OXFORD SERIES ON COGNITIVE MODELS AND ARCHITECTURES

Principles of Synthetic Intelligence

Psi: An Architecture of Motivated Cognition

Joscha Bach

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016
www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Bach, Joscha.
Principles of synthetic intelligence : Psi, an architecture of
motivated cognition / by Joscha Bach.
p. cm.—(Oxford series on cognitive models and architectures ; 4)
Includes bibliographical references and index.
ISBN 978-0-19-537067-6 (cloth : alk. paper)
1. Cognition. 2. Artificial intelligence. I. Title.
BF311.B23 2009
153—dc22

2008043048

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

Contents

This page intentionally left blank

Principles of Synthetic Intelligence

This page intentionally left blank

1

Machines to explain the mind

> *I propose to consider the question, "Can machines think?"*
>
> *This should begin with definitions of the meaning of the terms "machine" and "think."*
>
> Alan M. Turing (1950)

This book is an attempt to explain cognition—thought, perception, emotion, experience—in terms of a machine: that is, using a *cognitive architecture*. While this approach has gained acceptance in the cognitive sciences, it seemingly runs against many of our intuitions on how to understand the mind. As Gottfried Wilhelm Leibniz put it:

> Perception, and what depends on it, is inexplicable in a mechanical way, that is, using figures and motions. Suppose there would be a machine, so arranged as to bring forth thoughts, experiences and perceptions; it would then certainly be possible to imagine it to be proportionally enlarged, in such a way as to allow entering it, like into a mill. This presupposed, one will not find anything upon its examination besides individual parts, pushing each other—and never anything by which a perception could be explained. (Leibniz 1714 [translation by the author])

Cognitive architectures are indeed Leibnizean Mills: machines that are designed to bring forth the feats of cognition, and built to allow us to enter them, to examine them, and to watch their individual parts in motion, pushing and pulling at each other, and thereby explaining how a mind works.

Our particular cognitive architecture is based with a formal theory of human psychology, the *Psi theory*, which will be detailed in the following chapters. This theory has been turned into a computational model, called *MicroPsi*, which has been partially implemented as a computer program. The machine—the computer program and its formal specification—is subject to continuing research, while the Psi theory acts as its blueprint. The lessons that are learned from the workings and failures of the machine do, in turn, lead to improvements in the theory.

But before we discuss theory and implementation, let us note that computational models of the mind are still subject of philosophical and methodological controversies; just as in Leibniz' times, many philosophers and psychologists hotly disagree with the idea of interpreting cognition as the workings of a (computational) mill. Thus, it will be worthwhile to have a look at our main theme—cognition—first, and reflect on the methodological and some of the philosophical foundations of cognitive architectures.

When Leibniz tried to sketch the supposed activity of his mill, he used several related terms (perception, experience, and thought) to hint at what we now call *cognition*. Even today, cognition is not a strictly defined and concisely circumferenced subject. In fact, different areas in the cognitive sciences tend to understand it in quite different ways. In computer science, for instance, the terms "cognitive systems" and specifically "cognitive robotics" (Lespérance, Levesque et al. 1994) often refer loosely to situated, sometimes behavior-based agent architectures, or to the integration of sensory information with knowledge. In philosophy, cognition usually relates to *intentional* phenomena, which in functionalist terms are interpreted as mental content and the processes that are involved with its manipulation. The position that intentional phenomena can be understood as mental representations and operations performed upon them is by no means shared by all of contemporary and most traditional philosophy; often it is upheld that intentionality may not possibly be *naturalized* (which usually means *reduced to brain functions*). However, the concept that intentional states can be explained using a representational theory of the mind is relatively widespread and in some sense the foundation of most of cognitive science.

In psychology, cognition typically refers to a certain class of mental phenomena—sometimes involving all mental processes, sometimes

limited to "higher functions" above the motivational and emotional level, but often including these. Cognitive psychology acknowledges that the mind is characterized by internal states and makes these an object of investigation, and thus tends to be somewhat in opposition to behaviorist stances. Neuropsychology sometimes focuses on cognitive processing, and a substantial part of contemporary cognitive science deals with the examination of the biological processes and information processing of the brain and central nervous system. On the other hand, some psychologists argue that the neurobiological phenomena themselves take place on a functional level different from cognition (Mausfeld, 2003), and that although cognition is facilitated by brain processes and neurobiological correlates to mental (cognitive) processes have been identified, this relationship is spurious and should not mislead research into focusing on the wrong level of description. In this view, the relationship between cognition and neurobiological processes might be similar to the one between a car engine and locomotion. Of course, a car's locomotion is facilitated mainly by its engine, but the understanding of the engine does not aid much in finding out where the car goes. To understand the locomotion of the car, the integration of its parts, the intentions of the driver and even the terrain might be more crucial than the exact mode of operation of the engine. We will briefly revisit this discussion in the next section.

Traditionally, psychology tended to exclude emotion and motivation from the realm of cognition and even saw these as being in opposition. This distinction is now seen as largely artificial, and much research in cognitive psychology is devoted to these areas, as well as to higher level cognition (self-monitoring and evaluation, *meta-cognition*). Yet, the distinction is often still reflected on the terminological level, when reference is made to "cognitive and motivational processes" to distinguish, for instance, the propositional reasoning from action control.

Often it is argued that the cognitive processes of an organism do not only span brain and body, but also the environment—to understand cognition is to understand the interplay of all three. There are several reasons for this: for one thing, because cognition might be seen as a continuum from low-level physical skills to more abstract mental faculties (van Gelder & Port, 1995, p. viii–ix): Just as the motion of a limb might not be properly understood without looking at the nature of the environment of the organism, cognitive processes derive their semantics largely

from environmental interaction. Furthermore, the cognitive processes are not entirely housed within the substrate of the organism's nervous system, but, in part, literally in the interaction context with its habitat. While sometimes relevant aspects of the environment may be modeled within the organism (in the form of a neural "simulator"), these representations will tend to be incomplete and just sufficient for interaction, so parts of cognition will not work without the proper environmental functionality (Clark & Grush, 1999). It has also been argued that the acquired representations *within* the organism should be seen less as a part of the organism than of the environment to which it adapts (Simon 1981, p. 53). And finally, an organism might use tools that are specifically designed to interact with its cognitive core functionality, thus a part of the environment might become part of a mind.[3]

1.1 From psychology to computational modeling

As we see, it is difficult to put a fence around cognition. Why is the notion of cognition so immensely heterogeneous?—I believe this is because the term intends to capture the notion of mental activity, of what the mind does and how it gets it done. Because there is no narrow, concise understanding of what constitutes mental activity and what is part of mental processes, much less what has to be taken into regard to understand them, cognition, the cognitive sciences and the related notions span a wide and convoluted terrain. It might come as a surprise that most of this terrain now lies outside psychology, the science that originally subscribed to studying the mind. This methodological discrepancy can only be understood in the context of the recent history of psychology.

3 See, for instance, Clark (2002): "The sailor armed with hooey and alidade can achieve feats of navigation that would baffle the naked brain (...). And—perhaps more importantly for this discussion—the way such tools work is by affording the kinds of inner reasoning and outer manipulation that fit our brains, bodies and evolutionary heritage. Our visual acuity and pattern-matching skills, for example, far outweigh our capacities to perform sequences of complex arithmetical operations. The slide rule is a tool which transforms the latter (intractable) kind of task into a more homely one of visual cognition. Tools can thus reduce intractable kinds of problems to ones we already know how to solve. A big question about tools, of course, is how did they get here? If tools are tricks for pressing increased functionality out of biologically basic strategies, what kinds of minds can make the tools that make new kinds of minds?"

