasis on the evolutionary history of these individual modules. Their version of the modular mind is also more extreme than typical models, holding that there is no "governing" level arching over the specialized modules (of the type espoused by Fodor, 1983 and followers). The mind is entirely composed of many distinct, specialized

modules. Every module is domain-specific. Furthermore, this domain-specificity plays a crucial role in the arguments supporting this "massively" modular picture of the mind over a hybrid partly-modular, partly-general model such as Fodor's (1983) or a completely domain-general model such as some connectionists (Cosmides and Tooby, 1994). They argue that the domain-specific modules are the only ones natural selection could have operated to encourage during evolutionary history. Domain-generality is impossible for an evolved organism. Running an argument of this nature relies on the possibility of providing an anchoring account of domain-specificity, on which natural selection can operate (Atkinson and Wheeler, unpublished; Cowie, 2000).

Overall, then, domain-specificity plays an important role in quite a diversity of important programs. It is frequently invoked to analyze other key concepts—such as modularity or evolvability—but it is more rarely itself investigated in this manner. As a result, it is difficult to draw out from the various discussions a completely perspicuous characterization of this concept. I will argue below, in fact, that the gestures toward explication such as they are imply various *different* and *conflicting* accounts. But it is notable at this point that the murky nature of domain-specificity is not regarded as a crisis within the discipline. On the contrary, terms like "specialization", "domain", and "domain-specific" are freely and often used in characterizing cognitive capacities, modules, knowledge, mental mechanisms, sensory modalities, and other species of psychological structures.

The goal of this chapter is to problematize the concept of domain-specificity, offer an analysis of its constitutive features, and suggest a framework for domain-specificity claims. We will not end up with a sharp list of application criteria for the concept, unfortunately. Nonetheless, the aim is to make a useful advance by ruling out broad swaths of unpromising territory and keep one promising pathway open—a series of negative arguments as well as a defense what I will call the *informational* approach.

2. Features of Domain-Specificity

There is a general, intuitive conception of domain-specificity that is not tendentious, even among those theorists who try to give it a more technically precise account. Khalidi offers one fairly neutral way of putting it:

> To say that a cognitive capacity or set of beliefs or collection of ideas is domain-specific is to say that it is dedicated to solving a restricted class of problems relating to a certain field of inquiry or range of phenomena. (Khalidi, 2000: 194)

This characterization captures the intuitive sense of the concept frequently found in the literature. A mental structure is domain-specific because it has some proper area of application, it is "dedicated to" a particular type of input. This specialization suggests that it is better suited for handling this type of input than other mental structures might be, and this element of normativity suggests that the structure solves a problem rather than merely accepting a type of input. The vision system, for example, solves the problem of collecting visual information about the external world and identifying its relevant features

with limited resources. It solves this problem better than our other systems are able to, and this problem is its "proper" domain in that the vision system is not well-suited to handling any other types of problems.

The proposed formulation is also neutral about a number of tricky issues: (a) how "restricted" must a capacity's focus be to be "specific", (b) what type of restriction is the appropriate bound for a domain, (c) what sort of thing is part of a domain, (d) what sort of system can be domain-specific, and (e) how do we judge that a system is "dedicated" to a particular domain? There are several options for all of these, as well as for the questions on which take Khalidi's remark above to suggest a fixed treatment. The variety is a problem. Consider just the issue of how restricted a capacity has to be before it is "specific". Mathematical reasoning is sometimes described as a domain-specific capacity, though the class of mathematical problems is innumerably large. Then, by contrast, theorists have suggested that arithmetic reasoning alone, a subset of mathematical reasoning, is itself domain-specific. But then how to contrast this with the "specificity" of mathematical reasoning? Or does this mean that mathematics is in fact domain-general? It is not an idle question, since determination of specificity will imply the prospects for adaptationist or nativist investigation of the particular module, etc. As a start at least, the options for each issue fall along a few major parameters, which I think structure the possibilities for domain-specificity. The next several sections consider these parameters in more detail. Then, the following major section lays out a few assemblies of these options as the three major proposals for how to take domain-specificity.

2.1 Scope

A cognitive capacity is characterized as domain-specific in contrast to domain-general. Working only with the rough notion of domain-specificity, it is already clear that there will be a continuum of specialization between two idealized poles. Domain-specificity is a characterization of the *scope* of a cognitive function on some range of inputs or problems. It is unlikely that any system can be truly domain-general in the broadest sense, such that it operates on every possible input in every possible format or context. A Turing machine is frequently given as a canonical example of a domain-general machine, since it is capable of operating on problems of any subject matter. But this is misleading. Such machines still have restrictive conditions on input format even when the subject matter of the input is relevant to what the machine is programmed to do. Though not perfectly general, this type of system is still quite general when compared to something like a calculator or other specialized tool. Equally, regardless of the extent to which a system is specialized, it will still be possible to imagine a more specialized system that operates on only some subset of the inputs—an even more domain-specific system. A "verb-conjugator" might be specialized for the domain of "verbs", but one can easily imagine a more specialized device only for "irregular verbs", or only for "irregular verbs that start with A". Atkinson and Wheeler (2002) suggest that "domain-specific" and "domain-general" deserve a strictly "relative" construal, rather than what sometimes

seems to be a usage more like a category-designation.[71] A minimum conclusion to draw on the issue of scope, however, is that there are more than the two end-point options.

One way to improve this situation is to recognize "domains" as the unit of measure for assessing the scope of a particular capacity's domain-specificity. If the range of inputs on which a capacity operates can be made granular by outlining the various distinct domains, then domain-specificity can be measured as pertaining to one domain, pertaining to some or many domains, or pertaining to all domains. In this manner, one might be able to compare the relative domain-specificity of various capacities.

For this quantizing to work, "domain" has to be a substantive notion. If a domain is merely a "set" of problems or inputs, then finding the smallest possible domain requires individuating a single "problem" or "input". The problem of individuating a domain thus becomes a problem of individuating the problems or inputs on which capacities operate. This "grain problem" is perhaps more difficult at this lower level. On the other hand, a number of intuitive characterizations of domains are available, suggesting that a more technical account is possible at this level—"gross modalities" such as language or vision (Fodor, 1983; Arbib, 1987), functional areas such as folk psychology or analogical

[71] Another term sometimes encountered is "domain-neutral". This is often simply meant to contrast with domain-specific, in that the mechanism or information of the mental structure is neutral with respect to which domain it is applied on. In this sense, it must be equivalent to domain-general. But it seems a domain-neutral could be equally useless for all domains, even though it is not specialized for any one.

reasoning, or evolutionary problems such as “cheater-detection” (Cosmides and Tooby, 1992). The natural next step, then, is to look at domains.

2.2 Domains

The meaning of “domain” is surely central to an account of domain-specificity. We have seen in considering its scopal character that domain will have to be a countable or quantized category. Domains divide and classify the various inputs or problems on which cognitive capacities operate. The largest possible domain is simply the class of all information, objects, events or problems on which cognitive capacities can be said to operate. Saying this already makes a crucial advance a notion of domains simply as “subject matters”. Subject matter has no useful meaning in discussing psychological capacities unless we can describe the bounds of any particular subject. Yet this largest domain is a very general class of inputs indeed, and highly-dependant on how we describe these capacities (i.e. information processing systems, biochemical systems, evolutionary adaptations, etc.). Nonetheless, the simplest characterization of domain just collects various of these phenomena into different sets. For example, a domain could consist of “all electromagnetic radiation between red and purple”, the visible spectrum, or of “all items of all apparently living organisms ”, the class of folkbiological stimuli. This seems right but hardly enough, since we want the criterion to give us a way of dividing up the set membership as well. How many domains of input does the vision system accept? Perhaps just one, as in “all energy waves in such-and-such spectrum”; or perhaps some number matching the number of distinct sensory subsystems, as in “edges, movement, color, solidity, etc.” We need the account of domains to help us decide this and then compare the resulting groups of domains.

The goal for sharpening the notion of domain, then, will to be to find a standardized way of describing classes of psychological inputs in a way that leaves us with comparable units. When two capacities range over one domain each, we should be able to conclude they have the same scope. Second, the method should track our intuition that capacities really do vary in their relative domain-specificity. If every capacity has the same scope, then we have failed. Third, the class boundaries should be well motivated or "psychologically real". We want to know what *psychological* facts are relevant, not purely logical abstract facts.

The first option in adding substance to the idea of a domain is the question of who is to be master: the capacity or the domain? If the capacity, then we might define a domain as precisely that set of phenomena to which a given cognitive capacity applies. We can then distinguish the linguistic domain from any other phenomena by finding cases where the language module is invoked and where it is not. If instead the domain comes first, then domains will exist as they do regardless of how the mind is organized. We might easily find a mental capacity that straddles two domains, only partly solving problems of language and also of vision, for example. On this latter option, we will have to give criteria that a set of phenomena must meet to count as a domain. If we think language is a domain, regardless of how the mind happens to handle it, then we must be able to articulate abstract features of the set of problems—linguistic problems—that merit calling it one set instead of several sets or a mere part of a larger set. It could also turn out that

neither capacity nor domain is master, and that restrictions from both ends will go into delineating domain-membership.

Nearly everyone participating in this debate (Fodor, 2000; Cowie, 2000; Khalidi, 2000), has observed that "domain-specificity" risks an obvious and trivializing characterization. This risk is acute if we define domains by the way capacities divide up the tasks in the world. What is it for a capacity to be "specific" to a domain? If domain simply means, the class of problems to which a particular capacity pertains, then capacities can be nothing else but domain-specific. For if the language capacity pertained to anything more than what we intuitively consider linguistic phenomena, then these extra-linguistic phenomena would be technically defined into the "domain" of language. In this sense, even an idealized domain-general capacity is domain-specific – it pertains only to its own domain. This risk is unique to *capacity-dependant* accounts of domain, since a *capacity-autonomous* account will be able to add some restrictions that prevent the domain membership from slavishly following the cognitive capacity's range.

With a capacity-dependent account, we can avoid the path to triviality either by adding in autonomous criteria or by introducing a new distinction. The new distinction would separate a capacity's *actual* domain from its *proper* domain, a distinction familiar from discussions of biological function (Buller, 1999; Wright, 1973). The actual domain would be the trivial set of all phenomena on which the capacity performs. The proper domain, however, draws from some third source of justification—neither facts about the items in

the domain (as in the autonomous criteria case), nor facts about the capacity's actual performance.

An obvious third candidate might be the adapted function of the capacity as its proper domain (Cowie, 2000a; Cosmides and Tooby, 1994).[72] So we might say that the proper domain of language is the adaptive purpose for which it was designed by natural selection. Because we actually use language to treat problems well outside the range of those encountered in evolutionary contexts—such as using language for philosophy or astrophysics, perhaps—the language capacity is relatively domain-general. It functions outside its proper, adapted domain. One danger of an adaptive account is that an idealized "domain-general" system such as a connectionist network would not count as domain-general even if it indeed was selected by evolution precisely to be a general purpose problem solver (for example, in virtue of the Baldwin Effect, where increased plasticity is sometimes selected; cf. Godfrey-Smith, unpublished).

