OXFORD SERIES ON COGNITIVE MODELS AND ARCHITECTURES

ANATOMY OF THE MIND

Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture

RON SUN

OXFORD

Anatomy of the Mind

OXFORD SERIES ON COGNITIVE MODELS AND ARCHITECTURES

Integrated Models of Cognitive Systems
Edited by Wayne D. Gray

In Order to Learn: How the Sequence of Topics Influences Learning
Edited by Frank E. Ritter, Joseph Nerb, Erno Lehtinen, and Timothy O'Shea

How Can the Human Mind Occur in the Physical Universe?
By John R. Anderson

Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition
By Joscha Bach

The Multitasking Mind
By David D. Salvucci and Niels A. Taatgen

How to Build a Brain: A Neural Architecture for Biological Cognition
By Chris Eliasmith

Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies
Edited by Rosaria Conte, Giulia Andrighetto, and Marco Campennì

Social Emotions in Nature and Artifact
Edited by Jonathan Gratch and Stacy Marsella

Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture
By Ron Sun

Anatomy of the Mind

Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture

Ron Sun

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

First Edition published in 2016

Library of Congress Cataloging-in-Publication Data
Sun, Ron, 1960–
Anatomy of the mind : exploring psychological mechanisms and processes with the Clarion cognitive architecture / Ron Sun.
pages cm. — (Oxford series on cognitive models and architectures)
Includes bibliographical references and index.
ISBN 978–0–19–979455–3
1. Cognitive science. 2. Cognitive neuroscience. 3. Computer architecture. 4. Cognition—Computer simulation. I. Title.
BF311.S8148 2016
153—dc23
2015018557

9 8 7 6 5 4 3 2 1

Printed by Sheridan, USA

Contents

Preface

This book aims to understand psychological (cognitive) mechanisms, processes, and functionalities through a comprehensive computational theory of the human mind, namely, a computational "cognitive architecture," or more specifically, the Clarion cognitive architecture. The goal of this work is to develop a unified framework for understanding the human mind, and within the unified framework to develop process-based, mechanistic explanations of a substantial variety of psychological phenomena.

The book describes the essential Clarion framework, its cognitive-psychological justifications, its computational instantiations, and its applications to capturing, simulating, and explaining various psychological phenomena and empirical data. The book shows how the models and simulations shed light on psychological mechanisms and processes, through the lens of a unified framework (namely, Clarion).

While a forthcoming companion volume to this book will fully describe the technical details of Clarion (along with hands-on examples), the present book concentrates more on a conceptual-level exposition and explanation, but also describes, in a more accessible way, essential technical details of Clarion. It covers those technical details that are necessary for explaining the psychological phenomena discussed in this book.

The following may be considered the features of the present book:

- A scope broader than any other cognitive architecture, pointing to new possibilities for developing comprehensive computational cognitive architectures.

- Integration of multiple approaches and perspectives within this broad scope.
- Exploration of empirical data and phenomena through computational models and simulations, examining a variety of data from a variety of empirical fields.
- Balance of formal modeling and readability (i.e., accessibility to a multidisciplinary readership).

These features were designed with potential readers of the book in mind, who may include (in no particular order): (1) cognitive scientists (especially cognitive modeling researchers, or "computational psychologists" as one might call them) who might be interested in a new theoretical framework, a new generic computational model, as well as new interpretations of data through computational modeling; (2) experimental psychologists who might be interested in new possibilities of interpreting empirical data within a unified framework, new conceptual interpretations (or existing interpretations for that matter) being substantiated through computational modeling, and also new possibilities for further empirical explorations; (3) researchers from adjacent fields who might be interested in work on computational psychology (cognitive modeling) and how such research may shed light on the mind; (4) interested lay readers who might want to explore computational psychology and its implications for understanding the human mind . . . and so on. To put it simply, this book is for those who are interested in exploring and understanding the human mind through computational models that capture and explain empirical data and phenomena in a unified framework.

In fields ranging from cognitive science (especially cognitive modeling), to psychology, to artificial intelligence, and even to philosophy, academic researchers, graduate and undergraduate students, and practitioners of various kinds may have interest in topics covered by this book. The book may be suitable for graduate-level seminars or courses on cognitive architectures or cognitive modeling, but might also be suitable for the advanced undergraduate level.

A little history is in order here. The general ideas of a pair of books (this one and a companion technical book) on Clarion were drawn up in February 2009 after much rumination. I worked more on the ideas for the two books in May of that year. In November, between two trips, I wrote two book proposals. They were submitted to Oxford University

Press in January 2010. After a round of very thorough reviews of the book proposals by the publisher, the contracts for the two books were signed in May 2010. The writing of this book was sporadic and largely put off until the summer of 2011. Since that time, efforts were made to finish the book. The manuscript was sent to the publisher at the end of 2013.