Psychology, which originally had its roots as a natural science in the psychophysics of Fechner and Helmholtz, became an independent discipline when Helmholtz' pupil Wilhelm Wundt founded his experimental laboratory at the University of Leipzig in 1874 (Boring, 1929). The understanding of psychology as an experimental science was later challenged, especially by the psychoanalytic movement, starting in the 1890s, and because of the speculative nature of the psychoanalytic assumptions, psychology came under heavy fire from positivists and empiricists in the first half of the twentieth century (see Gellner 1985, Grünbaum 1984). The pendulum swung backwards so violently that the psychological mainstream turned away from structuralism and confined itself to the study of directly observable behavior. Behaviorism, as proposed by John B. Watson (1913) became very influential, and in the form of *radical behaviorism* (Skinner, 1938) not only neglected the nature of mental entities as on object of inquiry, but denied their existence altogether. At the same time, this tendency to deny the notion of mental states any scientific merit was supported by the advent of ordinary language philosophy (Wittgenstein, 1953, see also Ryle, 1949). Obviously, the negligence of internal states of the mind makes it difficult to form conclusive theories of cognition, especially with respect to imagination, language (Chomsky, 1959) and consciousness, so radical behaviorism eventually lost its foothold. Yet, *methodological* behaviorism is still prevalent, and most contemporary psychology deals with experiments of quantitative nature (Kuhl, 2001). Unlike physics, where previously unknown entities and mechanisms involving these entities are routinely postulated whenever warranted by the need to explain empirical facts, and then evidence is sought in favor of or against these entities and mechanisms, psychology shuns the introduction of experimentally ungrounded, but technically justified concepts. Thus, even cognitive psychology shows reluctance when it comes to building unified theories of mental processes. While Piaget's work (especially Piaget, 1954) might be one of the notable exceptions that prove the rule, psychology as a field has a preference for small, easily testable microtheories (Anderson, 1993, p. 69).

Psychology tends to diverge along the lines of the individual modeled fields into areas like developmental psychology, motivational psychology, linguistic development, personality theories and so on. Not that these disciplines would be invalidated by their restricted approach! Indeed, much of their credibility is even *due to* their focus on an area that allows a homogenous methodology and thus, the growth and establishment

of scientific routines, communities, and rules of advancement. But this strictness comes at a price: the individual fields tend to diverge, not just in the content that they capture, but also in the ways they produce and compare results. Thus, it not only becomes difficult to bridge the terminological gaps and methodological differences in order to gain an integrative understanding of an individual phenomenon—the results from different disciplines might completely resist attempts at translation beyond a shallow and superficial level.

It is not surprising that influences that lead to the study of genuinely mental entities and structures within psychology came from different fields of science: from information sciences and cybernetics, and from formal linguistics. They fostered an understanding that mental activity amounts to information processing, and that information processing can be modeled as a complex function—an algorithm—working over states that encode representations. In my view, the most important contribution of the information sciences to psychology was the extension of philosophical constructivism into functionalism and the resulting methodological implications.

Functionalist constructivism is based on the epistemological position of philosophical constructivism (see, for instance, von Foerster & von Glasersfeld, 1999) that all our knowledge about the world is based on what is given at our systemic interface. At this interface, we do not receive a description of an environment, but features, certain patterns over which we construct possible orderings. These orderings are functional relationships, systems of categories, feature spaces, objects, states, state transitions, and so on. We do not really *recognize* the given objects of our environment; we *construct* them over the regularities in the information that presents itself at the systemic interface of our cognitive system.

For example: if we take a glance out of the window on a cloudless day, we do not simply *perceive* the sun as given by nature, rather, we identify something we take as a certain luminance and gestalt in what we take to be a certain direction, relatively to what we take to be a point in time. A certain direction is understood as something we take as a characteristic body alignment to something we take as a certain place and which makes a certain set of information accessible that we take to be a certain field of view. In such a way, we may decompose all our notions into the functional features that are the foundation of their construction. Thus, all our notions are just attempts at ordering patterns: we take sets of

features, classify them according to mechanisms that are innate within our interpretational system and relate them to each other. This is how we construct our reality.

To perceive means on one hand to find order over patterns; these orderings are what we call *objects*. On the other hand, it amounts to the identification of these objects by their related patterns—this is intuitively described as the recognition of an object by its features, just as if we would observe the objects themselves instead of constructing them.

An opponent of this view (arguing, for instance, from an essentialist or realist perspective) might suggest that we intuitively do have access to physical objects in the world; but this argument may be tackled using a simple thought experiment: if someone would remove one of the objects of our world and just continue to send the related patterns to our systemic interface (for instance, to our retina) that correspond to the continued existence of the object and its interaction to what we conceptualize as other physical objects, we would still infer the same properties, and no difference could be evident. If, for instance, all electrons in the world would be replaced by entities that behave in just the same way, batteries would continue to supply electrical energy, atoms would not collapse and so on: no difference could ever become evident.[4] Now imagine the removal of the complete environment. Instead, we (the observers) are directly connected (for instance, by our sensory nerves) to an intricate pattern generator that is capable of producing the same inputs (i.e., the same patterns and regularities) as the environment before—we would still conceptualize und recognize the same objects, the same world as we did in the hypothetical world of "real" objects. There can be no difference, because everything that is given is the set of regularities (re-occurrence and seeming dependencies between the patterns).[5]

4 A similar example is supplied by Hilary Putnam (1975): Individuals in a hypothetical twin-world to earth on which all water has been replaced by a chemical compound XYZ with identical properties would arrive at the same observations and conceptualizations. Thus, the content of a concept that is encoded in a mental state refers to the functional role of the codified object.

5 This should be immediately clear to anyone who is familiar with controlling a robot: for the control program of the robot, the environment will present itself as vectors of data, attributable to sensory modalities by the different input channels. For all practical purposes, the world beyond the sensors is a pattern generator; nothing more, nothing less. The patterns will show regularities (some of these regularities may even be interpretable as feedback to motor actions), but the identification of structure and objects from these patterns happens due to the activity of the robot control program, not because of the specifics of the pattern origin. If the world is replaced by an artificial

The same restriction applies, of course, to the mental phenomena of the observer. The observer does not have an exclusive, intimate access to the objects of its cognition and representation that would enable it to witness "real" mental states. What we know about ourselves, including our first-person-perspective, we do not know because we have it available on "our side of the interface." Everything we know about ourselves is a similar ordering we found over features available at the interface; we know of mental phenoma only insofar as they are explicitly accessible patterns or constructed over these patterns. Even though our cognitive processes are responsible for the functionality of ordering/conceptualization and recognition, they are—insofar as they are objects of our examination—"out there" and only available as regularities over patterns (over those patterns that we take to be aspects of the cognitive processes).

From such a point of view, the Cartesian "*cogito ergo sum*" is a quite problematic statement. "*Cogito*" is just the expression of the belief of being in a certain state—and necessarily on the basis of certain perceived features. And naturally, these features may have been caused by something different than a cognitive process. The presupposition of a cognitive process is already an *interpretation* of procedurality, past, distribution and structure of these features. If we want to discover something about our minds, we will have to go beyond our Cartesian intuition and ask: what properties make up our respective concepts? What is the relationship between these concepts?

What the universe makes visible to science (and any observer) is what we might call *functionality*. Functionality, with respect to an object, is—loosely put—the set of causally relevant properties of its feature vector.[6] Features reduce to information, to discernible differences, and the notions we process in our perception and imagination are *systematically structured* information, making up a dynamic system. The description of such systems is the domain of *cybernetics* or *systems science* (Wiener, 1948; Ashby, 1956; von Bertalanffy, 1968; Bischof, 1968; Bateson, 1972;

---

pattern generator (a simulated environment), so that the input data show the same statistical properties with respect to the interpretation, the control program cannot know of any difference.

6 To be more accurate, the notion of *causality* should be treated with more care, because it is an attribution, not an intrinsic property of features. Because causality is an attributed structural property, functionality itself is constructed, even though the regularities classified as causality are not.

Klir, 1992). Systems science is a description of the constructive methods that allow the representation of functionality.

Thus, to understand our concept of mind, we have to ask how a system capable of constructing has to be built, what features and interrelations determine the relevant functionality. The idea of describing the mind itself as a functional system has had an enormous impact on a certain area on psychology and philosophy that has consequently been associated with the term *functionalism* (Fodor, 1987; Putnam, 1975, 1988). If a functionalist subscribes to representationalism (the view that the functional prevalence of a mental state entails its representation within a representing system) a functionalist model of cognitive processes might be implemented as a computer program (*computationalism*) and perhaps even verified this way, so functionalism often goes hand in hand with computer science's proposal of Artificial Intelligence.[7] Even if mental processes could not be modeled as a computational model—any detailed, formal *theory* on how the mind works certainly can (Johnson-Laird, 1988, p. 9).