The autonomous criteria for domains take a more abstract tack, but they avoid the need to distinguish a proper domain from the actual. The simplest autonomous set-like constraint is that a domain is a collection of some problems or inputs to a cognitive capacity; that is the rule that gave us the largest possible domain, a simple set of all possible facts or objects that could trigger psychological response. Looking for a ways to restrict this further, an obvious addition is the criterion that the constituents of a domain are *coherent*.

---

[72] We will return to this issue in discussing how a capacity can be "dedicated" to a particular domain. Being adapted to a domain is one such way.

The constituent elements should all fall under the same description. This restriction is like the ordinary usage of set, where some property applies to all set members. This minimum requirement just demands that the element have at least one description in common. Some sets will have fairly complicated membership conditions, such as the class of linguistic phenomena. In such cases, the rules that describe the property should maintain the required coherence.

Moreover, requiring coherence can demand that the collection of membership rules or the facts that constitute the set fit together in a structure such that all the elements form a network of conceptual dependencies.[73] Any two arbitrary membership rules or facts need not be connected directly, as long as there is some precedent principle on which they both rely, for example. This avoids motley sets of rules producing disjointed sets of items. The class of all linguistic facts might share a few common principles about carrying information, being syntactically structured, being generative, and so on. All and only those things in the set of linguistic facts will meet these linguistic principles. The coherence constraint, therefore, requires that elements of a domain fall under a common description, and that this description consist only of logically related principles.

---

[73] If the set of physical states is demarcated by *rules* about patterns of apparent light contrast, the rules should make use of interlinked concepts of light and dark, color, etc. If instead, the set of inputs is simply a collection of *facts* about light patterns, we should apply the same conceptual interdependence requirement. So the coherence requirement applies differently depending on what the constituent elements of the domain are.

The coherence constraint orders the potentially diverse subject matters of the domain, but it leaves the possibility that the set is too big or too small. The domain of "everything" does not seem like a suitable domain—that's just what it is to be domain-general. Yet it is tricky to characterize a restriction on size since the basic concept of set is intrinsically "nestable". There are a few concepts—like extension, non-contradiction, or existence—which seem to be common principles for all phenomena. Logically, every domain is a subdomain of a domain rooted in these most basic principles. This single super-set would be an undesirable domain to permit, since it makes domain-general mechanisms look *specific*. To be domain-general should mean to range across many domains. One response, though, might simply allow large sets, such as the domain of everything and insist that the size of a domain is what distinguishes relative domain-specificity. That runs into problems already discussed, however, of measuring the size of a domain.

A second tactic might force branches to count as their own domains based on some gradient of "relatedness". Perhaps phonology is a branch of linguistics, but not independent enough from linguistics' other branches to be its own domain. On the other hand, physics and chemistry are independent enough branches of "natural science" to qualify as independent domains. This seems a hopeless strategy in the psychological realm, no easier than other efforts to "carve nature at its joints" in philosophy of science. Let us leave it aside for now.

A third, more promising strategy has been suggested by Fodor (1983). This tactic depends on empirically contingent facts about the subject matters at hand. Some of them,

such as language, will be “eccentric”. The laws and principles that organize the domain of linguistic knowledge, for example, are eccentric in that they do not cover any phenomena outside this domain. Eccentricity is a feature of the basic principles in a domain, not a way to draw the boundaries. But since only linguistic facts will be coherent with x-bar phrase or clitic addition rules, the linguistic domain will not include facts about baseball or mathematics. We could define domains as the sets of elements defined by any eccentric group of principles.

An obvious remaining problem, of course, is that there may not be many eccentric domains. Language, for example, manages to be discontinuous with the rest of natural science because it is not a feature of the external world, but is itself mind-dependent. How the mind processes language is how language works. So while “vision” and “hearing” seem intuitively to be completely distinct faculties treating separate domains, the physical roots defining each domain are nonetheless obvious. Indeed, this is likely true for vast regions of mind-independent science. The remaining areas, like linguistics, may themselves fall into a large domain of all sciences of the mind, leaving only one integrated domain for the world and one for the mind, though each may be separate from the other

As a final proposal, we might think that domains do not have sharp borders but instead define degrees of relatedness to the core ideas of the domain. So as quasi-linguistic phenomena start diminishing the extent to which they are syntactic or generative, they fall less and less clearly in the domain of language. Vague borders for these groups, while

not ideal, are still consistent with the typical way these groups are used. Domains, on this proposal, would overlap with each other on this proposal, and be distinguished by their core or paradigmatic elements. Domain-membership on this account will always be a matter of degree, and so domain-specificity would become a matter of emphasis. Domain-general capacities would be those that ranged across very distant, very unrelated domains. While not as sharp as we might like, this is otherwise consistent with what we need from the concept of domain.

A related concern is that domains be too small. Should a single fact or problem count as a domain? The problem of "conjugating verbs that begin with A" probably meets the coherence constraint, and its domain is eccentric enough to keep out non-linguistic facts. But it seems unnatural to distinguish out this group as a subset from the more apparently natural "verb conjugation", or perhaps "language". It seems a domain should be the *maximal* coherent set possible, rather than simply any coherent set. Such a criterion will glom together the limitless subsets of a natural domain by adding any coherent fact not yet included in the domain, and leaving a more intuitive classification.

The usually discussed cognitive domains, such as language, vision, or folk psychology, make use of at least the most formal of the criteria we have been discussing. The elements of any domain are *coherent*, and include all available coherent information (*maximal*).. If we were to discover animal languages, even though humans could not understand them, they would surely fall into the linguistic domain. The domain is capacity-autonomous, structured on its conceptual foundations rather than the human

brain's features, such as generativity or compositionality.. Finally, the usually discussed domains do seem to segregate areas of physical phenomena that are strictly coherent with each other at the foundations, but they do so in a way that permits of unclear category membership. So while folk psychology is a different set of problems than self-knowledge, their reliance on certain common elements—such as belief-desire psychology—permits distinguishing the two domains while admitting overlap.

The most promising direction for treating domains, then, would pick up on the approach we have been using: a domain defines a maximal set of inputs that fall to varying degrees under a coherent group of principles. A cursory look at some standard domains of specialized cognitive capacities, such as language or folk psychology, suggests that this is useful way of characterizing the concept.

2.3 Contents of Domains

So far we have not set what sort of things will be element within domains. We only know that they will be psychological inputs. But there are a number of options for this, and the criteria developed in the preceding section will depend on our choices here. To define the *bounding conditions* of a domain, we must know what sort of criteria it makes sense to apply. The *capacity-autonomous* criteria on domains like those that seemed most promising apply easily to any collection of propositions or other logically-structured *information*. If the domain on which the language faculty operates is simply a body of linguistic facts, then this characterization seems apt. This kind of domain works well, for example, with Chomsky's type of cognitive module, where syntax is actually a body of information in the speaker's head. The subject matter of this knowledge are the rules and

principles that govern any natural language. Clearly, however, the syntax is not the input itself—the concrete objects which get encountered are utterances. Are these utterances themselves part of the domain of the language faculty? It seems the answer had better be yes if the domain is to be composed of actual psychological inputs, and not just *descriptions* of those inputs.

The more general issue is to describe what sort of stuff could possibly be in the domain of a cognitive capacity. Clearly, we know that the mind processes language, visual information, faces, animals, and so on. But at what level of description can we say that an animal or utterance is in the domain of human cognition? We might be tempted to say that the physical thing itself is in the domain, and this would surely be the Skinnerian treatment. A person encounters a physical animal after all, and any resulting cognitions are produced only by that. This would be a problem for the way we have treated domains thus far. Animals are not "coherent" with each other, even though they can be classed into sets.

This natural approach encounters other problems immediately, since the physical particulars themselves demonstrably matter less than more abstract features. A live cow, a convincing facsimile of a cow, and even a rough sketch all have equal power to trigger animal-specific reasoning, yet beef does not. The folkbiology capacity, which attributes certain essentialist features to anything considered biological, is equally triggered by any invocation of the concept of a cow. Similarly, linguistic input can be received in oral, visual-gestured, visual-written, or tactile-written form, and all of these forms involve the

language faculty's syntax, lexicon, and other units. In these cases, the clearest way of characterizing the domain of inputs is to describe particular features or properties shared by any common trigger. Indeed, this is also true for direct physical stimuli to the peripheral sensory modalities; the color red is the common property which causes sensations of red in the vision system.

The dominant view in psychology about how physical properties entrain cognitive processes takes up a broadly computational view of the mind. The mind itself is an information processing system implemented by a physical system, and so inputs have both physical and informational properties. A particular cognitive capacity will perform a function on some class of inputs. Spotting a cow, for example, invokes certain informational states that trigger the folkbiology capacity. At a lower level, light wavelengths that stimulate the retinal nerve in a particular way will be interpreted by the color-identifier as signals of redness, giving us a physicalist definition of "red" or "cow". This lets us take input descriptions above the purely physical level and translate them down into sets of physical stimuli. The domain of the faculty, then, can be characterized at a higher level of description by the range of features, concepts, or facts that invoked the faculty, but cashed out in terms of physical stimuli. For a red-detector, this is a very simple condition. For a cow detector it is more complex: there may be a horn, spots, smells and such to take into account. With the computational view we can integrate an informational level characterization of domain contents with the fact that psychological inputs themselves will be physical.

This level of description lets the *autonomous* informational criteria on domains—such as coherence, maximality, eccentricity—obtain. Certain basic features and laws bind together all objects of folkbiological reasoning, for example. Animals have particular manners of locomotion, endure through superficial change, reproduce with like species, and so on. Facts about prime numbers are not relevant, but facts about the present color of an animal's fur do cohere with the other facts by falling under a branch of facts relevant to the typical appearance of animals with fur. Similarly, with regard to vision, the fact that certain types of edges or contrasts exist or not is within the domain of the vision system's feature detector.

Domains are often characterized as subject matters or bodies of related information. This intuitive account works well with the criteria we have given thus far. The bounds of a domain are defined by informational properties of its contents, and so the contents themselves can be described as logically-structured pieces of information. This is also the general picture adopted by theorists who have otherwise varying commitments. For a typically Chomskyan cognitive capacity, one consisting of a body of rules and principles about a given subject, the natural characterization of a domain is patently one involving informational criteria. This is, in large part, the picture adopted by developmental psychologists studying various specific domains of cognition such as mathematics or folk psychology (Hirschfeld and Gelman, 1995).