The history of the Clarion cognitive architecture started, of course, much earlier than that. Back in the summer of 1994, the ONR cognitive science basic research program issued a call for proposals, which prompted me to put together a set of ideas that had been brewing in my head. That was the beginning of Clarion. The grant from the ONR program enabled the development and the validation of the initial version of Clarion. During the 1998–1999 academic year, I had my sabbatical leave at the NEC Research Institute. A theoretically oriented book on Clarion took shape during that period, which was subsequently published. Starting in 2000, research grants from ARI enabled the further development of a number of subsystems within Clarion. Then, from 2008 on, new grants from ONR enabled the extension of the work to social simulation and other related topics.

I would like to thank Frank Ritter for his solicitation of thorough reviews of the two book proposals and for his suggestions regarding the organizations of the books. Thanks also go to the eight reviewers of the book proposals for their helpful suggestions. Later I received detailed critiques of the entire book manuscript from Frank Ritter and two anonymous reviewers, whom I gratefully acknowledge as well. I would also like to acknowledge useful discussions that I have had with many colleagues, including Paul Bello, Michael Zenzen, Larry Reid, Jeff White, Jun Zhang, and Deliang Wang, regarding motivation, emotion, personality, ethics, learning, modeling, and so on. I am also indebted to my many collaborators, past and present, including Sebastien Helie, Bob Mathews, Sean Lane, Selmer Bringsjord, Michael Lynch, and their students. I also want to acknowledge my past and current graduate students: Jason Xi Zhang, Isaac Naveh, Nick Wilson, Pierson Fleischer, and others. Some other students contributed to the work on Clarion as well. The work described in this book is theirs as well as mine.

Clarion has been implemented as Java and C# libraries, available at (courtesy of Nick Wilson and Michael Lynch):

http://www.clarioncognitivearchitecture.com

Finally, the work described here has been financially supported, in part, by ONR grants N00014-95-1-0440, N00014-08-1-0068, and N00014-13-1-0342, as well as ARI grants DASW01-00-K-0012 and W74V8H-05-K-0002. Without these forms of support, this work could not have come into being.

Ron Sun
Troy, New York

Anatomy of the Mind

1

What Is A Cognitive Architecture?

In this chapter, as an introduction to what is to be detailed in this book, I will attempt to justify the endeavor of developing a generic computational model (theory) of the mind (i.e., a computational cognitive architecture), through addressing a series of questions. Then I will discuss a few issues fundamental to such an endeavor.

1.1. A Theory of the Mind and Beyond

Before embarking on this journey, it might help to make clear at the outset that what is to be described and discussed in the present book, including concepts, theories, models, and simulations, is centered on a particular theoretical framework—namely, the Clarion framework. It is worth noting that Clarion, in its full-fledged form, is a generic and relatively comprehensive theory of the human mind,[1] along with a computational implementation of the theory. It is thus a computational "cognitive

1. "Mind" is a complex notion. Rather than engaging in a philosophical discourse on the notion, the focus here is instead on mechanisms and processes of the mind. In turn, "mechanism" here refers to physical entities and structures and their properties that give rise to certain characteristics of the mind. Although living things often appear to have certain characteristics that have no counterpart in the physical universe, one may aim to go beyond these appearances (Thagard, 1996).

architecture" as is commonly referred to in cognitive science, cognitive psychology, or more generally in the "cognitive sciences".[2] In general, a cognitive architecture is a broad domain-generic cognitive-psychological model implemented computationally.

Clarion has been in continuous development for a long time, at least since 1994 (although its predecessors have had a longer history). It has been aimed to capture, explain, and simulate a wide variety of cognitive-psychological phenomena within its unified framework, thus leading (hopefully and ultimately) to unified explanations of psychological (and even other related) phenomena (as advocated by, e.g., Newell, 1990). The exact extent of cognitive-psychological phenomena that have been captured and explained within its framework will be discussed in detail in subsequent chapters. It is not unreasonable to say that Clarion constitutes a (relatively) comprehensive theory of the mind (or at least an initial version of such a theory).

In fact, Clarion, within itself, contains several different kinds of theories. First, it contains a core theory of the mind at a conceptual level. It posits essential theoretical distinctions such as implicit versus explicit processes, action-centered versus non-action-centered processes, and so on, as well as their relationships (Sun, 2002, 2012). With these distinctions and other high-level constructs, it specifies a core theory of the essential structures, mechanisms, and processes of the mind, at an abstract, conceptual level (Sun, Coward, and Zenzen, 2005).

Second, it also contains a more detailed (but still generic) computational model implementing the abstract theory. This implementation constitutes what is usually referred to as a computational cognitive architecture: that is, a generic computational cognitive (i.e., psychological) model describing the architecture of the mind, which by itself also constitutes a theory of the mind, albeit at a more detailed and computational level (as will be argued later; see also Sun, 2009b).