The idea of a full-featured model of the crucial components of human cognition was advanced by Alan Newell and Herbert Simon as a consequence of the *physical symbol system hypothesis* (Newell & Simon, 1976). According to this hypothesis, a physical symbol system, that is, an implemented *Turing machine*, "has the necessary and sufficient means for general intelligent action. By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence"[8] (Newell, 1987, p. 41).

7 Even though computationalism usually entails functionalism and representationalism, some philosophers maintain that it is possible to be a computationalist without being a functionalist (Block, 1995).

8 Is the physical symbol systems hypothesis equivalent to: "Iron ore is necessary and sufficient for building a locomotive?" On the surface, it goes way beyond that, because not every system built by intricately arranging iron molecules can be extended to pull a train. The physical symbol system hypothesis really refers to a functional, not a material relationship; a better metaphor might be that a steam engine has the necessary and sufficient means to drive a (steam) locomotive; that a steam engine will be found at the core of every steam locomotive, and that every conveniently sized steam engine could be suitably extended. Let's bear in mind, though, that the notion of computation is far more general than the principles of a steam engine. Colloquially speaking, it does not engender much more than *systematic regularity*.

A system capable of fulfilling the breadth of cognitive tasks required for *general intelligence* is a model of a *unified theory of cognition* (Newell, 1987), an implementation of a so-called *cognitive architecture*.

The development of cognitive architectures follows a different paradigm than strict experimental psychology: instead of posing an individual question, designing an experiment to find evidence for or against a possible answer and performing a study with a group of subjects, the cognitive modeler asks *how* a certain set of cognitive feats (for instance, in problem solving) could be possibly achieved and suggests a solution. This solution integrates previous research and might be even detailed enough to make specific predictions on task performance or neural correlates, which allow experimental falsification, either by behavioral studies or by neurobiological examinations (for instance brain imaging). Because the entities that are proposed in a cognitive architecture are usually not all empirically accessible, they have, to put it loosely, to be engineered into the system: the validity of the model depends on whether it works, in accordance to available empirical data, and whether it is sparse, compared to other available models explaining the same data.

This approach to understanding cognition equals the adoption of what Aaron Sloman has called the *constructionist stance* (Sloman, 2000), and bears a slight similarity to Daniel Dennett's suggestion of the *design stance*: "knowing how to design something like *X* is a requirement for understanding how *X* works," (Sloman & Chrisley, 2005).

In principle, a system might be described by identifying its physical makeup—this is what Dennett would term the "physical stance." With respect to the mind, such a description might entail a complete depiction of brain processes, which is usually regarded as unwieldy, perhaps even infeasible, and probably alludes to the wrong level of functionality, just as a thermodynamic description of air molecules might not be helpful to a meteorologist when forecasting tomorrow's weather. A different view is lent by the "design stance," which examines the components making up an artifact, such as buttons, levers, insulators, and so on. Such components might be replaced by other components that serve the same purpose. In a way, this engineering viewpoint is a teleological one, and it might also be applied to biological organisms with respect to organs and the roles they play within the organism. Dennett adds the "intentional stance", which is the description of a system in terms of attributed intentional states, such as beliefs, attitudes, desires and so on (Dennett,

1971). The intentional stance allows predictions about the behavior of the system, but is by no means a complete systematicdescription, because it does not explain how the intentional properties are realized. (Dennett himself does not maintain that the intentional description is always a functional description. Rather, it is an attribution, used by an external observer to characterize the system.[9]) Of course, the descriptions of a thing as either physical, designed or intentional are not mutually exclusive—it can be all these things at the same time, and the stance just marks a different way of looking at it. The physical properties of the system realize the properties of the abstract components that are part of its design, and the intentional properties of the system are eventually realized by the physical properties as well. To find a design description, a structural arrangement of components that realizes the intentional system of a mind might not be a bad description of what the creator of a cognitive architecture is up to.

The goal of building cognitive architectures is to achieve an understanding of mental processes by constructing testable information processing models. Every implementation that does not work, that is, does not live up to the specifications that it is meant to fulfill, points out gaps in understanding. The integration of regularities obtained in experimental psychology into the architecture is not just a re-formulation of what is already known but requires an additional commitment to a way this regularity is realized, and thus a more refined hypothesis, which in turn makes further predictions that can be taken into the lab of the experimental psychologist.

The difference to behaviorism is quite obvious. While the cognitive modeling of functionalist psychology is reluctant to propose and support entities that are not necessary to achieve a certain observable behavior (including everything that can be observed using behavioral and neuroscientific methods), functionalist psychology is essentially compatible with the ideas of scientific positivism, because it makes empirically falsifiable predictions of two kinds:

9 The intentional stance is *permissive*—for instance, a system has a belief in case its behavior can be predicted by treating it as a believer. This "maximally permissive understanding" (Dennett 1998, p. 331) makes no specific claims about inner structure or organization. Rather, Dennett suggests that the properties of a cognitive system are brought forth by a broad collection of "mind tools" which individually need not bear relationships to the outwardly interpretable functionality.

— The proposed model is capable of producing a specific behavior (or test subjects will show a previously unknown property of behavior predicted by the model).
— The model is the sparsest, simplest one that shows the specific behavior with respect to available observations.

If the predictions of the model are invalidated by observations or a more concise model is found, the original model will have to be revised or abandoned. Because cognitive architectures have many free variables, it is often possible to revise an obsolete model to fit conflicting data, so the methodological implications and criticisms arising are by no means trivial. As a result, cognitive architectures as theories do not behave as proposed by classical proponents of positivist methodology: they are often less predictive than integrative (Newell, 1973). But then, large scientific theories rarely do. Just as the extensive theoretical bodies of physics, chemistry, and so on, the unified theories of cognition are not isolated statements that are discarded when one of their predictions is being refuted. Rather, they are *paradigms*, viewpoints that direct a research program, and their adoption or abandonment depends on whether they can be characterized as what Imre Lakatos has called a "*progressive* research paradigm" (Lakatos, 1965); that is, if the shifts in their assumptions lead to more predictions that are substantiated with evidence instead of necessitating further repairs.[10]

The functionalist view on mental phenomena is by no means undisputed in philosophy (Block, 1978; Putnam, 1988). Attacks come from many directions. Especially famous is the position of John Searle, who attacks functionalism by claiming that mental processes, especially consciousness, would be a "causally emergent property" of the physical organism

10 These requirements are not reflected by all cognitive architectures. For instance, while Alan Newell claimed for his *Soar* architecture that it was Lakatosian in nature (Newell, 1990), he also stated: "There is no essential Soar, such that if it changes we no longer have the Soar theory. [...] The theory consists of whatever conceptual elements [...] it has at a given historical moment. It must evolve to be a successful theory at each moment, eliminating some components and tacking on others. [...] As long as each incremental change produces a viable [...] theory from the existing Soar theory, it will still and always be Soar." (Newell, 1992). I will not embark on this aspect of methodological discussion, the interested reader may consult (Cooper et al., 1996) for an introduction into the debate of methodological criticisms of cognitive architectures.

and stem from certain properties provided *only* by biological neurons (Searle, 1992, p. 112). Thereby, Searle ascribes properties to biological neurons that go beyond their otherwise identifiable functionality, that is, an artificial replacement for a neuron that would show the same reactions to neurochemicals and the same interactions with other neurons would not be capable of a contribution to consciousness, and thus, his argument marks an essentialist position (Laurence & Margolis, 1999) that is already incompatible with functionalism on epistemological grounds.[11] If an entity has to have a property that is not empirical itself (and being *biological* is not an empirical property *per se*) to contribute to some functionality, then this entity is conceptually inadequate to capture empirical phenomena in the eyes of a functionalist. Daniel Dennett, in an introduction to Gilbert Ryle's classic "Ghost in the machine" (Dennett, 2002), introduces the idea of a "zombank" to illustrate this. A *zombank* would be something that looks and acts like a financial institution, where people could have an account, store and withdraw money and so on, but which is not a *real bank*, because it lacks some invisible essence beneath its interface and functionality that makes a bank a bank. Just as the notion of a zombank strikes us absurd (after all, a bank is commonly and without loss of generality *defined* by its interface and functionality), Dennett suggests that the idea of a philosophical "zombie," a cognitive system that just acts as if it had a mind, including the ability for discourse, creative problem solving, emotional expression and so on, but lacks some secret essence, is absurd.