Equally, however, it fits well with the evolutionary psychologists' focus on adaptive problems (Atkinson and Wheeler, 2002). Rather than identifying subject areas with

relation to the concepts that organize a cluster of information in virtue of the information's own putatively intrinsic structure—as "physical", "military", or "financial" would carve out the joints of natural or social phenomena—evolutionary psychologists instead select "persisting evolutionary problems" as the phenomena to which all elements of a domain should be related (Cosmides and Tooby, 1992). Mate-selection, for example, would be a domain of issues about the cues for use in mate-assessment and their likely meanings. Even capacity-dependent accounts are likely to define the contents of domains by the informational properties of the inputs, while not relying on any *autonomous* restrictions on the character of a domain.

In essence, this general view of domains as coherent bodies of information reflects a consensus behind a roughly computational view of mind for characterizing the cognitive processes and their objects in spite of the lack of agreement on how precisely certain elements of domain-specificity should be elucidated. While there are radically non-computational views of the mind, such theorists are less interested in domain-specificity.

2.4 Capacities

Since domain-specificity is mainly an issue for theorists working with a computational model of mind, certain aspects of the concept are already fixed by this background theory. Just as the contents of domains get characterized informationally, the capacities themselves are characterized as information processing devices. This is a starting point for considering what sort of thing can be domain-specific at all. In the usage of all major discussants, it is a psychological, cognitive capacity that can be specific to a domain. Yet Khalidi's sketch above suggests a range of options to which we might attribute domain-

specificity: “cognitive capacity, set of beliefs or collection of ideas”; to which we might also add “database” and “mechanism” in the usage of evolutionary psychologists (Cosmides and Tooby, 1992; Atkinson and Wheeler, unpublished); and also “mental structure”, which willfully abstracts from the details (Samuels, 2000). There are other options too, since the fundamental concept of domain-specificity need not be constrained only to mental systems. Genes or cognitive architecture may also be domain-specific, even though the former is not mental and the latter is not itself a proper cognitive process (Elman et al. 1996). Finally, even non-biological things might be domain-specific, in the way a “bicycle pump” is specialized to perform a particular task.

The origin of this variety is a strategy of “hedging bets” around the ultimate, to-be-agreed-upon meaning of “cognitive capacity” or “mental structure”. At one level, it is entirely uncontroversial that the objects of study are psychological faculties. Psycholinguistics, for example, is clearly about a cognitive capacity for the comprehension and production of language. So any account of domain-specificity is at least willing to say that cognitive capacities are the sort of things that are specialized on a domain. But saying more precisely what a “capacity” is creates difficulty.

Capacities come in at least two broad categories: mechanisms and knowledge, or what Carruthers and Smith (1996) call “processors” and “subject matter”.[74] Mechanisms are un-intelligent systems that take physical input and produce a legible output into another system. Pain-detecting nerves or a light-detecting rod in the eye are extreme examples of

[74] This is discussed in more detail in Chapter 3 with respect to folk psychology.

simple mechanisms. Some consider more complicated capacities to be implemented by mechanisms, such as certain parts of memory or the entire vision system. Insofar as non-cognitive things can also be specialized, in the way biological traits can be designed by natural selection or artificial tools can be designed by engineers, domain-specificity can apply to mechanism-type cognitive capacities as well as to bicycle pumps.

Knowledge capacities are distinct precisely because mechanisms do not contain or represent any knowledge-like states. Chomsky's grammar module is the classic example of a body of knowledge-like states invoked to explain the functioning of a psychological capacity. This type of system is domain-specific in virtue of the informational relationship between its explicitly represented contents and the elements of its domain. Since it is just a body of knowledge-like states, such states can equally be domain-specific if they are represented outside the mind (for example, in a computer or written in a textbook) as merely a set of propositions. So while domain-specific mechanisms can include any mental or non-mental tools specialized to treat a range of inputs, any mental or non-mental knowledge-like system is domain-specific if it pertains to only a few domains.

Many types of informational states will count as knowledge-like states, not only explicitly represented propositions in the mind. Given what is known about the brain, much of this is likely to be implicit rather than spelled out in discrete physical symbols. Furthermore, these states will fail to function like garden-variety knowledge. Many are likely to be tacit and isolated from access by general cognition. As a result, knowledge-

like states become difficult to distinguish from mechanisms. Indeed, since any knowledge state must ultimately be implemented in a physical system, it is likely that the two merely describe the same functions at different levels of description.

One consequence of maintaining a view of mechanisms as brute, non-informational systems is a difficulty in explaining their domain-specificity. A system that holds knowledge of the rules of grammar, for example, is itself part of the domain of grammar that it operates on. Its contents are coherent with the grammatical facts and principles underlying speech acts that it encounters. Hence, the knowledge is about the domain to which it relates. Linguistic knowledge is specific to the domain of language. A mechanism on the other hand is not itself coherent with the domain it operates on. Instead, mechanism are relevant to a domain purely insofar as they operate directly on objects governed by the principles of the domain. So a red-detector cone in the retina operates on reflected light of a certain wavelength. Yet insofar as a mechanism is maintained as a brute tool, and not as an implementation of an implicit red-detection rule, the mechanism itself contains no knowledge-like proposition that could be relevant or coherent with the external world. It simply operates on a given set of physical phenomena (light waves of a certain wavelength).

One option at this point is to simply admit, for this and other reasons, that mechanisms and knowledge systems do not describe genuinely contrasting types of mental systems; at best the difference pertains to the level of description chosen, as I have urged in Chapter 3 from independent motivations.

However, a second option, is to jettison the informational account of domains and focus instead on the non-autonomous, *capacity-dependent* account for mechanisms. On this view, the domain of a mechanism is simply the range of phenomena on which it operates. This strategy has, in fact, affinities with the explicit strategy of the evolutionary psychologists. They invoke both mechanisms and knowledge bases, and identify their domains by looking to evolutionary history. Since the methods of evolutionary biology provide independent grounds for identifying human adaptive problems, this taxonomy of cognitive tasks can be applied to organize the space of domains. So organized, cognitive capacities can be judged domain-specific or domain-general based on which problems they address.

Elman et al. (1996) provide a useful discussion of the "levels" at which we can talk about domain-specificity. They highlight five: (a) tasks – particular means-end activities, (b) behavior – a pattern of actions or discrete stimulus-response loop, (c) representations – explicit knowledge-like states, (d) processing mechanisms – like a red-detector or a small quantity estimator, or (e) genes. Each is a stage in the causal story for any person's encounter with the world. A task, like "finding similar pairs", might be domain-general, while "spotting predators", might be highly specific to a particular domain of concern While a useful set of distinctions, the literature in psychology and cognitive science mains focuses on (c) and (d).[75] The categories (a) and (b) play an indirect role in

[75] Just above and in Chapter 3 I argue that this distinction is mistaken. Nevertheless, Elman et al. 1996 are clearly right to distinguish these two levels as different as theoretical instruments of explanation. For the

determining the range for a capacity at the levels (c) or (d). A cognitive capacity—either mechanism or representation—functions through particular behaviors or performs particular tasks. Insofar as the behaviors or tasks it relates to are restricted to certain informational domains, we can call the capacity itself similarly restricted.[76]

In sum, domain-specificity is a property of cognitive capacities, which are broadly of either "knowledge-like" or "mechanism" types. The *capacity-autonomous* informational account of domain membership of the previous sections works best with knowledge-like capacities. If we take all capacities to have a fundamentally knowledge-like structure, then *autonomous* criteria can help us to define domain-specificity. If, however, we insist that there are completely non-knowledge-like mechanistic modules, then we must appeal to *capacity-dependent* criteria or other third sources like evolution.

---

purposes, however, of assigning domain-relevance, it may still be better to treat them each as alternative ways of describing information processing devices. The confusion does not start with deploying "knowledge" vs. "mechanism" in theories—the confusion starts with holding this opposition to follow a deep distinction in the cognitive function of one system as opposed to another.

[76] Genes are not at the heart of this debate; they do come up frequently in discussions of nativism, but primarily (as in Chomsky's usage) as a gesture to the likely biological basis of an innate capacity. The question of whether a gene is domain-specific seems to be more relevant to biology than psychology, as in "genes for depression" or "genes for language".

2.5 Specificity

In the forgoing sections we have been developing a notion of "domain", about the sorts of things that can be elements in a domain and by what sort of criteria they should be grouped. We also considered how the characterization of cognitive capacities influences our understanding of the domains *on which they operate.* The remaining issue relates to understanding the link between capacities and their domains. Domain-specificity is a claim that there is a special link between a capacity and some domain, yet this loose idea leaves us with many ways to interpret it. We need to come up with a non-trivial way of describing the special relationship between a capacity and the domains in its range to which it is specific. The intuitive term "specialized" is often used in a way that is nearly synonymous with a capacity's being domain-specific (e.g. Cosmides and Tooby, 1992). This usage raises the issue of distinguishing between mere pertinence or relevance, on the one hand, and a stronger sense suggested by specialization or specificity. While the contrast between "specific" and "general" describes the scope of a capacity across a continuous range of widths, specificity is also meant to resist a merely trivial quality of pertinence. One way to trivialize domain-specificity is to define a *domain* simply as the body of phenomena on which a capacity operates, as discussed above. We saw that this missed a key aim of the domain-specificity concept, which was to show a normative connection between a particular capacity and its domain rather than a merely accidental one.

A second way to trivialize the concept is to let any domain to which a capacity pertains count as among those to which it is "*specific*". For example, the archetypical domain-

general mechanism is a connectionist network, yet even it will count as domain-specific if its input system only feeds it information about a single domain. Then every system will be trivially "specific" to whatever domain it happens to treat, even when the system has no intrinsic specializations for that function. Therefore we need to adopt more rigorous criteria for domainhood, in order to get granular, countable domains.

Of course, this weak criterion benefits from the restrictions already put in place. First, capacities can be differentiated by their scope, so not all domain-specific capacities are automatically equally domain-specific. So if the connectionist network is restricted only to receiving visual and auditory information, it can at least be judged more domain-specific than a vision-only processor. Second, domains are further required to have certain features, so we do not risk the proliferation of innumerable arbitrary domains. The network deals with two discrete and countable domains, rather than also being domain-specific to red-detection, rabbit-hunting, chess-playing, etc. Third, we know to judge domain-specificity on the level of gross capacities rather than arbitrarily small units of mentality. A single idea, such as the idea that "Napoleon loved art", will not count as a separate trivially domain-specific mental faculty.