2. In the narrow sense, "cognition" refers to memory, learning, concepts, decision making, and so on—those aspects of the individual mind that are not directly related to motivation, emotion, and the like. In the broadest sense, it may refer to all aspects of the individual mind, especially when methods and perspectives from contemporary cognitive science are used in studying these aspects. In the latter case, I often use a hyphenated form, "cognition-psychology", to make it clear. However, the plural form, "cognitive sciences," is often used to refer to all fields of cognitive, behavioral, and psychological sciences, applying the broadest sense of the term. Similarly, in the term "cognitive architecture," the word "cognitive" should be interpreted in the broadest sense.

Third, with the generic computational cognitive architecture, one may construct specific models and simulations of specific psychological phenomena or processes. That is, one may "derive" specific computational models (namely, specific computational theories) for specific psychological phenomena or processes, from the generic computational model (theory). So, the generic theory leads to specific theories.

Clarion encompasses all of the above simultaneously. Thus, it synthesizes different types of theories at different levels of theoretical abstraction (Sun, 2009b). Below I will refer, alternately, to Clarion in these different senses, at different levels of abstraction, as appropriate.

1.2. Why Computational Models/Theories?

Why would one want computational models for the sake of understanding the human mind? Why are computational models useful exactly?

Generally speaking, models of various forms and complexities may be roughly categorized into computational, mathematical, and verbal-conceptual varieties (Sun, 2008). Computational models present algorithmic descriptions of phenomena, often in terms of mechanistic and process details. Mathematical models present (often abstract) relationships between variables using mathematical equations. Verbal-conceptual models describe entities, relations, or processes in informal natural languages (such as English). A model, regardless of its genre, might often be viewed as a theory of whatever phenomena that it purports to capture. This point has been argued extensively before (by, e.g., Newell, 1990 and Sun, 2009b).

Although each of these types of models has its role to play, I am mainly interested in computational modeling. The reason for this preference is that, at least at present, computational modeling appears more promising in many respects. It offers the expressive power that no other approach can match, because it provides a wider variety of modeling techniques and methodologies. In this regard, note that mathematical models may be viewed as a subset of computational models, because normally they can lead readily to computational implementations (even though some of them may be sketchy, not covering sufficient mechanistic or process details). Computational modeling also supports practical applications (see, e.g., Pew and Mavor, 1998; Sun, 2008).

Computational models are mostly mechanistic and process oriented. That is, they are mostly aimed at answering the questions of how human performance comes about, by what psychological structures, mechanisms, and processes, and in what ways.[3] The key to understanding cognitive-psychological phenomena is often in fine details, which computational modeling can describe and illuminate (Newell, 1990; Sun, 2009b). Computational models provide algorithmic specificity: detailed, exactly specified, and carefully worked-out steps, arranged in precise and yet flexible sequences. Thus, they provide clarity and precision (see, e.g., Sun, 2008).

Computational modeling enables and, in fact, often forces one to think in terms of mechanistic and process details. Instead of verbal-conceptual theories, which may often be vague, one has to think clearly, algorithmically, and in detail when dealing with computational models/theories. Computational models are therefore useful tools. With such tools, researchers must specify a psychological mechanism or process in sufficient detail to allow the resulting models to be implemented on computers and run as simulations. This requires that all elements of a model (e.g., all its entities, relationships, and so on) be specified exactly. Thus it leads to clearer, more consistent, more mechanistic, more process-oriented theories. Richard Feynman once put it this way: "What I cannot create, I do not understand." This applies to the study of human cognition-psychology. To understand is to create, in this case on a computer at least.

Computational models may be necessary for understanding a system as complex and as internally diverse as the human mind. Pure mathematics, developed mainly for describing the physical universe, may not be sufficient for understanding a system as different as the human mind. Compared with theories developed in other disciplines (such as physics), computational modeling of the mind may be mathematically less "elegant", but the human mind itself may be inherently less mathematically elegant when compared with the physical universe (as argued by, e.g., Minsky, 1985). Therefore, an alternative form of theorizing may be necessary—a form that is more complex, more diverse, and more algorithmic in nature. Computational modeling provides a viable way

3. It is also possible to formulate so called "product theories", which provide a functional account of phenomena but do not commit to a particular psychological mechanism or process. Thus, product theories can be evaluated mainly by product measures. One may also term product theories *black-box theories* or *input-output theories*.

of specifying complex and detailed theories of cognition-psychology. Therefore, they may be able to provide unique explanations and insights that other experimental or theoretical approaches cannot easily provide.

A description or an explanation in terms of computation that is performed in the mind/brain can serve either as a fine-grained specification of cognitive-psychological processes underlying behavior (roughly, the mind), or as an abstraction of neurobiological and neurophysiological data and discoveries (roughly, the brain), among other possibilities that may also exist. In general, it is not difficult to appreciate the usefulness of a computational model in this regard, in either sense, especially one that summarizes a body of data, which has been much needed in psychology and in neuroscience given the rapid growth of empirical data.