Physicalism (or materialism, the philosophical idea that everything is either material or supervenes on the material) is often associated with functionalism—there is not much controversy between functionalists and materialists, functionalists are usually proponents of physicalism (Maslin, 2001, p. 184; Kim 1998).[12]

11 See Preston and Bishop (2002); a point that deserves particular recognition may be Searle's claim that semantics is something which is not reducible to syntax, and that symbol processing systems can only ever know syntax, while intentionality is about semantics (Searle, 1980).

12 Functionalism does not have materialism as a strong requirement, at least not in the sense that states the necessity of matter as a res extensa in the Cartesian sense (Block, 1980). For functionalism to work it is sufficient to have a computational system, and assumptions about the nature of this system beyond its capabilities with respect to computationability are entirely superfluous and speculative. There is also a functionalist emergentist proposal that attempts to construct a non-physical functionalism (Koons, 2003). On the other hand, the position usually called *type physicalism* opposes

If we choose to depict the mind as a dynamic system of functional dependencies, we are not necessarily at an agreement of what to model and how to do it. There are many possible positions that might be taken with regard to the level of modeling, the entities on that level, and of course, to the question as to what makes up a mind. However, the path of designing, implementing, and experimentally testing cognitive architectures seems to be the only productive way to extend philosophy of mind beyond its given bi-millennial heritage, which constrains each theory to the mental capability of an individual thinker. The knowledge embodied in the materials, structure, and assembly of almost any complex industrial artifact like a car, a notebook computer, or a skyscraper goes way beyond of what an individual designer, material scientist, planner, or construction worker may conceive of or learn in their lifetime, but is the result of many interlocking and testable sub-theories within sub-domains and on different levels of abstraction, and the same applies to the large theoretical bodies in physics, biology, computer programming, and so on. Yet in the field of the philosophy of mind, theories are typically associated with and constrained to individual thinkers. If understanding the mind is not much simpler than the design of the plumbing of a skyscraper, then there may be reason to believe that any theory of mental functioning that fits into a single philosopher's mind and is derived and tested solely by her or his observations and thought-experiments is going to be gravely inadequate. Pouring theories of mental functioning into formal models and testing these by implementing them may soon become a prerequisite to keep philosophy of mind relevant in an age of collaborative and distributed expertise.

On the other hand, cognitive modeling is lacking approaches that are broad enough to supply a foundation for theoretical bodies of a philosophy of mind. Broad and not too shallow theories of cognition will be a requirement for substantial progress in understanding the mind.

1.2 Classes of cognitive models

Models of cognition can be classified in various ways (Logan, 1998; Pew & Mavor 1998; Elkind et al., 1989; Morrison, 2003; Ritter et al., 2002).

functionalism and instead maintains that mental states are identical to physical states (Fodor, 1974; Papineau, 1996).

Architectures that attempt to model mental faculties form several methodological groups.

They might be divided into

— Classical (symbolic) architectures, which are essentially rule-based. These sprang up after Newell's call for a revival of unified theories in psychology (Newell, 1973a, 1987). Classical architectures concentrate on symbolic reasoning, bear influences of a relatively strict language of thought concept, as suggested by Fodor, and are often implemented as production based language interpreters. Gradually, these architectures have been modified to allow for concept retrieval by spreading activation, the formation of networks from the initial rules and have occasionally even been implemented based on neural elements.

— Parallel distributed processing (PDP) (subsymbolic) architectures. This term was introduced by James McClelland (Rumelhart, McClelland et al., 1986); here, it is used to refer to nonsymbolic distributed computing (usually based on some or several types of recurrent neural networks). Where classical architectures strive to attain the necessary complexity by carefully adding computational mechanisms, PDP systems are inspired by biological neural systems. Their contemporary forms essentially work by constraining a chaotic system enough to elicit orderly behavior. While PDP architectures do not necessarily differ in computational power from classical architectures, it is difficult to train them to perform symbolic calculations, which seem to be crucial for language and planning. On the other hand, they seem to be a very productive paradigm to model motor control and many perceptual processes.

— Hybrid architectures may use different layers for different tasks: a reasoning layer that performs rule-based calculations, and a distributed layer to learn and execute sensory-motor operations. Hybrid architectures are usually heterogenous (i.e., they consist of different and incompatible representational and computational paradigms that communicate with each other through a dedicated interface), or they could be homogenous (using a single mode of representation for different tasks). The

latter group represents a convergence of classical and PDP architectures, and our own approach follows this direction.

— Biologically inspired architectures, which try to directly mimic neural hardware—either for a complete (simple) organism, or as a layer within a hybrid approach.

— In my view, emotion and motivation are vital parts of a cognitive system, but this distinction does not take care of how they are introduced into the system. This is because most existing models either ignore them or treat them as separate entities, situated and discussed outside the core of the model. Exceptions to the rule exist, of course, for instance Clarion (Sun, 2003, 2005), PURR-PUSS (Andreae, 1998) and of course the Psi theory, which all treat emotion and motivation as integral aspects of the cognitive system. For many other cognitive architectures, separate additions exist, which provide an emotional or motivational module that interfaces with the cognitive system (Belavkin, Ritter, & Elliman, 1999; Norling & Ritter, 2004; Franceschini, McBride, & Sheldon, 2001; Gratch & Marsella, 2001; Jones, 1998; Rosenbloom, 1998).

As noted before, cognitive modeling is not constrained to the realm of psychology, yet most existing approaches have their origins in psychological theory. Many interesting contributions, however, came from Artificial Intelligence (AI). AI as a field arguably does not seem much concerned with full-blown models of cognition (Anderson, 1983, p. 43), and most AI architectures do not attempt to model human performance, but strive to solve engineering problems in robotics, multi-agent systems, or human-computer interaction. On the other hand, contemporary AI architectures tend to start out from an agent metaphor, building an autonomous system that acts on its own behalf and is situated in an environment, whereas low-level architectures in psychology usually deal with isolated or connected modules for problem solving, memory, perception and action, but leave out motivation and personality. There are psychological theories of motivation and personality, of course (Kuhl, 2001; Lorenz, 1965, 1978), but they rarely visit the lowly realms of computational models. There is no strict boundary between AI architectures and cognitive architectures in psychology,

however, and most of the latter are based on representational mechanisms, description languages, memory models, and interfaces that have been developed within AI.

1.2.1 Symbolic systems and the Language of Thought Hypothesis

Research in the field of cognitive architectures traditionally focused on symbolic models of cognition, as opposed to subsymbolic, distributed approaches. Classical, symbolic architectures are systems that represent and manipulate propositional knowledge. If there are things to be represented and manipulated that are not considered propositional knowledge, they are nonetheless represented in the form of propositional rules (productions). Let us make the philosophical commitment behind this approach more explicit: symbolic architectures are proponents of a symbolic *Language of Thought* (LOT).

The *Language of Thought Hypothesis* (LOTH) is usually attributed to Jerry Fodor (1975), and it strives to explain how a material thing can have semantic properties, and how a material thing could be rational (in the sense of how the state transitions of a physical system can preserve semantic properties). (A summary is given by Aydede, 1998.)

Fodor gives the following answers:

— Thought and thinking take place in a mental language. Thus, thought processes are symbolic, and thinking is syntactic symbol manipulation.
— Thoughts are represented using a combinatorial syntax and semantics.
— The operations on these representations depend on syntactic properties.

LOTH is, by the way, not concerned with questions like "how could anything material have conscious states?," "what defines phenomenal experience?," or "how may qualia be naturalized?"

Fodor did not exactly state something new in 1975, and thus did not open up a new research paradigm in cognitive science. Rather, he spelled out the assumptions behind artificial intelligence models and cybernetic models in psychology: Perception is the fixation of beliefs, the learning of concepts amounts to forming and confirming hypotheses, and decision making depends on representing and evaluating the consequences of actions depending on a set of preferences. If all these aspects of cognition

can be seen as computations over certain representations, then there must be a language over which these computations are defined—a language of thought. Fodor was also not the first to express this idea (see, for instance, Ryle, 1949), but he narrowed it down to an argument that sparked a debate about the nature of the language of thought, a debate that is far from over.