Still, there is no further way to guarantee that capacities will only be "domain-specific" to those domains for which they are specialized, or at least where their domain-specificity is "psychologically real" (Khalidi, 2001). There are two types of threats here. The first is *incidentality.* Cognitive capacities are specialized to operate on certain domains of inputs, but it should be possible that some areas on which the capacity *can* operate are not the areas where it is *specialized* to operate. Indeed, the capacity may frequently be invoked in

treating a particular domain, but only *incidentally*, in service of the more specific capacity. One version of incidentality is a case where the object-recognition capacity surprisingly turns out to be invoked in memorizing cloud shapes. Here the capacity is useful in an area where it is not in fact specialized. Another type of incidentality is where the capacity merely "scaffolds" or "conducts" one aspect of a phenomenon without being specialized to the domain of which it is a part. For example, folkbiology requires use of vision to gather input, but the vision system is not "specific" to the domain of folkbiology.

A second type of threat is the threat of *inefficacy*. In this case, a phenomenon may well fall squarely within a capacity's domain. However, the capacity may be ineffective in handling it. For example, humans have been shown to have trouble reasoning with certain basic logical or rational puzzles, such as the Wason task or transitivity (Wason, 1966; Kahneman and Tversky, 1982). So whatever capacity is invoked to handle such puzzles—practical reasoning, or perhaps abstract reasoning—is not very good at abstract logical reasoning, but only at some diminished class of such problems. It seems incorrect, then, to say that this capacity is domain-specific to logical reasoning. Incidentality and inefficacy seem related of course; a capacity is specific to the domain it treats best, and probably less effective in its treatment of incidental domains. But it is also possible that a capacity will treat an incidental domain very well, or that a capacity will be ineffective on the sole domain to which it could be judged specific.[77]

---

[77] Some capacities may not be domain-specific at all. If they are domain-general, and relatively ineffective, they are likely to treat some domains uniquely, though poorly.

To address incidentality, there are several options for enriching the meaning of specialization. The simplest requirement is logical pertinence, that the capacity embody some rules or principles relevant to the content of the domain. This already is enough to control for cases where a capacity play a mere supporting role in cognition about a particular domain, as where vision supplies input for folkbiological reasoning. The second feature is actual range, a capacity should only be judged specific to those domains on which it actually operates or could actually operate. These two features exhibit two ends of a range of possibility with respect to the capacity's context. The language system, for example, sits behind a row of more peripheral sensory capacities. It relies on audition and vision to collection information which it will process. Though it is possible to express language in modalities other than those in the range of human perception, the language system can only process it within a certain range due to certain contingent limitations. Given the contents of the language capacity, for example the syntax and lexicon, it is logically possible to process any natural language presented in any format. Adding in the various contingent limitations on any individual speaker's language ability—facts about his cognitive architecture, about his linguistic experience, about the linguistic community, and so on—there will be a limited range of inputs which he can *actually* process. Actuality matters, since it should not matter in deciding domain-specificity that the lexicon's word-recognizer is in principal a domain-general connectionist network in principle when it is only used to process words. The actual domain-specificity of the process is what matters.

While it seems reasonable to require that the capacity is at least logically relevant to the domain in question, this relevance is not alone enough. If we add the criteria that there be some actual interaction between capacity and domain, a further question arises. If a capacity can only be specific to a domain it actually encounters, arbitrary historical facts may interfere with proper analysis of some categories. For example, a sensory impairment may cut off input to a particular section of sensory cognition. Lack of actual information flow would not make the vision system any less specific to visual phenomena. Rather than taking simply the actual situation, it seems the relevant reference should be somewhat less restrictive. A reasonable bound would be the nomological possibility given the typical biological and psychological facts: if the brain is wired so that a particular connectionist network only functions as a lexicon, then that network is domain-specific to word learning.

These criteria look at the actual facts about a capacity—it wiring, its logical range of relevance—to determine which domains it is specialized to. Alternatively, we might look to historical facts about the capacity, such as its design or evolution. In most cases, insofar as evolution can be shown to have designed a capacity to function on a particular domain, it seems unlikely that this criterion will be much different than the former one. If language ability were selected to operate on linguistic subject matters, this fact could only be demonstrated if in fact the language system was actually capable of acting on that domain. The exception, in this case, would be where the former adaptive function of the system had been "exapted" to some new function (Gould and Vrba, 1982). This is the interesting case: where a capacity like folk psychology finds itself usefully applied to

reasoning about the probable behavior of computers, for example. In such an instance, the advantage of an evolutionary frame is that it lets us distinguish between a capacity that is domain-specific to reasoning about human psychology but also useful to reasoning about other things. This may also be its disadvantage: the history seems irrelevant if the system is effective with its present domains. Both options seem workable.

The final concern about specialization is that a capacity may not actually execute the desired functions very well. For example, we may find, in the human reasoning system, some sort of abstract logical reasoning capacity (or perhaps it may be a feature of an existing capacity, like language). Imagine that this capacity is entirely dedicated to logical deductions. Given humans' demonstrated limitations with abstract logic, it seems inappropriate to call any such discovered system "specific" for logic. Empirical evidence suggests we are actually quite bad at such logic. Surely we would expect any system that was logic-specific, i.e. specialized for logic, to be actually *good* at such reasoning. But so far our domain-specific concept only guarantees that a capacity takes a certain domain as its input; it does not guarantee that the capacity processing this input effectively (regardless of how we interpret effectiveness). In the case of this hypothetical logic capacity, then, at best it seems an instance where the capacity is relevant to logic, actually invoked by logic inputs, and perhaps even selected by evolution specifically to handle logical problems. But we cannot take the further step of actually calling the system effective, which is a strange result. Of course, we could simply take the position that a poorly performing mechanism can be called specialized. This is not too damaging. But if we want to require specialized systems to perform at some particular level of

effectiveness, the only remaining option is to stipulate that level. Simply by stipulation, a domain-specific capacity might be required to be effective in operating on the domain to which it is specialized.

2.6 Recap

This section considered a variety of parameters for characterizing domain-specificity in more detail, as well as some of the principal options in each case:

(a) scope – domain-specificity is a relative characterization of the range of a capacity over domains;
(b) domains – can be defined autonomously or capacity-dependently; in the former case, informational or evolutionary criteria might be applied, while the latter risks trivialization;
(c) contents of domains – probably characterized as bodies of information or subject matters;
(d) capacities – domain-specificity is a property of capacities, which can be mechanisms or knowledge systems; each must be consistent with the account of domains;
(e) specificity – capacities are dedicated to a domain if they actually can engage the domain, if they are relevant to the domain, and if they are effective in their contact with the domain.

This does not constitute a neat menu of options, but shows the family of issues involved. The three main accounts to be considered in the next section can be evaluated using this framework.

3. Different Accounts

The cognitive science literature has deployed a great variety of views about domain-specificity, though many have common elements and all are very briefly sketched (Carey and Spelke, 1995; Leslie, 1995; Hirschfeld and Gelman, 1995; Chomsky, 1980; Fodor, 1983, 2000; Cosmides and Tooby, 1992, 1995; Khalidi, 2001; Elman et al. 1996; Karmiloff-Smith, 1992; Atkinson and Wheeler, unpublished; Sterelny and Griffiths,

1998; Botterill and Carruthers, 2000; Cowie, 1999, 2000; Keil, 2000). This section proposes to look at and evaluate only three in some detail: the "adaptive" account from evolutionary psychology; the "subject matter" account; and Fodor's (2000) mixed "rule range" account.

The intuitive characterization of domain-specificity is not very reliable. It leaves open a long list of features that need to be clarified, as we have seen, to avoid a proliferation of domain-specific capacities. On the one hand, there is this risk of innumerable trivially domain-specific capacities. Equally, there may be a large class of capacities that might be so labeled on insufficient or inconsistent grounds. Given the role of domain-specificity in diagnosing the existence of a cognitive module, evolutionary module, or innate knowledge, this can have serious results.

3.1 Adaptive Domain-Specificity

Evolutionary psychologists have argued that the mind is made up of a large collection of domain-specific modules. Some of these are mechanisms, and others are cognitive "databases", or bodies of knowledge-like states. Domain-specificity on their account is fundamentally a focus of each module on one type of evolutionary problem, since that is the only way natural selection can develop such capacities (Atkinson and Wheeler, unpublished). For example, Cosmides and Tooby (1992) have argued for the existence a cheater-detection module, used to identify violations of deontic rules.

In their usage, every capacity has a very precisely defined scope: one evolutionary problem. Insofar as evolutionary problems can be identified and segregated successfully,

each module will be equally domain-specific. Their usage does not appear to have relative grade, though it seems unlikely that they would specifically prohibit the possibility. The account just fails to present a way of comparing the scope of adaptive problems (e.g. is predator avoidance or food gathering strategy a wider adaptive problem?). Domains are defined by the problem areas, which do not consist of concrete physical entities but rather of challenges or means for better reproduction. Some are abstract, like cheater-detection. As such, the adopted level of description for domains permits the informational criteria discussed above: the relevant concepts with which the problems are posed, like cheater or mate, can be subjected to the coherence, maximality, and eccentricity criteria to produce clear cut domains of adaptive problems.

Capacities also are partly described by the rules and procedures they embody, though evolutionary psychologists do not fully distinguish mechanisms from knowledge-like systems. While some commentators have taken this and a general commitment to computational psychology to imply adherence to a cleavage between mechanisms and knowledge, this may be mistaken (e.g. Atkinson and Wheeler, unpublished). On the contrary, the overall position is better tenable when both types of phenomena are interpreted to be variant implementations of knowledge-like instructions for dealing with inputs. Mechanisms transform input information according to complex rules, e.g. the way a cheater-detector would take observed facts, link them together, add certain decision rules, and produce a conclusion. Knowledge stores, or "databases" as evolutionary psychology typically calls them, are likely to be simpler devices that keep lists of information or rules available and deliver them in response to specific requests. Either

type of system is a computational mechanism, a rule-governed, symbol-processing machine. So while the computers perform different tasks, both the mechanism-types and the processor-types invoked in evolutionary psychology's characterization of capacities are computational. As a result, there is no trouble integrating the informational account of specificity developed in this paper, where we look for coherence and actual relevance to a domain.

One implication of this approach is that domain-specificity is easy to measure for the evolutionary psychologist, as it should be. If a capacity treats its adaptive problem, it is domain-specific; if it treats anything else, it is general. It is also consistent with the information-based criteria on domain-borders, domain-membership and capacity characterization discussed above. Capacities and their domains are described as sets of rules and information, so they can be held to the coherence, maximality, eccentricity and actual relevance standards to avoid trivialization threats. This is a far better situation than if we had to rely on classifying groups of physical events or objects, since we have seen how difficult to clearly characterize the key concepts in those terms. Nonetheless, adaptive domain-specificity faces a number of challenges.

First, there is a methodological inconsistency with much of the literature. As Chomsky illustrates well, domain-specificity is supposed to be the sort of thing one can "read off" the capacity itself. Chomsky is vigorously uninterested in the evolutionary past of the language system, to the point where he even suggests it may originate in quantum mechanics. Nonetheless, he claims that the system is domain-specific, a claim only

possible if evolutionary history is not relevant. More broadly, the force of this criticism seems important. The analogue in biology is the problem of apparent design. Apparent design can be read off an artifact or organism, such that it implies the existence of a designer. Here, domain-specificity seems to be something that evolution should explain, not define. Completely in ignorance of the system's history, we assess that it is domain-specific.