In particular, understanding the mind (at the psychological level) through computational modeling may be very important. One would naturally like to know more about both the mind and the brain. So far at least, we still know little about the biology and physiology of the brain, relatively speaking. So for this reason (and others), we need a higher level of abstraction; that is, we need to study the mind at the psychological level in order to make progress toward the ultimate goal of fully understanding the mind and the brain.

Trying to fully understand the human mind purely from observations of human behavior (e.g., strictly through behavioral experiments) is likely untenable (except perhaps for small, limited task domains). The rise and fall of behaviorism is a case in point. This point may also be argued on the basis of analogy with the physical sciences (as was done in Sun, Coward, and Zenzen, 2005). The processes and mechanisms of the mind cannot be understood purely on the basis of behavioral experiments, which often amount to tests that probe relatively superficial features of human behavior, further obscured by individual and cultural differences and other contextual factors. It would be extremely hard to understand the human mind in this way, just like it would be extremely hard to understand a complex computer system purely on the basis of testing its behavior, if one does not have any prior ideas about the inner workings and theoretical underpinnings of that system (Sun, 2007, 2008, 2009b).

Experimental neuroscience alone may not be sufficient either, at least for the time being. Although much data has been gathered from empirical work in neuroscience, there is still a long way to go before all the details of the brain are identified, let alone the psychological functioning on that basis. Therefore, as argued earlier, at least at present, it is important to

understand the mind/brain at a higher level of abstraction. Moreover, even when we finally get to know all the minute details of the brain, we would still need a higher-level, yet precise, mechanistic, process-based understanding of its functioning. Therefore, we still need a higher level of theorizing. In an analogous way, the advent of quantum mechanics did not eliminate the need for classical mechanics. The progress of chemistry was helped by the discoveries in physics, but chemistry was not replaced by physics. It is imperative that we also investigate the mind at a higher level of abstraction, beyond neuroscience. Computational modeling has its unique, indispensable, and long-term role to play, especially for gaining conceptually clear, detailed, and principled understanding of the mind/brain.

It might be worth mentioning that there have been various viewpoints concerning the theoretical status of computational modeling. For example, many believed that a computational model (and computational simulation on its basis) may serve as a generator of phenomena and data. That is, they are useful media for hypothesis generation. In particular, one may use simulation to explore process details of a psychological phenomenon. Thus, a model is useful for developing theories, constituting a theory-building tool. A related view is that computational modeling and simulation are suitable for facilitating a precise instantiation of a preexisting verbal-conceptual theory (e.g., through exploring possible details for instantiating the theory) and consequently detailed evaluations of the theory against data. These views, however, are not incompatible with a more radical position (e.g., Newell 1990; Sun 2009b) that a computational model may constitute a theory by itself. It is not the case that a model is limited to being built on top of an existing theory, being applied for the sake of generating data, being applied for the sake of validating an existing theory, or being used for the sake of building a future theory. According to this more radical view, a model may be viewed as a theory by itself. In turn, algorithmic descriptions of computational models may be considered just another language for specifying theories (Sun, 2009b; Sun, 2008).[4] The reader is referred to Sun (2009b) for a more in-depth discussion of this position.

4. Constructive empiricism (van Fraasen, 1980) may serve as a philosophical foundation for computational cognitive modeling, compatible with the view of computational models as theories (Sun 2009b).

In summary, computational models (theories) can be highly useful to psychology and cognitive science, when viewed in the light above (and when the issues discussed below are properly addressed).

1.3. Questions about Computational Models/Theories

There are, of course, many questions that one can, and should, ask about any computational model before "adopting" it in any way.

One important question about any particular computational model is this: how much light can it really shed on the phenomena being modeled? There are a number of aspects to this question, for instance:

- Do the explanations provided by the computational model capture accurately human "performance" (in a Chomskian sense; Chomsky, 1980)? That is, does it capture and explain sufficiently the subtleties exhibited in the empirical data? If an explanation is devoid of "performance" details as observed in empirical data, it will be hard to justify the appropriateness of such an explanation, especially when there are other possible ways of describing the data.[5]
- Does the model take into consideration higher-level or lower-level constraints (above or below the level of the model in question)? There are usually many possible models/theories regarding some limited data. Higher-level or lower-level considerations, among other things, may be used to narrow down the choices.
- Does the model capture in a detailed way psychological mechanisms and processes underlying the data? If a model lacks mechanistic, process-oriented details, it may be less likely to bring new insights into explaining the dynamics underlying the data.
- Do the primitives (entities, structures, and operations) used in the model provide some descriptive power and other advantages over and above other possible ways of describing human behavior and performance (but without being overly generic)?

5. This is not the case for Noam Chomsky's theory of language, which thus serves as an exception.

- Does the model provide a basis for tackling a wide set of cognitive-psychological tasks and data? If a model is insufficient in terms of breadth of coverage, it cannot claim to be a "general" theory.