The Language of Thought Hypothesis makes three main assumptions:

First, the *representational theory of the mind* (Field, 1978, p. 37; Fodor, 1987, p. 17), which consists of two claims—the representational theory of thought (i.e., thoughts are mental representations), and the representational theory of thinking (the processes that operate on the thoughts are causal sequences of instantiations, or *tokenings*, of mental representations), in other words: thinking consists in processing mental representations in an algorithmic manner.

Second, LOTH asks that these representations reside somehow in the subject's physical makeup. This amounts to functionalist materialism (i.e., mental representations are realized by physical properties of the subject, or, colloquially put, mental representations are somehow and only stored in the physical structures of the brain and body). This does not necessarily imply that all propositional attitudes need to be represented explicitly (Dennett, 1981, p. 107); it is sufficient if they are functionally realized. On the other hand, not all explicit representations within a cognitive system need to be propositional attitudes (because not all of them are in a proper psychological relation to the subject; see Fodor, 1987, p. 23–26).

The next assumption of LOTH is, at least as far as cognitive science is concerned, the most controversial one: Mental representations have a *combinatorial syntax and semantics*, with structurally simple, atomic constituents making up structurally complex, molecular representations in a systematic way, whereby the semantics of the complex representations is a function of the semantics of the atomic constituents and their formal structure. This claim about represented mental content is complemented by a claim about operations over this content: the operations on mental representations are causally sensitive to the formal structure defined by the combinatorial syntax; the semantics follow formal, combinatorial symbol manipulation.

According to LOTH, a thinking system is characterized by representational states (the "thoughts") and semantically preserving transitions between them (the "thought processes"), which can be described as a formal language with combinatorial syntax, that is, a computational engine. This immediately raises the question: How does the representational structure of a Language of Thought acquire its meaning? This is commonly called the *symbol grounding problem* (Harnad, 1987, 1990; Newton 1996). LOTH proponents respond in two ways: either, the atomic symbols can somehow be assumed to have a meaning, and the molecular symbols inherit theirs by a Tarski-style definition of truth conditions according to the syntactic operations that make them up of atomic components (Field, 1972; Tarski, 1956), or the semantics arise from the constraints that are imposed by the computational roles the individual components assume in the syntactic structure.[13] (For a critical discussion, see Haugeland, 1981, and Putnam, 1988.)

How does Fodor back up the strong claim that mental representations are following the rules of a formal language with combinatorial syntax? Obviously, a system may represent and compute things without obeying the requirement of combinatorial syntax (i.e., nonsymbolic) or with limited structural complexity. Fodor (1987, see also Fodor & Pylyshyn, 1988) points out that:

1. Thinking is *productive*. While one can only have a finite number of thoughts in their lifetime (limited performance), the number of possible thoughts is virtually infinite (unbounded competence). This can be achieved by systematically arranging atomic constituents, especially in a recursive fashion.
2. Thoughts are *systematic* and *compositional*. Thoughts come in clusters, and they are usually not entertained and understood in isolation, but because of other thoughts they are based on and related to. Thoughts are usually not atomic, but are syntactically made up of other elements in a systematic way. Systematically related thoughts are semantically related, too.

13 If a cognitive system is temporarily or permanently disconnected from its external environment, does its mental content cease to be meaningful? If not, then the semantics will have to reside entirely within the conceptual structure, i.e. they are determined by the constraints that individual representational components impose onto each other via their syntactic relationships.

3. Thinking itself is systematic (argument from *inferential coherence*). For instance, if a system can infer *A* from *A and B*, then it is likely to be able to infer C from C *and D*, so thoughts are obviously not just organized according to their content, but also according to their structure. A syntactically operating system takes care of that.

The mindset behind the Language of Thought Hypothesis clearly sets the scene for symbolic architectures. Their task consists of defining data structures for the different kinds of mental representations, distinguishing and defining the relations that these data structures have within the system (laying out an architecture that handles beliefs, desires, anticipations and so on), and specifying the set of operations over these representations (the different kinds of cognitive processes).

LOTH also provides a watershed between symbolic and connectionist approaches: Not all theorists of cognitive modeling, even though they tend to accept functionalist materialism and the representational theory of mind, agree with Fodor's proposal. Many connectionists argue that symbolic systems lack the descriptive power to capture cognitive processes (for a review, see Aydede, 1995). Yet they will have to answer to the requirements posed by productivity, systematicity, and inferential coherence by providing an architecture that produces these aspects of mental processes as an emergent property of nonsymbolic processing. Fodor (Fodor & Pylyshyn, 1988) maintains that a connectionist architecture capable of productivity, systematicity and inferential coherence will be a functional realization of a symbolic system (i.e., the connectionist implementation will serve as a substrate for a symbolic architecture).

A weighty argument in favor of connectionism is the fact that low-level perceptual and motor processes—which may be regarded as subcognitive (Newell, 1987)—are best described as distributed, nonsymbolic systems, and that the principles governing these levels might also apply to propositional thinking (Derthick & Plaut, 1986). Are Language of Thought systems just too symbolic? A connectionist description might be better suited to capture the ambiguity and fuzziness of thought, where a symbolic architecture turns brittle, fails to degrade gracefully in the face of damage or noise, does not cope well with soft constraints, and has problems integrating with perceptual pattern recognition.

Connectionists might either deny strict systematicity and compositionality of thought (Smolensky, 1990, 1995; see also Chalmers, 1990,

1993), or regard them as an emergent by-product of connectionist processing (Aizawa, 1997).

This is the line where classical and connectionist models of cognition fall apart. Where Fodor states that because of the productivity, systematicity and compositionality of thought, symbolic languages are the right level of functional description, connectionists point at the vagueness of thought and argue that the level of symbolic processes is not causally closed (i.e., cannot be described without resorting to nonsymbolic, distributed operations) and is therefore not the proper level of functionality (see: Rumelhart & McClelland, 1986; Fodor & Pylyshyn, 1988; Horgan & Tienson, 1996; Horgan, 1997; McLaughlin & Warfield, 1994; Bechtel & Abrahamsen, 2002; Marcus, 2002).

Are classicist and connectionist approaches equivalent? Of course, all computational operations that can be performed with a connectionist system can be implemented in a symbol manipulation paradigm, and it is possible to hard-wire an ensemble of connectionist elements to perform arbitrary symbol manipulation, but the difference in the stance remains: symbolic processes (especially recursion), which seem to be essential for language, planning, and abstract thought, are difficult to model with a connectionist architecture, and many relations that are easy to capture in a connectionist system are difficult to translate into a symbolic, rule-based system, without emulating the connectionist architecture. In practice, however, the line between classical and connectionist models is not always clear, because some classical models may represent rule sets as spreading activation networks, use distributed representations and even neural learning, and some connectionist systems may employ localist representations for high-level, abstract operations.

Hybrid systems may combine connectionist and symbolic architectures, either by interfacing a symbolic control layer with subsymbolic perceptual and motor layers (Konolidge, 2002; Feldman, 2006), or by using a common (semi-symbolic) mode of representation that allows for both kinds of operations (Sun, 1993; Wermter, Palm et al., 2005). The latter method treats symbolic representations as a special (highly localized) case of distributed representations, and because the author

believes that both are required in a unified framework, our own approach (PSI and MicroPSI) also falls into this category.

1.2.2 Cognition without representation?

Apart from the connectionist attack, there is another front against Fodor's proposal in cognitive science, which denies the second assumption of the Language of Thought Theory—representationalism. This position is exemplified in earlier works of Rodney Brooks (Brooks, 1986, 1989, 1991, 1994; Brooks and Stein 1993) and denies Fodor's dictum of "*no cognition without representation*" (1975), by stating that "*the world is its own best model*" and the relevant functional entities of cognitive processes would not be information structures stored in the nervous system of an individual, but emergent properties of the interaction between the individual and its environment. Therefore, a functional cognitive model either requires the inclusion of a sufficiently complex model of the environment, or the integration of the model of mental information processing with a physical (or even social) environment (Dreyfus, 1992). The proponents of *behavior-based robotics* (Beer, 1995; Arkins, 1998; Christaller, 1999; Pfeifer & Bongard, 2006) sometimes reject the former option and insist on a physical environment, either because of objections to functionalism (i.e., because they think that the simulation of a physical environment is *in principle* an impossibility), or just because they consider a sufficiently complex environmental simulation to be practically impossible. Taken to the extreme, behavior-based approaches even become behaviorist and deny the functional relevance of mental representations altogether, treating them as an irrelevant epi-phenomenon (Brooks, 1992; van Gelder, 1995; Beer, 1995; Thelen & Smith, 1994). Even in their nonradical formulation, behavior-based approaches sometimes deny that the study of cognition may be grounded on a separation of system and environment at the level of the nervous system. Without the inclusion of an environment into the model, the low level configurations of the nervous system do not make any sense, and because high-level configurations are inevitably based on these low-level structures, a study of cognition that draws a systemic line at the knowledge level, at the neural level, or at the interface to the physical world, is doomed from the start.