Second, this account makes "domain-general" systems trivially impossible. If any system evolved precisely to treat a very wide range of problems, this account would simply define that range as the domain to which it was "specific". Only if one system could develop in response to the separate pulls of multiple, diverse problems would we end up with a "domain-general" capacity that treated a number of distinct domains. Yet if those several problems susceptible to characterization as a single problem—the problem of preparing for the unexpected, for example—then we would mislabel a general problem solver as a domain-specific tool. In effect, this problem is very similar to a criticism raised by Atkinson and Wheeler (unpublished) in following Sterelny and Griffiths (1999): there is a grain problem in differentiating distinct domains since there are so many ways to describe a single evolutionary problem.

A family of problems also comes handed down from debates on biological function (Godfrey-Smith, unpublished; Cowie, 2001). One in particular is due to Fodor (1990). Evolutionary methods cannot in principle distinguish between an F-detector and an F-or-G-detector. Since evolution creates a poison-response mechanism that mistakenly rejects

harmless allergens like pollen, we cannot be sure which of its actual functions is actually its "proper" function. One way philosophers of biology have approached this problem is to historicize function: the actual problems that this capacity addressed in its evolutionary history constitute its problem domain. Language addressed communication, not mating (Miller, 2001). The result is a historical concept of domain-specificity: though domain-general when first applied to mating games, we should now say of language ability that mate selection has in fact played a role in its development (however weak that may be). It seems wrong to say that capacity, barely changed, has shifted its proper domain simply in virtue of being put to a new use across some generations.

3.2 Rule Range

Fodor (2000) proposes an account that mixes informational measures and constraints from the contingent situation of the cognitive capacity. Any cognitive capacity has some logical range of application. For example, modus ponens applies to all situations where X→Y and X obtain. This logical range is intrinsic in the capacity itself—language applies to all systems of symbols with certain organizing principles of minimalist linguistic theory. Though in fact the system is format-dependent, on visual symbols or temporal expression, the linguistic theory is not. As such, the logical range of the language system is one thing, while the contingent range of application is more narrow. It is narrower due to issues of cognitive architecture, input mechanisms and so on. These constraints diminish the generality with which it can be applied. In the case of modus ponens, the logical form IF X→Y and X, THEN Y appears in some more restricted form, e.g. IF 2→Y and 2, THEN Y. For Fodor, if a capacity is actually implemented in a way that restricts it from operating on its full, logical range, then it is domain-specific. Modus

ponens in this case is specific to the "2" domain. The rule's range determines its general scope, and any limitation constitutes specification.

On this view, domain-generality is only achieved with perfect generality. Anything narrower is a matter of degree. This is roughly consistent with the scopal treatment considered above, though there seems to be a clear limit to how general a capacity can get. This view is also broadly compatible with the informational criteria developed for domain-definition and domain-contents. Fodor's account of capacities also appeals directly to an informational specification of the implemented function. The difficulty, however, comes in the way it matches with specialization.

Fodor's account is based only on the intrinsic range of the capacity. But we have already elaborated the need that the capacity effectively treat a domain. Simply because modus ponens is triggered by "2" does not mean it is at all effective. Falling within the logical or even actual range of the capacity should not be enough to qualify as the subject of its specialization. If the wind triggers my speech recognition capacity (so that I hear "words" being said), it should not thereby fall into the domain of the capacity.[78] The account does include two of the required features for specialization: since the actual implementation's

[78] Two types of errors: one because the logical form of the "competence" is too broad, a second because it is implemented poorly. The former is relevant here. The idealized capacity makes this "error" of hearing the wind talk. But "wind talk" is not in the domain—surely it should be "human speech", with some suitable range of producers.

logical range is considered, this formulation respects the need that the capacity is actually able to treat the domain. Equally, we can be sure that the domain is relevant.

A second problem with this account is that nothing can be intrinsically domain-specific. That is, a capacity can only be domain-specific in virtue of some limitation imposed on top of its pure structure. Yet, linguistic syntax has no proper subject matter other than syntax. It is likely to be "eccentric", or logically unrelated to any other discipline of inquiry that is liable to be encountered. As such, in its plenary implementation, it is still domain-specific to language. But Fodor's treatment classes it as domain-general. This is a grave problem for the account.

Fodor converts all the nomological or accidental issues around the capacity into a formulation of the principle which the mechanism expresses. In doing this, he sticks to a treatment of capacities as implementing knowledge-like rules. A perfectly general modus ponens mechanism simply implements the general rule ("IF P→Q and P, THEN Q"). One can imagine restricting this implementation in various ways. Perhaps the rule written in explicit symbols is simply "IF X=2→Q and X=2, THEN Q". Or perhaps there is a psychological input mechanism set in front of the modus ponens device that can only pass on "X=2", such that the modus ponens device is *implicitly* implementing this "X=2" rule. Or finally, it may simply be that the accidental non-psychological facts of the universe are such that the device can or does simply never encounter inputs other than "2". Equally, in these cases where the machine never receives "3" or "blue" as inputs, it is surely accurate to describe this machine as implementing the "2"-rule. There is no

counterfactual case. The machine is a *de facto* "2"-rule machine, though it is logically possible for it to process non-"2" inputs.

This much is consistent with the informational view we have been using, where a mechanism implicitly implements a knowledge-like rule or procedure. On the account so far, we would judge domain-specificity simply on this logical relevance. But Fodor pares back another level: he claims the "2"-rule is just a domain-based instance of the more fundamental "prime" rule: modus ponens. As such, the instance is constrained by accidental logical facts; the "2"-constraint is accidental to the logical essence of the rule, according to him. The underlying structure of the rule is "modus ponens", and the particular input it is about does not change that. This is a distinction that seems difficult to make. In the case of the "2"-rule, it seems easy to read off the underlying prime rule. Yet, it seems hard to imagine what prime rules will apply to edge-detection, phoneme-parsing, and other functions. Surely all of them will be ultimately reducible to AND-, and NOT-operators. But that seems regressing too far. It is not clear how this notion of prime rules helps us determine what is fully general and what is accidentally specific.

3.3 Subject Matters

Chomsky (1980) developed a view of linguistics as a unique science of the mind, with a unique subject matter and therefore describing a domain-specific cognitive ability. As Carey and Spelke (1995) put it, picking up this account, "each system of knowledge is organized around a distinct body of core principles". From the start, this account has been developed for use with knowledge-like modules, though here it may be convenient to extend the interpretation to mechanisms by analyzing the functions they implement. The

idea, however, is that each cognitive capacity is a body of rules and principles that fits into a domain of such principles. The domains are distinct from each other at the "core", though perhaps there is peripheral overlap. An extreme version of this thesis, in Fodor (1983), is that some domains are "eccentric", or completely disconnected from other domains, as Chomsky seems to claim about language. Khalidi (2001) emphasizes that this type of structure ensures that a capacity is not "generalizable" beyond its immediate domain.

A virtue of this approach is that domain-specificity is easy to read off a capacity. Using the informational criteria developed above, one can divide up the realms of knowledge into rough domains centered around particular core principles. Domain-specificity means applying to only one domain, while any number of further domains linearly increases the degree of domain-generality. Since the account is scopal, it does not have the problem of defining a "proper" domain, and defining domain-generality as exceeding that bound into an "appropriated" domain. Therefore it avoids giving an account of "proper", or the domain a capacity is meant to handle, as distinct from the domains it can actually handle, some of those being "appropriated". As such, there is no problem of historicizing the concept, or overrunning the methodological convention that domain-specificity is judged on the present (though, actual and counterfactual) properties of the capacity.

The principal difficulty may be with finding eccentric domains, or the perennial problem in the sciences of carving up nature. Absent that, it may be possible to define domains by

the clustering of certain distinctive concepts. For example, the concept of phrasal structure would sit near the center of the linguistic domain.

A second difficulty arises from the difficulty of assessing efficacy for capacities judged domain-specific. The informational relations between capacity and domain do not guarantee that the capacity does anything useful. The idea that specialization implies a function undercuts this goal-free account.

3.4 Recap

Most treatments of domain-specificity have attempted simplified accounts turning on a single criterion. As a general strategy, this underestimates the many possibilities for trivializing the concept. Some background structure needs to be in place to put limits on the various dimensions at risk. A theme of this paper has been that treating cognitive capacities as informational systems, and domains as bodies of information, permits a coherent approach to these risks. Some of the prominent uses of the domain-specificity concept do not adopt this approach explicitly, leaving them open to difficulties. The positive result, however, is the observation that if we begin with informational accounts of domains—using the ideas of *coherence*, *maximality*, and *eccentricity*—and an informational treatment of capacities, we can use *coherence* and actual *relevance* to outline a fairly good framework for the concept of domain-specificity.

Bibliography

Anthony, Louise. 2001. Empty Heads. *Mind and Language*, 16, 2.

Ariew, Andre. 1996. Innateness and Canalization. *Philosophy of Science*, 63 (Proceedings), S19-S27.

Ariew, Andre. 1999. Innateness is canalization. In Hardcastle, 1999.

Ariew, Andre. Forthcoming, 2003. Natural Selection Doesn't Work That Way: Fodor on Adaptationism. *Mind and Language*.

Atkinson, Anthony and Michael Wheeler. Forthcoming, 2002. The Grains of Domains: The evolutionary-psychological case against domain-general cognition. <http://www.wkac.ac.uk/psychology/staff/GrainDomains_M&L_Final.pdf>.

Atran, Scott. 1990. *Cognitive Foundations of Natural History: Towards an Anthropology of Science*. Cambridge and New York: Cambridge University Press.

Atran, Scott. 1994. Core domains versus scientific theories: Evidence from systematics and Itza-Maya folkbiology. In Hirschfeld and Gelman, 1994, 316-340.

Atran, Scott. 1999. Itzaj Maya Folkbiological Taxonomy: Cognitive Universals and Cultural Particulars. In Medin and Atran, 1999, 119-203.

Ayers, M. and D. Garber, eds. 1998. *Cambridge History of Seventeenth Century Philosophy*. Cambridge: Cambridge University Press.

Baron-Cohen, S. 1995. *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Baron-Cohen, Simon, Tager-Flusberg, Helen, Cohen, Donald J. 1993. *Understanding other minds. Perspectives from autism*. Oxford: Oxford University Press.

Baron-Cohen, S., A. Leslie, and U. Frith. 1985. Does the autistic child have a 'theory of mind'? *Cognition*, 21:37-46.