It should be noted that, in relation to the issue of generality, one should be aware of the danger of over-generality. That is, a model might be so under-constrained that it may match practically any possible data, realistic or unrealistic. To address this problem, many simulations in a wide range of domains are needed, in order to narrow down choices and to constrain parameter spaces (more on this in Chapter 8).

From the point of view of the traditional cognitive science, a model/theory at the computational or knowledge level (in Marr's [1982] or Newell's [1990] sense) can provide a formal language for describing a range of cognitive-psychological tasks. Indeed, in the history of cognitive science, some high-level formal theories were highly relevant (e.g., Chomsky's theory of syntactic structures of language). So, a further question is:

- How appropriate is the model/theory in terms of providing a "formal language" for a broad class of tasks or data? Does it have realistic expressiveness (sufficient for the target tasks or data, but not much more or less) and realistic constraints (of various types, at various levels)?

Furthermore, what is more important than a formal (e.g., mathematical or computational) language for describing cognition-psychology is the understanding of the "architecture" of the mind, especially in a mechanistic (computational) sense. That is, one needs to address the following question:

- How do different components of the mind interact and how do they fit together? Correspondingly, how do different components of a computational model/theory interact and how do these different components fit together, instead of just a mere collection of limited models?

Studying architectural issues may help us to gain new insight, narrow down possibilities, and constrain the components involved.

Moreover, different components and different functionalities of the mind, for example, perception, categorization, concepts, memory,

decision making, reasoning, problem solving, planning, communication, action, learning, metacognition, and motivation, all interact with and depend on each other. Their patterns of interaction change with changing task demands, growing personal experiences, varying sociocultural contexts and milieus, and so on. Some argue that cognition-psychology represents a context-sensitive, dynamic, statistical structure that, on the surface at least, changes constantly—a structure in perpetual motion. However, complex dynamic systems may be attributed to its constituting elements. Thus, one may strive for a model that captures the dynamics of cognition-psychology through capturing its constituting elements and their interaction and dependency. So, an important question is:

- How does a model/theory account for the dynamic nature of cognition-psychology?

Finally, one has to consider the cost and benefit of computational modeling:

- Is the complexity of a model/theory justified by its explanatory utility (considering all the questions above)?

These questions cannot be addressed in abstraction. My specific answers to them, in the context of Clarion, will emerge in subsequent chapters, as details of Clarion emerge in these chapters.

1.4. Why a Computational Cognitive Architecture?

Among different types of computational cognitive-psychological models/theories, computational cognitive architectures stand out. A computational cognitive architecture, as commonly termed in cognitive science, is a broadly scoped, domain-generic cognitive-psychological model, implemented computationally, capturing the essential structures, mechanisms, and processes of the mind, to be used for broad, multiple-level, multiple-domain analysis of behavior (e.g., through its instantiation into more detailed computational models or as a general framework; Newell, 1990; Sun, 2007).

Let us explore this notion of cognitive architecture with a comparison. The architecture for a building consists of its overall structural design and major constituting structural elements such as external walls, floors, roofs,

stairwells, elevator shafts, and so on. Furniture can be easily rearranged or replaced and therefore may not be part of the architecture. By the same token, a cognitive architecture includes overall structures, essential divisions of modules (e.g., subsystems), essential relations between modules, basic representations and algorithms within modules, and a variety of other major aspects (Sun, 2007; Langley, Laird & Rogers, 2009). In general, a cognitive architecture includes those aspects that are relatively invariant across time, domains, and individuals. It deals with them in a structurally and mechanistically well-defined way.

A cognitive architecture can be important to understanding the human mind. It provides concrete computational scaffolding for more detailed modeling and exploration of cognitive-psychological phenomena and data, through specifying essential computational structures, mechanisms, and processes. That is, it facilitates more detailed modeling and exploration of the mind. As discussed earlier, computational cognitive modeling explores cognition-psychology through specifying computational models of cognitive-psychological mechanisms and processes. It embodies descriptions of cognition-psychology in computer algorithms and program codes, thereby producing runnable models. Detailed simulations can then be conducted. In this undertaking, a cognitive architecture can be used as the unifying basis for a wide range of modeling and simulation. Note that here I am mainly referring to psychologically oriented cognitive architectures (as opposed to software engineering oriented cognitive architectures, which are quite different in terms of purpose).

A cognitive architecture serves as an initial set of (relatively) generic assumptions that may be applied in further modeling and simulation. These assumptions, in reality, may be based on empirical data, philosophical arguments, or computational considerations. A cognitive architecture is useful and important because it provides a (relatively) comprehensive yet precise foundation that facilitates further modeling in a wide variety of domains (Cooper, 2007).