By highlighting low-level control of interaction with a physical environment, behavior-based systems achieve fascinating results, such as passive walkers (e.g., Kuo, 1999; Pfeifer, 1998; Collins et al., 2005), which produce two-legged walking patterns without the intervention of

a cognitive system. The credo of such approaches might be summarized as "physics is cognition's best friend," and they sometimes see cognition primarily as an extension of such low-level control problems (Cruse, Dean, & Ritter, 1998; Cruse, 1999).

I see two objections to radical behavior-based approaches, which in my view limit their applicability to the study of cognitive phenomena: First, while a majority of organisms (*Drosophila*, the fruitfly, for instance) manages to capitalize on its tight integration with physical properties of its environment, only a small minority of these organisms exhibits what we might call cognitive capabilities. And second, this majority of tightly integrated organisms apparently fails to include famous physicist Stephen Hawking, who is struck with the dystrophic muscular disease ALS and interacts with the world through a well-defined mechatronic interface—his friendship with physics takes place on an almost entirely knowledge-based level. In other words, tight sensor-coupling with a rich physical environment seems neither a sufficient nor a necessary condition for cognitive capabilities.

Also, dreaming and contemplation are being best understood as cognitive phenomena; and they take place in isolation from a physical environment. The physical environment may have been instrumental in building the structures implementing the cognitive system and forging the contents of cognition, and yet, after these contents are captured, it does not need to play a role any more in defining the semantics of thought during dreaming, meditation, and serendipitous thinking. Even when high-level cognitive processing is coupled with the environment, it does not follow that the nature of that coupling has a decisive influence on this processing.[14]

14 Andy Clark and Josefa Toribio (2001), in a commentary on O'Reagan and Noë's "sensorimotor account of vision and visual consciousness", have denounced the view that conscious processing could only be understood in conjunction with environmental coupling as "sensorimotor chauvinism." They point out the example of a ping-pong playing robot (Andersson, 1988), which does not know visual experience, and yet performs the task—and on the other hand, they argue that it is implausible that all changes to our low-level perception, for instance, in the speed of saccadic movement, would influence conscious experience. Because there seems to be no *a-priori* reason to believe that this is the case, actual environmental coupling is not only an insufficient condition, but likely also not a necessary condition for high-level cognition and consciousness. For high-level mental activity, higher level mechanisms (Prinz, 2000) such as memory retrieval, planning, and reasoning should be constitutive. Of course, this view contradicts a lot of contemporary arguments in the area of behavior-based robotics.

For reasons of technical complexity, it might be easier to couple a cognitive model with a physical environment instead of a simulation, and a lot may be learned from the control structures that emerge from that connection. And yet, the organization and structuring of a cognitive system might be an entirely different story, according to which the division of the modeled system and the given environment at the somatic level or even above the neural level might be just as appropriate as the intuitions of symbolic and sub-symbolic cognitivists suggest.

1.3 Machines of cognition

> *"Every intelligent ghost must contain a machine."*
>
> Aaron Sloman (2002)

Cognitive architectures define computational machines as models of parts of the mind, as part of the interaction between cognitive functions and an environment, or as an ongoing attempt to explain the full range of cognitive phenomena as computational activity. This does not, of course, equate the human mind with a certain computer architecture, just as a computational theory of cosmology—a unified mathematical theory of physics—maintains that the universe is possessed by a certain computer architecture. It is merely a way of expressing the belief that scientific theories of the mind, or crucial parts of research committed to a better understanding of the mind, may be expressed as laws, as rules, as systematized regularities, that these regularities can be joined to a systematic, formal theory, and that this theory can be tested and expanded by implementing and executing it as a computer program.

1.3.1 Cognitive science and the computational theory of mind

If we take a step back from the narrow issue of whether we should use a symbolic computational engine to describe cognition, or if we should aim at specifying a symbolic computational engine that describes a nonsymbolic architecture that takes care of producing cognitive functionality (and this is, in my view, what the question boils down to), the fact remains that cognitive modeling is committed to a computational theory of mind (see Luger, 1995 for an introduction). There are two viewpoints in cognitive science with respect to the computational

theory of mind (i.e., that the mind can be described as a computational engine). The theory may be seen as an ontological commitment (in the form that either the universe itself is a computational process (e.g., Wolfram, 2002), and thus everything within it—such as minds—is computational too, or that at least mental processes amount to information processing). But even if one does not subscribe to such a strong view, the theory of mind may be treated as a methodological commitment. This second view, which I would like to call the "weak computational theory," has been nicely formulated by Johnson-Laird, when he said:

> Is the mind a computational phenomenon? No one knows. It may be; or it may depend on operations that cannot be captured by any sort of computer. (...) Theories of the mind, however, should not be confused with the mind itself, any more than theories about the weather should be confused with rain or sunshine. And what is clear is that computability provides an appropriate conceptual apparatus for theories of the mind. This apparatus takes nothing for granted that is not obvious. (...) any clear and explicit account of, say, how people recognize faces, reason deductively, create new ideas or control skilled actions can always be modelled by a computer program. (Johnson-Laird, 1988)
>
> Indeed, cognitive models can be seen as the attempt to elucidate the workings of the mind by treating them as computations, not necessarily of the sort carried out by the amiliar digital computer, but of a sort that lies within the broader framework of computation (*ibid*, p. 9).[15]

15 This does not mean that a digital computer is incapable of performing the computations in question. Here, Johnson-Laird hints at parallel distributed processing as opposed to sequential binary operations in a von-Neumann computer. The operations that are carried out by a parallel distributed system can be emulated on a digital computer with sufficient speed and memory with arbitrary precision. Computationally, parallel distributed operations do not fall into a different class than those executed by a traditional von-Neumann computer; both are instances of deterministic Turing machines with finite memory. An exception would be a system that employs certain quantum effects (non-locality and simultaneous superposition of states). Such a quantum computer may be in more than one state at once and thus execute some parallel algorithms which a deterministic Turing machine performs in non-polynomial time in linear time (Deutsch, 1985). Indeed, some theorists maintain that such quantum processes play a role in the brain and are even instrumental in conscious processes (Lockwood, 1989; Penrose, 1989, 1997; Stapp, 1993; Mari & Kunio, 1995). However, there is little evidence both for quantum computing facilities in the human brain or

Thus, a complete explanation of cognition would consist of a computational model that, if implemented as a program, would produce the breadth of phenomena that we associate with cognition. In that sense, the computational theory of mind is an empirical one: it predicts that there may be such a program. Unfortunately, this does not mean that the computational model of the mind could be falsified based on its predictions in any strict sense: If there is no computational model of the mind, it may just mean that it is not there *yet*. This lack of falsifiability has often been criticized (Fetzer, 1991). But does this mean that the computational theory of mind is of no empirical consequence at all and does not have any explanative power, as for instance, Roger Binnick (1990) states? Binnick applies the same criticism to Chomsky's theory of language (1968), even though

> linguistics constitutes (apart from the theory of vision and perhaps a few corners of neuropsychology) just about the only cognitive system for which we can say we have something like a formal and explicit theory of its structure, function, and course of development in the organism (S. R. Anderson, 1989, p. 810)

From the viewpoint of natural sciences, this criticism is surprising, and in most cases may be assumed to originate in a misunderstanding of the notion of computation. All theories that are expressed in such a way that they may be completely translated into a strict formal language are computational in nature. The ontological or methodological assumption that is made by the computational theory of mind is not unique to cognitive science, but ubiquitously shared by all nomothetic (Rickert, 1926) sciences, that is, all areas that aim at theories that describe a domain exhaustively using strict laws, rules, and relations. This is especially the case for physics, chemistry, and molecular biology.