Barkow, J., L. Cosmides and J. Tooby, eds. 1992. *The Adapted Mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Bates, E. 1994. Modularity, domain specificity and the development of language. *Discussions in Neuroscience*, 10:136-149.

Bateson, P. 1991a. Are there principles of behavioural development? In Bateson, 1991b, 19-39.

Bateson, P., ed. 1991b. *The Development and Integration of Behaviour*. Cambridge: Cambridge University Press.

Berko, Jean. 1958. "The Child's Learning of English Morphology," *Word* 14:150-77.

Bever, Thomas G. 1974. The Psychology of Language and Structuralist Investigations of Nativism. In Harman, 1974.

Bickerton, Derrick. 1981. *Roots of Language*. Ann Arbor, Michigan: Karoma.

Bickerton, Derrick. 1983. Creole Languages. *Scientific American*, July.

Block, Ned. Unpublished. The Mind as the Software of the Brain. <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/msb.html>

Bock, K. & W. Levelt. 1994. Language production: Grammatical encoding. In Gernsbacher, 1994.

Bonner, J. T. 1988. *The Evolution of Complexity*. Princeton University Press, Princeton, NJ.

Botterill, G. and Carruthers, P. 1999. *Philosophy of Psychology*. Cambridge: Cambridge University Press.

Boyer, Pascal. 1994. Cognitive constraints on cultural representations: Natural ontologies and religious ideas. In Hirschfeld and Gelman, 1994.

Boyer, Pascal. 2000. Evolution of the modern mind and the origins of culture: religious concepts as a limiting case. In Carruthers and Chamberlain, 2000, 93-112.

Broad, C.D. 1975. *Leibniz: An Introduction*. Cambridge: Cambridge University Press.

Buller, David and Valerie Gray Hardcastle. 2000. Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity. *Brain and Mind*, 1:307-325.

Buller, David, ed. 1999. *Function, Selection and Design*. Albany: SUNY Press.

Butterfield, J. 1986. *Language, Mind and Logic*. Cambridge: Cambridge University Press.

Campbell, J.I.D. 1994. “Locality, modularity and numerical cognition,” *Behavioral and Brain Science*, 17, 1:63-64.

Carey, Susan and Elizabeth Spelke. 1994. "Domain specific knowledge and conceptual change" in Hirschfeld and Gelman, 1994a, 169-200.

Carey, Susan and R. Gelman, eds. 1991. *Epigenesis of mind: Studies in biology and cognition.* Hillsdale, NJ: Erlbaum.

Carey, Susan. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, Susan. 1988. “Conceptual differences between children and adults”. *Mind & Language,* 3:167-181.

Caramazza, Alfonso, Argye Hillis, Elwyn C. Leek and Michele Miozzo. 1994. The organization of lexical knowledge in the brain: Evidence from category- and modality-specific deficits. In Hirschfeld and Gelman, 1994a.

Carruthers, P. and J. Boucher, eds. 1998. *Language and Thought*. Cambridge: Cambridge University Press.

Carruthers, P. and A. Chamberlain, eds. 2000. *Evolution and The Human Mind: Modularity, Language and Meta-Cognition.* Cambridge: Cambridge University Press.

Carruthers, P. and P.K. Smith, eds. 1996. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Carruthers, P. 1996a. *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. 1996b. Simulation and self-knowledge: a defence of theory-theory. In Carruthers and Smith, 1996, 22-38.

Carruthers, P. 1998a. Thinking in language? Evolution and a modularist possibility. In Carruthers and Boucher, 1998, 94-119.

Chalmers, David. Unpublished. “A Computational Foundation for the Study of Cognition”. <http://www.u.arizona.edu/~chalmers/ai-papers.html>

Chater, N. 1994. Modularity, interaction, and connectionist neuropsychology. *Behavioral and Brain Science*, 17, 1:66-67.

Chemero, Anthony. 2000. Anti-Representationalism and the Dynamical Stance. *Philosophy of Science*, 67, December:625-647.

Chomsky, Noam. 1959. Review of B.F. Skinner's *Verbal Behavior.* Reprinted in Fodor and Katz, 1964.

Chomsky, Noam. 1965. *Aspects of a Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, Noam. 1966. *Cartesian Linguistics.* New York: Harper and Row.

Chomsky, Noam. 1975. *Reflections on Language.* New York: Pantheon Books.

Chomsky, Noam. 1980. *Rules and Representations.* New York: Columbia University Press.

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Holland: Foris Publications.

Chomsky, Noam. 1984. *Modular Approaches to the Study of the Mind.* San Diego: State University Press.

Chomsky, Noam. 1986. *Knowledge of Language*. New York: Praeger Publishers.

Chomsky, Noam. 1987. *Language and Problems of Knowledge*: *Managua Lectures.* Cambridge, MA: MIT Press.

Clark, Andy and Josefa Toribio. 1998. *Machine Intelligence: Perspectives on the Computational Model*. New York: Garland.

Coltheart, M. 1999. Modularity and Cognition. *Trends in Cognitive Science*, 3, 3:115-120.

Cosmides, L. and J. Tooby, 1992. Cognitive adaptations for social exchange. In Barkow et al. 1992

Cosmides, L. and J. Tooby. 1987. From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré, 1987.

Cosmides, Leda and John Tooby. 1994. Origins of domain specificity: the evolution of functional organization. In Hirschfeld and Gelman, 1994a, 85-116.

Cowie, Fiona. 1999. *What's Within?* Oxford: Oxford University Press.

Cowie, Fiona. 2000a. Domain-specificity revisited: response to Keil. *A Field Guide to Philosophy of Mind: E-symposium on Fiona Cowie.* <http://host.uniroma3.it/progetti/kant/field/cowiesymp.htm>

Cowie, Fiona. 2000b. Whistlin' Dixie: Reply to Fodor. *A Field Guide to Philosophy of Mind: E-symposium on Fiona Cowie.* <http://host.uniroma3.it/progetti/kant/field/cowiesymp.htm>

Cowie, Fiona. 2001. Cussing in Church: In Defence of What's Within? In *Mind and Language*, 16, 2.

Crain, Stephen and Mineharu Nakayama. 1987. Structure dependence in children's language. *Language,* 62:522-543.

Crain, Stephen. 1991. Language acquisition in the absence of experience," *Brain and Behavioral Sciences,* 14.

Currie, Greg. 1995. Visual imagery and the simulation of vision. *Mind and Language*, 10, 25–44

Davies, M. 1981a: *Meaning, Quantification, Necessity: Themes in Philosophical Logic*. London: Routledge and Kegan Paul.

Davies, M. 1981b. Meaning, structure, and understanding. *Synthese*, 48, 135–61.

Davies, M. 1987. Tacit knowledge and semantic theory: Can a five per cent difference matter? *Mind*, 96, 441–62.

Davies, M. 1989. Tacit knowledge and subdoxastic states. In George, 1989, 131–52. Reprinted in C. Macdonald and G. Macdonald (eds.), 1995.

Davies, M. 1995a. Two notions of implicit rules. In Tomberlin, 1995, 153–83.

Davies, M. 1995b. Consciousness and the varieties of aboutness. In C. Macdonald and G. Macdonald, eds. 1995, 356–92.

Davies, M. 1996. The mental simulation debate. In Peacocke, 1996. Also reprinted in W.G. Lycan (ed.), *Mind and Cognition: An Anthology*. Second Edition, Oxford: Blackwell Publishers, 1998.

Davies, M. 2000. Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society*, 1, 000–00.

Davies, M. and T. Stone, eds. 1995a: *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell Publishers.

Davies, M. and T. Stone, eds. 1995b: *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

Davies, M. and T. Stone. 1995c. Introduction. In Davies and Stone, 1995a.

Davies, M. and T. Stone. 1998. Folk psychology and mental simulation. In O'Hear, 1998, 53-82.

Davies, M. and T. Stone. Unpublished (2001). Mental simulation, tacit theory, and the threat of collapse. Presented at NYU, January 30, 2001. <http://www.nyu.edu/gsas/dept/philo/courses/content/papers/davies.pdf>

Deacon, Terrence W. 1997. *The Symbolic Species: the coevolution of language and human brain.* London: Penguin.

Dennett, Daniel. 1987. Making sense of ourselves. In *The Intentional Stance*. Cambridge, MA: MIT Press.

Descartes, R. 1955. *The Philosophical Works of Descartes,* trans. E. S. Haldane and G.R.T. Ross, New York: Dover Publications. Cited in Chomsky, 1966.

Descartes, Renee. 1985. *The Philosophical Writings of Descartes*, Vol. 1, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge UP. Cited in Cowie, 1999.

Donald, Merlin. 1993. *Origins of the Modern Mind*. Cambridge, MA: Harvard University Press.

Donald, Merlin. 2001. *A Mind So Rare: The Evolution of Human Consciousness*. New York: W. W. Norton.

Dretske, Fred. 1985. Machines and the Mental. Presidential Address to the 83rd Western Division of the American Philosophical Association. Reprinted in Clark and Toribio, 1998, 267-277.

Duffy, S. A., R. K. Morris & K. Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27, 429-446.

Dupré, John, ed. 1987a. *The Latest on the Best*. Cambridge, MA: Bradford.

Dupré, John. 1987b. "Human Kinds" in Dupré, 1987a, 327-348.

Easton, Patricia, ed. 1987. *Logic and the Workings of the Mind: The Logic of Ideas and Faculty Psychology in Early Modern Philosophy.* Atascadero, Calif.: Ridgeview Publishing Co.

Elman, Jeffrey L., et al. 1996. *Rethinking Innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Erneling, C.E. and D. M. Johnson, eds. Forthcoming. *Mind as a Scientific Object: Between Brain and Culture.* Oxford, Oxford University Press.

Fodor, Jerry and Jerrold Katz, 1964a. Introduction. In Fodor and Katz, 1964b.

Fodor, Jerry and Jerrold Katz, eds. 1964b. *The Structure of Language*. Englewood Cliffs, NJ: Prentice Hall.

Fodor, Jerry. 1975. *The Language of Thought.* New York: Crowell.

Fodor, Jerry. 1981a. *Representations*. Cambridge, MA: MIT Press.

Fodor, Jerry. 1981b. Introduction: Something on the State of the Art. In Fodor, 1981a.

Fodor, Jerry. 1981c. The Present Status of the Innateness Controversy. In Fodor 1981a.

Fodor, Jerry. 1983a. *The Modularity of Mind.* Cambridge, MA: MIT Press, Bradford.

Fodor, Jerry. 1985a. Précis of Modularity of Mind. In *Behavioral and Brain Sciences*. Reprinted in Fodor, 1990.

Fodor, Jerry. 1985b. Fodor's Guide to Mental Representation. In *Mind*, Spring: 66-97. Reprinted in Fodor, 1990.

Fodor, Jerry. 1985c. Why should the mind be modular? In George, 1989.

Fodor, Jerry. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays.* Cambridge, MA: MIT Press.