In exploring cognitive-psychological phenomena, the use of cognitive architectures forces one to think in terms of mechanistic and process-oriented details. Instead of using often vague and underspecified verbal-conceptual theories, cognitive architectures force one to think more clearly. Anyone who uses cognitive architectures must specify a cognitive-psychological mechanism or process in sufficient detail to allow the resulting models to run as simulations. This approach encourages more detailed and clearer theories. It is true that more specialized,

narrowly scoped computational models may also serve this purpose, but they are not as generic and as comprehensive. Consequently, they are not as generally useful. Cognitive architectures are thus crucial tools (Pew and Mavor, 1998; Sun, 2007).

A cognitive architecture may also provide a deeper level of explanation (Sun, 2007). Instead of a model specifically designed for a specific task (which is often ad hoc), a cognitive architecture naturally encourages one to think in terms of the mechanisms and processes available within a generic model that are not specifically designed for a particular task, and thereby to generate explanations of the task that are not centered on superficial, high-level features of the task (as often happens with specialized, narrowly scoped models)—that is, to generate explanations of a deeper kind. To describe a task in terms of available mechanisms and processes of a cognitive architecture is to generate explanations based on primitives of cognition-psychology envisioned in the cognitive architecture, thereby leading to deeper explanations.

Because of the nature of such deeper explanations, this approach is also more likely to lead to unified explanations for a wide variety of data and phenomena, because potentially a wide variety of tasks, data, and phenomena can be explained on the basis of the same set of primitives provided by the same cognitive architecture (Sun, 2007). Therefore, a cognitive architecture is more likely to lead to a unified, comprehensive theory of the mind, unlike using more specialized, narrowly scoped models (Newell, 1990).

Although the importance of being able to reproduce the nuances of empirical data is evident, broad functionalities in cognitive architectures are even more important (Newell, 1990). The human mind needs to deal with all of the necessary functionalities: perception, categorization, memory, decision making, reasoning, planning, problem solving, communication, action, learning, metacognition, motivation, and so on. The need to emphasize generic models capable of broad functionalities arises also because of the need to avoid the myopia often resulting from narrowly scoped research.

For all of these reasons above, developing cognitive architectures is an important endeavor in cognitive science. It is of essential importance in advancing the understanding of the human mind (Sun, 2002, 2004, 2007). Existing cognitive architectures that have served this purpose include ACT-R, Soar, Clarion, and a number of others (see, e.g., Taatgen and Anderson, 2008 for a review).

In addition, cognitive architectures also, in a way, support the goal of general AI, that is, building artificial systems that are as capable as human beings. In relation to building intelligent systems, a cognitive architecture may provide the underlying infrastructure, because it may include a variety of capabilities, modules, and subsystems that an intelligent system needs. On that basis, application systems may be more readily developed. A cognitive architecture carries with it theories of cognition-psychology and understanding of intelligence gained from studying the human mind. In a way, cognitive architectures reverse engineer the only truly intelligent system around—the human mind. Therefore, the development of intelligent systems on that basis may be more cognitively-psychologically grounded, which may be advantageous in some circumstances. The use of cognitive architectures in building intelligent systems may also facilitate the interaction between humans and artificially intelligent systems because of the relative similarity between humans and cognitively-psychologically based intelligent systems. It was predicted a long time ago that "in not too many years, human brains and computing machines will be coupled together very tightly and the resulting partnership will think as no human brain has ever thought . . ." (Licklider, 1960). Before that happens, a better understanding of the human mind is needed, especially a better understanding in a computational form.

There are, of course, questions that one should ask about cognitive architectures, in addition to or instantiating questions about computational modeling in general as discussed earlier. For instance, a cognitive architecture is supposed to include all essential psychological capabilities and functionalities. As mentioned before, those functionalities may include perception, categorization, memory, decision making, reasoning, problem solving, communication, action, and learning. They may involve all kinds of representation (in a broad sense). There are also motivational and metacognitive processes. However, currently, most cognitive architectures do not yet support all of these functionalities, at least not fully. So, what is minimally necessary? How should these functionalities interact? To what extent are they separate? And so on. There are no simple answers to these questions, but they will be addressed along the way in this book.

In this regard, a question concerning any capability in a cognitive architecture is whether the cognitive architecture includes that capability as an integral part or whether it includes sufficient basic functionalities that allow the capability to emerge or to be implemented later on.

This may be determined by what one views as an integral part of a cognitive architecture and what one views as a secondary or derived capability. Sun (2004) provides a discussion of the relation between a cognitive architecture and the innate structures in the human mind and the notion of minimality in a cognitive architecture. These ideas may help to sort out what should or needs to be included in a cognitive architecture (Sun, 2004). The outcomes of the deliberation on this and other questions will be presented in the subsequent chapters.