Of course, there are areas of scientific inquiry that do not produce insights of such nature, but are descriptive or hermeneutic instead. These sciences do not share the methodology of natural sciences. Indeed, the rejection of a computational stance with respect to a subject marks that the field of investigation is one of the cultural sciences (humanities). To treat psychology as a natural science means to subscribe to the

---

the explanatory power of such states for cognitive processes or consciousness, which is questionable.

computational theory of mind—either in its weak or even in its strong form (see also Dörner, 1999, p. 16).

This view has also been expanded upon by Aaron Sloman (see Sloman & Scheutz, 2001; Sloman & Chrisley, 2005). Sloman characterizes the task of describing the world as a quest for suitable ontologies, which may or may not supervene on each other (Kim, 1998). When describing systems that describe other systems, we will create second-order ontologies. If such systems even describe their own descriptions, recursive third-order ontologies will need to be employed (this is where it ends—further levels are addressed by recursion within the third). Conceptualizations of second order and third order ontologies are creations of *virtual machines*. A virtual machine is an architecture of causally related entities that captures the functionality of an information processing subject or domain, and if mental phenomena can be described as information processing, then a theory of cognition will be a complex virtual machine.

Contrary to the intuition that machines are always artifacts, here, a machine is simply seen as a system of interrelated parts that are defined by their functionality with respect to the whole:

> Machines need not be artificial: organisms are machines, in the sense of "machine" that refers to complex functioning wholes whose parts work together to produce effects. Even a thundercloud is a machine in that sense. In contrast, each organism can be viewed simultaneously as several machines of different sorts. Clearly organisms are machines that can reorganize matter in their environment and within themselves, e.g. when growing. Like thunderclouds, windmills and dynamos, animals are also machines that acquire, store, transform and use energy. (Sloman & Chrisley, 2005)

For a given system (given by a functional description with respect to its environment), however, it is not always clear what the functional parts are—there is not even a guarantee that there is sufficient modularity within the system to allow its separation into meaningful parts. An ontology that specifies parts needs to be justified with respect to completeness—that the parts together indeed provide the functionality that is ascribed to the whole—and partitioning—that it does not misconstrue the domain. For example, if the gearbox of a car is described as the part that takes a continuous rotational movement with a certain angular

momentum from the crankshaft and transforms it into a variety of different rotational movements with different momentums to drive the wheels, this might be a good example for a functional element. If the gearbox is removed and replaced by a different unit that provides the same conversion, the function of the overall system—the car—might be preserved. Such a separation is often successful in biological systems too. A kidney, for instance, may be described as a system to filter certain chemicals from the bloodstream. If the kidneys are replaced by an artificial contraption that filters the same chemicals (during *dialysis*, for instance), the organism may continue to function as before. There are counterexamples, too: a misconstrued ontology may specify the fuel of the car simply as an energy source. If the fuel tank would be replaced by an arbitrarily chosen energy source, such as an electrical battery, the car would cease to function, because fuel is not just an energy source—to be compatible with a combustion engine, it needs to be a very specific agent that when mixed with air and ignited shows specific expansive properties. The car's fuel may perhaps be replaced with a different agent that exhibits similar functional properties, such as alcohol or natural gas, provided that the compatibility with the engine is maintained. Even then, there might be slight differences in function that lead to failure of the system in the long run, for instance, if the original fuel has been providing a lubricating function that has been overlooked in the replacement. Similarly, the mind is not just an information processing machine (for instance, a Turing Machine). Still, it may in all likelihood be described as an information processing machine as well, in the same way as fuel in a car may be depicted as an energy source, but this description would be far too unspecific to be very useful! The difficulty stems from the fact that there is little agreement in cognitive science and psychology as towhat, exactly, defines mental activity (i.e., what the properties of the whole should be). Even if we limit our efforts to relatively clearly circumscribed domains, the ontologies that we are using to describe what takes place on different levels and which supervene on each other are not necessarily causally closed.[16] For

16 Causal closure may best be explained by an example: in graphical user interfaces, widgets indicating similar functions may be implemented by different programming libraries. Nevertheless, a click on the closing icon of a main window usually ends an associated application, no matter which interface programming library realizes the functionality. This allows for the user to neglect the programming level of the application and use the abstraction of the interface when describing the system. But what happens if clicking the closing icon fails to close the application? Sometimes, the reason

instance, language processing may be difficult to study in isolation from the representation of and abstraction over perceptual content (Feldman, et al., 1996), perception may be impossible to study without looking at properties of neural circuitry with respect to synchronization and binding (Engel & Singer, 2000; Singer, 2005), and even relatively basic perceptual processing like the formation of color categories may depend on language capabilities (Steels & Balpaeme, 2005).

The study of cognitive architecture somehow has to cope with these difficulties—either by specifying a very complex, mainly qualitative architecture that does not lend itself to quantitative experiments (see Sloman & Scheutz, 2003; Baars, 1993; Franklin, Kelemen, & McCauley 1998), by attempting to simplify as much as possible by reducing the architecture to a small set of organizational principles that can be closely fitted to experimental data in narrow domains (Laird, Newell, & Rosenbloom 1987; Anderson & Lebiére, 1998; Touretzky & Hinton, 1988; Smolensky, 1995), or by an attempt to find a middle ground (Sun, 2005, Dörner, 1999, Feldman, 2006).

1.3.2 Classical (symbolic) architectures: Soar and ACT-R

Alan Newell committed himself strongly to the Language of Thought Hypothesis, when he stated his own version in 1976 (Newell & Simon, 1976): "A physical symbol system has the necessary and sufficient means for general intelligent action," a dictum that has since been known as the *Physical Symbol Systems Hypothesis* (PSSH). According to Newell, a symbol system is made up of

- memory, which contains the symbol information
- symbols, which supply patterns to index information and give references to it
- operators, to manipulate the symbols
- interpretations, which specify the operations over the symbols.

---

resides on the level of the application interface, for instance, because the application still holds an unsaved document. In this case, the causal frame of the application interface is not broken. But if the window fails to close because of the hidden interaction of the programming library with a different application that uses the same instance of the programming library, then the behavior of the graphical user interface can only be understood if the different programming libraries are taken into account. The frame of the graphical user interface is no longer a self-contained ontology but needs to be expanded by elements of the level it supposedly supervenes on.

Table 1.1 Layers of Description of a Cognitive System (Newell, 1990)

Scale (seconds)	System	Stratum
10^7		
10^6		Social
10^5		
10^4		
10^3	Tasks	Rational
10^2		
10^1	Unit Tasks	
10^0	Operations	Cognitive
10^{-1}	Deliberative Acts	
10^{-2}	Neural Circuitry	
10^{-3}	Neurons	Biological
10^{-4}	Organellae	

To function, a symbol system has to observe some basic requirements: it needs sufficient memory, and it has to realize composability and interpretability. The first condition, composability, specifies that the operators have to allow the composition of any symbol structure, and interpretability asks that symbol structures can encode any valid arrangement of operators.

A fixed structure that implements such a symbol system is called a *symbolic architecture*. The behavior of this structure (that is, the program) only depends on the properties of the symbols, operators and interpretations, not on the actual implementation; it is independent of the physical substrate of the computational mechanism, of the programming language and so on.

The advantages of a symbolic architecture are obvious: because a large part of human knowledge is symbolic, it may easily be encoded (Lenat, 1990); reasoning in symbolic languages allows for some straightforward conceptualizations of human reasoning, and a symbolic architecture can easily be made computation complete (i.e., Turing computational: Turing 1936).

According to Newell (1990), cognitive acts span action coordination, deliberation, basic reasoning and immediate decision-making—those mental operations of an individual that take place in the order of hundreds of milliseconds to several seconds (insert table 1.1). Long-term behavior, such as the generation and execution of complex plans, the acquisition of a language, or the formation and maintenance of a social

role, go beyond the immediately modeled area and are facilitated by many successive cognitive acts. The neurobiological level is situated below the cognitive band and falls outside the scope of a functional theory of cognition.