Fodor, Jerry. 1999. Doing Without What's Within: A Critique of Fiona Cowie's *What's Within? A Field Guide to Philosophy of Mind.* <http://host.uniroma3.it/progetti/kant/field/cowiesymp.htm> Reprinted in *Mind*, 2000.

Flourens, Pierre. 1851. *Examen de la Phrénologie*. Paris : Hachette.

Frazier, L. 1987. Theories of sentence processing. In Garfield, 1987.

Friedman, Michael. 2001. *Dynamics of Reason: The 1999 Kant Lectures at Stanford University*. Palo Alto: CSLI Publications. (Pagination from Draft)

Gall, F. J. And G. Spurzheim. 1809. *Recherches sur les système nerveux en général, et sur celui du cerveau en particulier.* Schoell : Paris.

Gall, F. J. and G. Spurzheim. 1811. *Des dispositions innées de l'âme et de l'esprit.* Schoell : Paris.

Gall, F. J. 1825. *Sur l'origine des qualités morales et des facultés intellectuelles de l'homme.* Paris : JB. Baillière.

Garfield, J. 1987. *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, MA: MIT Press.

Gazzaniga, M., ed. 2000. *The New Cognitive Neurosciences* . Cambridge, MA: MIT Press.

Gelman, S. A., and Hirschfeld, L. A. 1999. How Biological is Essentialism? In Medin and Atran, 1999, 403-446.

George, A. ed. 1989. *Reflections on Chomsky*. Oxford: Blackwell.

Gerken, L. 1994. Child phonology: Past research, present questions, future directions. In Gernsbacher, 1994.

Gernsbacher, M.A., ed., 1994. *Handbook of psycholinguistics*. San Diego: Academic Press.

Gerrans, Philip. 2002. "Modularity Reconsidered." *Language and Communication* 22, 259-268.

Giere, Ronald D. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gigerenzer, Gerd, Peter Todd, and ABC Research Group. 1999. *Simple Heuristics that Make Us Smart.* Oxford: Oxford University Press.

Godfrey-Smith, Peter. 1994. *Complexity and the Function of Mind in Nature.* Cambridge UP: Cambridge.

Godfrey-Smith, Peter. Forthcoming. Between Baldwin Skepticism and Baldwin Boosterism. In Weber and Depew, Forthcoming. <http://www-philosophy.stanford.edu/fss/papers/baldwin.pdf>

Godfrey-Smith, Peter. 1994. A Modern History Theory of Functions. *Noûs*, 28:344-362.

Godfrey-Smith, Peter. 1994. A Continuum of Semantic Optimism. In Stich and Warfield, 1994, 259-277.

Gold, E. 1967. "Language identification in the limit." *Information and Control*, 10, 47-474.

Gold, Ian and Daniel Stoljar. 1999. A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Sciences*, 22, 5.

Goldman, Alvin. 1989. Interpretation Psychologized. *Mind and Language*, 4:165-82.

Goldman, Alvin. 1992a. In Defense of Simulation Theory. *Mind and Language*, 7:104-119.

Goldman, Alvin. 1992b. Empathy, Mind and Morals. *Proceedings and Addresses of the American Philosophical Association*, 66:17-41. Reprinted in Davies and Stone, 1995b, 185-208.

Goldman, Alvin. 1993. The Psychology of Folk Psychology. *Behavioural and Brain Sciences*, 16.

Gopnik, A. and A. Meltzoff. 1997. *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gopnik, Alison & Meltzoff, Andrew N. 1998. Theories vs. modules: to the max and beyond. A reply to Poulin-Dubois and to Stich and Nichols. *Mind and Language*, 13: 450-456.

Gopnik, A. and Henry M. Wellman. 1994. The theory theory. In Hirschfeld and Gelman, 1994a.

Gopnik, A. and Henry M. Wellman. 1995. Why the child's theory of mind really *is* a theory. *Mind and Language*, 7, 145–71. Reprinted in Davies and Stone, 1995a.

Gordon, R. 1986. Folk Psychology as Simulation. *Mind and Language*,1, 158-71.

Gordon, Robert. 1992. The Simulation Theory: Objections and Misconceptions. *Mind and Language*. Vol. 7, No. 1-2.

Gordon, Robert. 1992. Reply to Stich and Nichols. *Mind and Language*. 7, 1-2.

Gordon, R. 1995. The Simulation Theory: Objections and Misconceptions. In Davies and Stone, 1995a, 100-122.

Gordon, R. 1996. 'Radical' Simulation. In Carruthers and Smith, 1996, 11-21.

Gould, S. J. and E. S. Vrba. 1982. Exaptation: A missing term in the science of form. *Paleobiology* 8, 1:4-15.

Grantham, Todd and Shaun Nichols. 1999. Evolutionary Psychology: Ultimate Explanations and Panglossian Predictions. In Hardcastle, 1999, 47-66.

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

Griffiths, 2001. See pdf file. Submitted to *The Monist*. Special Issue edited by Kim Sterelny.

Hacking, Ian. Domain-Specificity.

Hardcastle, Valerie Gray. 1999. *Where Pyschology Meets Biology: Conjectures, Connections, Constraints*. Cambridge, MA: MIT Press.

Harman, Gilbert. 1974. *On Noam Chomsky*. Garden City, New York: Anchor Books.

Harris, P. 1994. Thinking by children and scientists: False analogies and neglected similarities. In Hirschfeld and Gelman, 1994a.

Harris, P. 1995. From Simulation to Folk Psychology: The Case for development. In Davies and Stone, 1995a, 207-231.

Hatfield, Gary and Stephen Kosslyn. 1984. Representation without Symbol Systems. *Social Research*, 51: 1019-1045.

Hatfield, Gary. 1997. The Workings of the Intellect: Mind and Psychology. In Easton, 1997, 21-45.

Hatfield, Gary. 1998. The Cognitive Faculties. In Ayers and Garber, 1998, 953-1002.

Haugeland, John. 1985a. Semantic Engines: An Introduction to Mind Design. In Haugeland, 1985b. Reprinted in Clark and Toribio, 1998.

Haugeland, John. 1985b. *Mind Design: Philosophy, Psychology, Artificial Intelligence,* Cambridge: MIT Press.

Heal, J. 1986. Replication and functionalism. In Butterfield, 1986, 135–50. Reprinted in Davies and Stone, 1995a.

Heal, J. 1994. Simulation vs. theory theory: What is at issue? In Peacocke, 1994, 129–44.

Heal, J. 1995a. Replication and Functionalism. In Davies and Stone, 1995a, 45-59.

Heal, J. 1995b. How to Think About Thinking. In Davies and Stone, 1995b, 33-52.

Heal, J. 1996. Simulation and cognitive penetrability. *Mind and Language*, 11, 44–67.

Heal, J. 1998a. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language*, 14, 477–98.

Heal, J. 1998b. Understanding other minds from the inside. In O'Hear, 1998, 83–99.

Heal, J. 2000: Other minds, rationality and analogy. *Proceedings of the Aristotelian Society, Supplementary Volume,* 74, 1–19.

Hirschfeld, Lawrence A. Forthcoming. Human kinds.

Hirschfeld, Lawrence A. 1994. Is the acquisition of social categories based on domain-specific competence or on knowledge transfer? In Hirschfeld and Gelman, 1994a, 201-234.

Hirschfeld, Lawrence A. and Susan A. Gelman, eds. 1994a. *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York: Cambridge University Press.

Hirschfeld, Lawrence A. and Susan A. Gelman. 1994b. Towards a topography of mind: An introduction to domain specificity. In Hirschfeld and Gelman, 1994a, 3-36.

Hornstein, N. and D. Lightfoot, eds. 1981. *Explanation in Linguistics: The Logical Problem of Language Acquisition.* Longman, London.

Hornstein, N. and D. Lightfoot. 1981. Introduction. In Hornstein and Lightfoot, 1981.

Israel, David and John Perry. 1998. What is Information? In Clark and Toribio, 1998, 221-239.

Jackson, F.C. 1999. All That Can Be at Issue in the Theory-Theory Simulation Debate. *Philosophical Papers*, 28, 77–96.

Jusczyk, P.W. & R.N. Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.

Kahneman, D. and Tversky, A. 1982. The Simulation Heuristic. In Kahneman, Slovic and Tversky, 1982.

Kahneman, D., P. Slovic and A. Tversky, eds. 1982. *Judgment Under Uncertainty*. Cambridge: Cambridge University Press.

Karmiloff-Smith, Annette. 1992. *Beyond Modularity*. Cambridge, MA: MIT Press.

Kaye, L. 1999. Empiricist Sour Grapes. *A Field Guide to Philosophy of Mind.* <http://host.uniroma3.it/progetti/kant/field/cowiesymp.htm>.

Keil, F. C. 1989. *Concepts, Kinds and Cognitive Development*. Cambridge, MA.: Bradford Books/MIT Press.

Keil, F. C. 1994. The birth and nurturance of concepts by domains: The origins of concepts of living things. In Hirschfeld and Gelman, 1994a.

Keil, F. C. 1999. Nativism. *MIT Encyclopedia of the Cognitive Sciences.* Cambridge, MA: MIT Press.

Keil, F. C. 1999. Nurturing Nativism. *A Field Guide to Philosophy of Mind.* <http://host.uniroma3.it/progetti/kant/field/cowiesymp.htm>

Khalidi, M. A. 2001. Innateness and Domain-Specificity. *Philosophical Studies*, 105, 191-210.

Khalidi, M. A. 2002. Nature and Nurture in Cognition. *British Journal for the Philosophy of Science*, 53, 2:251-272.

Kirsh, David. 1998. When Is Information Explicitly Represented? In Clark and Toribio, 1998, 240-265.

Kitcher, Patricia. 1988. Marr's computational theory of vision. *Philosophy of Science*, 55:1-24.

Kitcher, Phillip. 1977. The Nativist's Dilemma. *Philosophical Quarterly*, 28: 1-16.

Leekam, S. & Perner, J. 1991. Does the autistic child have a metarepresentational deficit? *Cognition*, 40:203-218.

Van Lehn, K., ed. 1991. *Architectures for Intelligence*. Hillsdale: Lawrence Erlbaum Associates Inc.

Leibniz, G.W. 1981 (1765). *New Essays on Human Understanding.* Peter Remnant and Jonathan Bennet, trans. New York: Cambridge.

Lennenberg, Eric. 1964. The Capacity of Language Acquisition. In Fodor and Katz, 1964.

Leslie, Alan M. 1987. Pretense and representation: the origins of 'theory of mind'. *Psychological Review*, 94:412-426.

Leslie, A. and Frith, U. 1988. Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology*, 6:315-24.

Leslie, Alan M. 1994a. Pretending and believing: issues in the theory of ToMM. *Cognition*, 50, 211-238.