1.5. Why Clarion?

Among existing cognitive architectures, why should one choose Clarion? In a nutshell, one might prefer Clarion, for (the totality of) the following reasons:

- Clarion is a cognitive architecture that is more comprehensive in scope than most other cognitive architectures in existence today (as will become clear later).
- Clarion is psychologically realistic to the extent that it has been validated through simulating and explaining a wide variety of psychological tasks, data, and phenomena (as detailed in chapters 5, 6, and 7).
- Its basic principles and assumptions have been extensively argued for and justified, in relation to a variety of different types of evidence (as detailed in chapters 2, 3, and 4).
- It has major theoretical implications, as well as some practical relevance. It has provided useful explanations for a variety of empirical data, leading to a number of significant new theories regarding psychological phenomena (e.g., Sun, Slusarz, & Terry, 2005; Helie & Sun, 2010).
- In addition to addressing problems at the psychological level, it has also taken into account higher levels, for example, regarding social processes and phenomena, as well as lower levels (Sun, Coward, & Zenzen, 2005).

More specifically, Clarion has been successful in computationally modeling, simulating, accounting for, and explaining a wide variety of psychological data and phenomena. For instance, a number of well-known

skill-learning tasks have been simulated using Clarion that span the entire spectrum ranging from simple reactive skills to complex cognitive skills. The simulated tasks, for example, include serial reaction time tasks, artificial grammar learning tasks, dynamic process control tasks, alphabetical arithmetic tasks, and Tower of Hanoi (e.g., Sun, Slusarz, & Terry, 2005; Sun, 2002). In addition, extensive work has been done in modeling complex and realistic skill-learning tasks that involve complex sequential decision making (Sun et al., 2001). Furthermore, many other kinds of tasks not usually dealt with by cognitive architectures—reasoning tasks, social simulation tasks, as well as metacognitive and motivational tasks—have been tackled by Clarion. While accounting for various psychological tasks, data, and phenomena, Clarion provides explanations that shed new light on underlying cognitive-psychological processes. See, for example, Sun et al. (2001), Sun, Slusarz, and Terry (2005), Sun, Zhang, and Mathews (2006), and Helie and Sun (2010) for various examples of such simulations and explanations.

These simulations, more importantly, provided insight that led to some major new theories concerning a number of important psychological functionalities. Some new theories resulting from Clarion include:

- The theory of bottom-up learning (from implicit to explicit learning), as developed in Sun et al. (2001).
- The theory of the implicit-explicit interaction and their synergistic effects on skill learning, as developed in Sun, Slusarz, and Terry (2005).
- The theory of creative problem solving, as described in Helie and Sun (2010).
- The theory of human motivation and its interaction with cognition, as described in Sun (2009), as well as in related simulation papers (e.g., Wilson, Sun, & Mathews, 2009; Sun & Wilson, 2011; Sun & Wilson, 2014)
- The theory of human reasoning (based on implicit and explicit representation and their interaction), as developed in Sun (1994, 1995) and Sun and Zhang (2006).

These theories are standalone, conceptual-level theories of psychological phenomena. However, these theories are also an integral part of Clarion. They have been computationally instantiated. They have led not only to

numerical (quantitative) simulations but also to major qualitative (theoretical) predictions.

I should mention here that two meta-principles have guided the development of this cognitive architecture: (a) completeness of functionalities (to include as many functionalities as possible), but (b) parsimony of mechanisms (to reduce the number of distinct mechanisms and their complexity as much as possible). Or to put it another way, the goal for Clarion has been: maximum scope and minimum mechanism. That goal and the associated meta-principles have led to the aforementioned theories and explanations by Clarion.

Given all of the above, Clarion is worthy of further exploration and examination. In particular, its comprehensive scope should be examined in more detail. Thus a book-length treatment is required.

1.6. Why This Book?

Although a substantial number of articles, including journal and conference papers, have been published on Clarion and its modeling of psychological data of various kinds, there is currently no one single volume that contains all of the information, especially not in a unified and accessible form. Therefore, it seems a good idea to put together a single volume for the purpose of cataloguing and explaining in a unified and accessible way what has been done with regard to Clarion and why it might be of interest.

Furthermore, a book may contain much more material than a typical journal or conference paper. It may describe not only details of Clarion but also many detailed models of psychological phenomena based on Clarion. It may summarize materials published previously, in addition to new materials. A book may also provide theoretical and meta-theoretical discussions of issues involved. Above all, a book may provide a gentler introduction to Clarion and its exploration of psychological mechanisms and processes, which may be of use to some readers.

The present book will present a unified (albeit preliminary and still incomplete) view of the human mind, and interpret and explain empirical data on the basis of that view. The focus will be on broad interpretations of empirical data and phenomena, emphasizing unified explanations of a wide variety of psychological tasks and data. Thus

exact parameter values and other minute technical details will be minimized.

For the sake of clarity, I will proceed in a hierarchical fashion. In other words, there will be a series of progressively more detailed descriptions. First, a high-level conceptual sketch will be given; then a more detailed description will be provided. After that, there will be an even more detailed, more technical description. (However, the most technically exact and complete description, with a code library, can be found in a forthcoming companion technical book on Clarion.) In this way, the reader may stop at any time, up to the level where he or she feels comfortable.