Alan Newell has set out to find an architecture that—while being as simple as possible—is still able to fulfill the tasks of the cognitive level, a minimally complex architecture for *general intelligence* (i.e., with the smallest possible set of orthogonal mechanisms). To reproduce results from experimental psychology, so-called *regularities* (covering all conceivable domains, be it chess-playing, language, memory tasks, and even skiing), algorithms would be implemented *within* these organizational principles. Newell's architecture (Newell, 1990; Laird, Newell, & Rosenbloom, 1987) is called *Soar* (originally an acronym that stood for *State, Operator and Result*) and originated in his conceptions of human problem solving (Newell 1968; Newell & Simon, 1972). Soar embodies three principles: heuristic search for the solution of problems with little knowledge, a procedural method for routine tasks, and a symbolic theory for bottom-up learning, implementing the *Power Law of Learning* (Laird, Newell, & Rosenbloom, 1986).

Central to Soar is the notion of *Problem Spaces*. According to Newell, human rational action can be described by

— a set of knowledge states
— operators for state transitions
— constraints for the application of operators
— control knowledge about the next applicable operator.

Consequently, a problem space consists of a set of states (with a dedicated start state and final state) and operators over these states. Any task is represented as a collection of problem spaces. Initially, a problem space is selected, and then a start state within this problem space. The goal is the final state of that problem space. During execution, state transitions are followed through until the goal state is reached or it is unclear how to proceed. In that case, Soar reaches an *impasse*. An impasse creates and selects a new problem space, which has the *resolution* of the impasse as its goal. The initial problem spaces are predefined by the modeler.

Problem spaces are also defined independently from Soar, for example in STRIPS (*Stanford Research Institute Problem Solver*; Fikes & Nilsson, 1971), and generally contain a set of goals (with the top-level goal being the task of the system), a set of states (each of which is realized as a set of

literals describing knowledge and world model) and a set of valid operators and constraints.

The actual problem solving work in Soar is delivered by the *operators*. Operators are algorithms that describe how to reach the next state; they are executed upon the filling of the context slots of a problem space. Soar can develop new operators on its own, but its models typically work with a set of predefined operators (often augmented with a library of about 50 default rules for planning and search, including means-end analysis, hill-climbing, alpha-beta search, branch and bound); the system may learn which one to apply in a given context. This represents a considerable extension over Newell's and Simon's earlier attempt at a universal problem-solving mechanism, the *General Problem Solver* (1961), which did, among many other restrictions, only have a single problem space and two operators: means-end analysis and sub-goaling to find a new operator. Also, it lacked the impasse mechanism to recognize missing knowledge (see also Newell, 1992).

As a strictly symbolic architecture, Soar stores knowledge in the form of rules (*productions*, also called "chunks"), even though a neuro-symbolic implementation exists (Cho, Rosenbloom, & Dolan, 1993). Perception and action are originally not integral parts of the Soar architecture—they are supplied by independent, asynchronous modules (e.g., from EPIC, Chong, & Laird, 1997). Despite many successful applications (e.g., Gratch & Marsella, 2004; Ritter & Bibby, in press), Soar is frequently criticized for being more of an AI programming language (Ritter, Baxter, et al., 2002) than it is a model of human cognition.

Many of the criticisms that apply to Soar have later been addressed by John Anderson's *ACT theory*[17] (Anderson, 1983, 1990; Anderson & Lebiere, 1998). ACT is—next to Soar—currently the most extensively covered and applied model in the field of symbolic cognitive architectures, and probably the one best grounded in experimental psychological research literature (Morrison, 2003, p. 30). Just as Soar, ACT-R is based on production rules, but unlike Soar, it allows for real-valued activations (instead of a binary on-off), which are biologically more plausible,

17 ACT is an acronym that supposedly stands for the *Adaptive Character of Thought* (it meant the *Adaptive Control of Thought* earlier, has also been reported to abbreviate *Atomic Components of Thought* (Morrison, 2003) and perhaps, it just refers to *Anderson's Cognition Theory*). The 'R' abbreviates *Rational* and refers to Anderson's *rational analysis* (Anderson, 1990, 1991).

because the spread of activation is governed by time, not by programming steps (Anderson, 1978, 1983).

The ACT theory has its roots in a model of human associative memory (HAM, Anderson and Bower 1973), which was an attempt to provide a descriptive language of mental content, made up of hierarchies of connected nodes (called "chunks") in a semantic network and featuring associative recall. In the course of its development, it was also extended by perceptual and motor facilities (PM, Byrne & Anderson, 1998; Byrne, 2001).

ACT-R (and its predecessor, ACT*) have both claimed to bridge the gap between neural-like implementation and symbolic computation. The connectionist implementation of ACT-R, ACT-RN (Lebière & Anderson, 1993), is an attempt to substantiate that claim. ACT-RN's implementation of a declarative memory makes use of a simplified Hopfield network (Hopfield, 1984) with real values, with each chunk acting as a node in the network. To limit the number of links, in the connectionist implementation, the declarative memory is split into several areas. Within each area, all chunks are fully connected to each other. ACT-RN has been used in several cognitive models, but has been abandoned nonetheless, because it was considered too unwieldy for the intended applications—the development of current ACT-R versions focuses on symbolic implementations. Even so, the retrieval of chunks partially follows a sub-symbolic paradigm: spreading activation.

In addition to the declarative memory, ACT-R proposes a procedural memory. Such a distinction has, for instance, been suggested by Squire (1994), but is far from being undisputed in the literature of psychology (Müller, 1993). Procedural memory consists of production rules, which coordinate the cognitive behavior using a goal stack that is laid out in working memory (see Figure 1.1).

Using chunks and productions, ACT-R can encode temporal strings (which are somewhat like scripts, see Schank & Abelsson, 1977), spatial images (similar to schemas; Minsky, 1975) and abstract propositions. The activity of the system is determined by a probabilistic, goal-oriented matching process of productions, which leads to the acquisition of new procedures (productions) and the manipulation of declarative knowledge.

ACT-R has been designed to mimic human performance more closely than Soar and attempts to be an integrated theory of the mind (Anderson et al., 2004). Recently, John Anderson's perspective on modeling cognition shifted even further from symbolic abstraction towards modeling

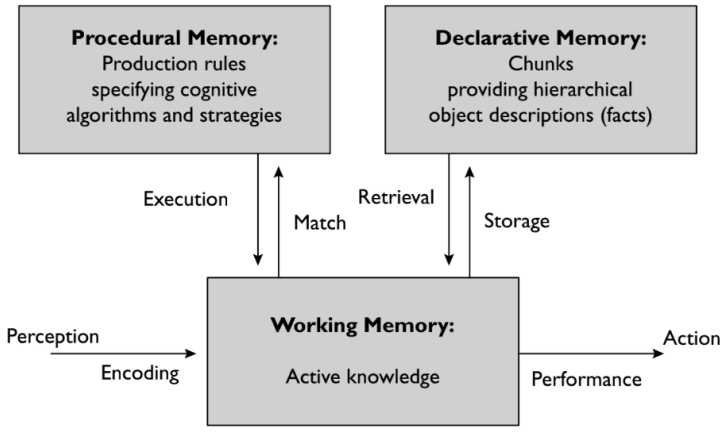

Figure 1.1 ACT-R memory organization (simplified, see Anderson 1983, p. 19)

brain functions and maintains that "a *cognitive architecture* is a specification of the structure of the brain at a level of abstraction that explains how it achieves the function of the mind" (Anderson, 2007, p. 7). But ACT-R is not a model of a complete cognitive system. ACT-R models have captured many regularities of the behavior of subjects in psychological experiments, from visual perception tasks to the learning of mental arithmetic, and its success stems not least from the fact that it allows for testing its cognitive models by comparing computation times with those of human subjects, without making more than a few very basic assumptions on the speed of activation spreading. On the other hand, ACT-R models are usually not autonomous or motivated—goals of the system are given explicitly and beforehand by the experimenter.

Anderson maintains that ACT-R is a *hybrid architecture*, because it combines the explicit learning of discrete memory structures with Bayesian reasoning supplied by its associative memory structures. However, Anderson's semantic networks are strictly localist,[18] and distributed representations only play a role in external modules, which are not an integral part of the architecture, so its current implementations put it into the realm of symbolic models. Yet, even though it can be

18 Ron Sun (2003, p. 5) characterizes sub-symbolic units as not being individually meaningful. In this sense, the distinction between symbolic and sub-symbolic systems does not allude to whether link weights are strictly binary or real-valued, but whether the links implement a distributed representation of concepts.