Leslie, Alan M. 1994b. ToMM, ToBy, and Agency: Core architecture and domain specificity. In Hirschfeld and Gelman, 1994a, 119-148.

Leslie, A. 2000. 'Theory of mind' as a mechanism of selective attention. In Gazzaniga, 2000, 1235-1247.

Lorenz, Konrad. 1963 *On Aggression.* New York: Harcourt, Brace & World.

Macdonald, C. and G. Macdonald, eds. 1995. *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford: Blackwell Publishers.

Markman, E. M. 1990. Constraints children place on word meaning. *Cognitive Science*, 14, 57-77.

Marr, D. 1982. *Vision*. New York: H. Freeman and Co.

Matthews, R. 2001. Cowie's Anti-Nativism. *Mind and Language*, 16, 2: 215-230.

McCloskey, M. 1992. Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia," *Cognition*, 44, 107-57.

Medin, D.L. and S. Atran, eds. 1999. *Folkbiology*. Cambridge, MA: MIT Press.

Miller, Geoffrey. 2001. *The Mating Mind*. New York: Doubleday.

Mithen, Steven. 1996. *The Prehistory of the Mind*. New York: Routledge

Morán, F., A. Morán, J.J. Merelo, and P. Chacón, eds. 1995. *Advances in Artificial Life*. Springer Verlag, Berlin.

Murphy, D. and Stich, S. 2000. Darwin in the madhouse: evolutionary psychology and the classification of mental disorders. In Carruthers and Chamberlain, 2000.

Newell, Alan and Herb Simon. 1972. *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nichols, Shaun, Stephen Stich, Alan Leslie, and David Klein. 1996. Varieties of Off-line Simulation. In Carruthers and Smith, 1996, 39-74.

O'Hear, A., ed. 1998. *Contemporary Issues in Philosophy of Mind*. Cambridge: Cambridge University Press.

Orr, H. 2003. Darwinian Storytelling. *New York Review of Books*, February 27.

Osherson, D. and H. Lasnik eds. 1990. *Language: An invitation to cognitive science*, Vol. 1. Cambridge, MA: MIT Press.

Oyama, S. 1990. The idea of innateness: effects on language and communication research. *Developmental Psychobiology, 2*3, 7:741-747.

Peacocke, C. 1986. Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language*, 1:101–23.

Peacocke, C. 1989. When is a grammar psychologically real? In George, 1989, 111–30.

Peacocke, C., ed. 1996. *Objectivity, Simulation and the Unity of Consciousness: Current Issues in the Philosophy of Mind* (Proceedings of the British Academy vol. 83). Oxford: Oxford University Press.

Peacocke, C. 1996a. Introduction: The issues and their further development. In Peacocke, 1996*,* xi–xxvi.

Perner, J. 1991. *Understanding the representational mind.* Cambridge, MA: Bradford Books/ MIT-Press.

Perner, Josef. 1993. The theory of mind deficit in autism: rethinking the metarepresentation theory. In Baron-Cohen et al., 1993, 112-137.

Perner, J. 1994. The necessity and impossibility of simulation. In Peacocke, 1994.

Perner, J. 1996. Simulation as Explicitation of Predication-Implicit Knowledge about the Mind: Arguments for a Simulation-Theory Mix. In Carruthers and Smith, 1996, 90-104.

Pinker, S. and Bloom, P. 1990. Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 4:707-784.

Pinker, Steven. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Pinker, S. 1990. Language Acquisition. In Osherson and Lasnik, 1990.

Pinker, S. 1994. *The Language Instinct*. New York: Harper Collins.

Pinker, Steven. 1998. *How the Mind Works.* New York: Norton.

Pinker, S. 1999. *Words and Rules.* New York: Weidenfeld & Nicholson.

Pinker, S. 2000. *Words and Rules*. New York: Basic Books.

Plotkin, H. 1997. Evolution in Mind. London: Alan Lane.

Posner, Michael A., ed. 1989. *Foundations of Cognitive Science*. Cambridge, MA: Bradford.

Premack, D., & Woodruff, G. 1978. Does the Chimpanzee Have a 'Theory of Mind'? *Behavioural and Brain Sciences*, 4, 515-526.

Pylyshyn, Z. 1980. Computation and Cognition: Issues in the Foundations of Cognitive Science. *Behavioral and Brain Sciences,* 3:111-132.

Pylyshyn, Z. 1984. *Computation and Cognition.* Cambridge, MA: MIT Press.

Pylyshyn, Z.W. 1991. The role of cognitive architecture in theories of cognition. In Van Lehn, 1991.

Quartz, Steve and Terence Sejnowski. 1997. The Neural Basis of Cognitive Development: A Constructivist Manifesto. *Behavioral and Brain Sciences*, 20, 4: 537-596.

Quine, E.g. Methodological Reflections on Current Linguistic Theory. In Harman, 1974.

Ryle, G. 1949. *The concept of mind.* London: Hutchinson & Company.

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* London: MIT Press.

Samuels, R. 1998. Evolutionary psychology and the massive modularity hypothesis". *The British Journal for the Philosophy Science* 49: 575-602.

Samuels, R. 2000. Massively modular minds: evolutionary psychology and cognitive architecture. In Carruthers and Chamberlain, 2000.

Samuels R. 2002. Nativism in Cognitive Science. *Mind and Language*, 17, 3, 233-265.

Searle, John. 1974. Chomsky's Revolution in Linguistics. In Harman, 1974.

Searle, John. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, , 417-457.

Searle, J. R. 1984. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Searle, John, 1990a. Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association,* 64: 21-37

Searle, John, 1990b. Is the Brain's Mind a Computer Program? *Scientific American,* 262, 1, 20-25.

Searle, John, 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Segal, Gabriel. 1996. The Modularity of Theory of Mind. In Carruthers and Smith, 1996.

Shallice, T. 1988. *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Shepard, Roger N. 1987. Evolution of a Mesh between Principles of the Mind and Regularities of the World. In Dupré, 1987a, 251-276.

Skinner, B. F. 1957. *Verbal Behavior*. New York: Appleton-Century-Crofts.

Singleton, J. & Newport, E. 1994. When learners surpass their models: The acquisition of American Sign Language from impoverished input. Manuscript, University of Rochester.

Smith, Neil and Ianthi-Maria Tsimpli. 1995. *The Mind of a Savant*: *Language Learning and Modularity*. Oxford: Basil Blackwell.

Sober, E. 1980. Evolution, Population Thinking and Essentialism. *Philosophy of Science, 47,* 3, 350-383.

Spelke, E.S. 1990. Principles of object perception. *Cognitive Science*, 14, 29-56.

Spelke, E.S. 1991. Physical knowledge in infancy: Reflections on Piaget's theory. In Carey and Gelman, 1991.

Sperber, D. & Wilson, D. 1995. *Relevance: Communication and cognition*, Second Edition. Oxford: Blackwell.

Sperber, D. & Wilson, D. 1996. Fodor's frame problem and relevance theory. *Behavioral and Brain Sciences*, 19:3, 530-532

Sperber, Dan. 1994. The modularity of thought and the epidemiology of representations. In Hirschfeld and Gelman, 1994a, 39-67.

Sperber, Dan. 1997. Citation from Botterill and Carruthers, 1999. Talk on cognition and logic module.

Sperber, Dan. Unpublished. In Defense of Massive Modularity. <http://www.dan.sperber.com>

Sterelny, Kim. Forthcoming. *Cognition in a Hostile World.*

Sterelny, Kim and Paul Griffiths. 1999. *Sex and Death*. Chicago: University of Chicago Press.

Sterelny, Kim. 1991. *The Representational Theory of Mind.* London: Blackwell Publishers.

Stich, Stephen and Ted A. Warfield. 1994. Introduction. In Stich and Warfield, 1994, 1-8.

Stich, Stephen and Ted A. Warfield, eds. 1994. *Mental Representations: A Reader*. Cambridge, MA: Basil Blackwell.

Stich, Steven, ed. 1975. *Innate Ideas*. Berkeley: University of California Press.

Stich, Steven. 1975a. The idea of innateness. In Stich, 1975, 1-24.

Stich, Stephen. 1985. Could Man Be an Irrational Animal? Some Notes on the Epistemology of Rationality? *Synthese*, 64, 1:115-135.

Stich, Stephen. 1994. What Is a Theory of Mental Representation? In Stich and Warfield, 1994, 347-364.

Stone, T. and Martin Davies. 1996. The Mental Simulation Debate: A Progress Report. In Carruthers and Smith, 1996.

Symons, D. 1992. On the use and misuse of Darwinism in the study of human behavior. In Barkow et al., 1992, 137-159.

Tomberlin, J.E., ed., 1995. *Philosophical Perspectives, 9: AI, Connectionism, and Philosophical Psychology*. Atascadero, CA: Ridgeview Publishing Company.

Turing, A.M. 1950. Computing Machinery and Intelligence. *Mind*, LIX, 236:433-460. Reprinted in Clark and Toribio, 1998.

van Gelder, Tim. 1994. Playing Flourens to Fodor's Gall. In Farah, 1994.

van Gelder, T. 1995. What might cognition be, if not computation? *The Journal of Philosophy* XCI, 345-381.

van Gelder, T. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 1-14.

van Gelder, T. and R. Port, eds. 1995a. *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

van Gelder, T. J., & Port, R. 1995. It's About Time: An Overview of the Dynamical Approach to Cognition. In Port and van Gelder, 1995, 1-43.

Wagner, G. P. 1996. Homologues, natural kinds and the evolution of modularity. Am. Zool. 36: 36-43.

Wagner, G.P. 1995. Adaptation and the modular design of organisms. In Morán et al. 1995.

Wagner, G.P. and L. Altenberg 1996. Complex adaptations and the evolution of evolvability. *Evolution* 50:967-976.

Weber, B. and D. Depew, eds. Forthcoming. *Learning and Evolution: The Baldwin Effect Reconsidered*. Cambridge: MIT Press.

Wernicke, Carl. 1874. *Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis*. Breslau: Cohen und Weigert.

Whitney, P. 1998. *The Psychology of Language*. Boston: Houghton Mifflin.

Wilson, David Sloan. 2002. *Darwin's Cathedral*. Chicago: University of Chicago Press.

Wimmer, H. and J. Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103-128.

Wimsatt, W. C. 1999. Generativity, Entrenchment, Evolution, and Innateness: Philosophy, Evolutionary Biology, and Conceptual Foundations of Science. In Hardcastle, 1999, 139-179.

Wright, Larry. 1973. Functions. *Philosophical Review*, 82, 2:139-168.

Zawidzki, T. and Bechtel, W. (in press). "Gall's Legacy Revisited Decomposition and Localization in Cognitive Neuroscience." In Erneling and Johnson, Forthcoming. <http://mechanism.ucsd.edu/~bill/research/gall.html>