I will start with the overall Clarion framework and then move on to individual components or aspects. To achieve clarity, I will limit the amount of details discussed to only those that are minimally necessary. (Fortunately, the technical book will provide full technical specifications.) With regard to technical details, especially in relation to simulations, I will have to strike a balance between conceptual clarity and technical specificity. Of course, both are important. To achieve conceptual clarity, a high-level conceptual explanation will be provided. To achieve some technical specificity, a more technical (computational) description or explanation will also be provided, corresponding to the high-level conceptual explanation.

1.7. A Few Fundamental Issues

To start, I will quickly sketch a few foundational issues. My stands on these issues form the meta-theoretical basis of Clarion. (Details of the cognitive architecture will be explained in subsequent chapters.)

1.7.1. Ecological-Functional Perspective

The development of a cognitive architecture needs to take into consideration of what I have called the ecological-functional perspective. As discussed in Sun (2012) and Sun (2002), the ecological-functional perspective includes a number of important considerations on human cognition-psychology, especially in relation to ecological realism of

cognitive-psychological theories or models. They may be expressed as dictums such as:

- taking into account ecological niches (evolutionarily or at the present), and focusing attention on characteristics of everyday activities that are most representative of the ecological niches (Sun, 2002; more later);
- taking into account the role of function, because cognitive-psychological characteristics are often, if not always, functional, useful in some way for everyday activities within an ecological niche;
- taking into account cost-benefit trade-offs of cognitive-psychological characteristics (such as implicit versus explicit processes)[6], as psychological characteristics are often selected based on cost-benefit considerations (evolutionarily or at the present).

In particular, these dictums imply that human cognition-psychology is mostly activity-based, action-oriented, and embedded in the world. They also seem to point toward implicit (subconscious or unconscious) psychological processes, as opposed to focusing exclusively on explicit processes. Humans often interact with the world in a rather direct and unmediated way (Heidegger, 1927; Dreyfus, 1992; Sun, 2002).

These dictums, serving as meta-heuristics for developing cognitive architectures, will become clearer in the next chapter, when the justifications for the essential framework of Clarion are discussed.

1.7.2. Modularity

Fodor (1983) argued that the brain/mind was modular and its modules worked largely independently and communicated only in a limited way. However, evidence to the contrary has accumulated that modules and subsystems in the brain/mind may instead be more richly interconnected, anatomically and functionally (Damasio, 1994; Bechtel, 2003).

Nevertheless, starting off with a modular organization might make the task of understanding the architecture of the human mind more tractable.

6. For instance, compared with implicit processes, explicit processes may be more precise but may be more effortful. See more discussions in Chapter 3.

Connections, communications, and interactions, if necessary, may be added subsequently. At a minimum, some cognitive-psychological functionalities do appear to be specialized and somewhat separate from others (in a sense). They may be so either because they are functionally encapsulated (their knowledge, mechanisms, and processes do not transfer easily into other domains) or because they are physically (neurophysiologically) encapsulated. Modularity can be useful functionally, for example, to guarantee efficiency or accuracy of important or critical behaviors and routines (whether they are a priori or learned), or to facilitate parallel operations of multiple processes (Sun, 2004). Hence we start with a (circumscribed) modular view.

1.7.3. Multiplicity of Representation

With modularity (i.e., with the co-existence of multiple modules), multiple different representations (either in terms of form or in terms of content) may co-exist.

Here I use the term "representation" to denote any form of internal encoding, either explicitly and individually encoded or embodied/enmeshed within a complex mechanism or process. Thus this notion of "representation" is not limited to explicit, individuated symbolic entities and their structures (as often meant by "representationalism"). Because it is not limited to symbolic forms, it includes, for example, connectionist encoding, dynamic system content, and so on. So the term should be interpreted broadly here.

In terms of representational form, there are, for example, symbolic-localist representation and distributed connectionist representation. Symbolic-localist representation implies representing each unique concept by a unique basic representational entity (such as a node in a network). Distributed representation involves representing each concept by an activation pattern over a shared set of nodes in a network (Rumelhart et al., 1986). Different forms of representations have different computational characteristics: for example, crisp versus graded, rule-based versus similarity-based, one-shot learning versus incremental learning, and so on, as will be discussed in more detail later.

In terms of representational content, there may be the following types: procedural representation, declarative representation, metacognitive representation, motivational representation, and so on. Each of these types

is necessary for a full account of the human mind. In subsequent chapters when I discuss each of these types in turn, I will present arguments why each of them is needed. Each type may in turn involve multiple representational forms within.

On the other hand, one may question why an individual needs multiple representational forms after all. There are a number of potential advantages that may be gained by involving multiple representational forms. For example, in incorporating both symbolic-localist and distributed representation (for capturing explicit and implicit knowledge, respectively, as will be detailed later), one may gain

- synergy in skill learning from dual procedural representation
- synergy in skill performance from dual procedural representation
- synergy in reasoning from dual declarative representation

and so on. These advantages have been demonstrated before in previous publications; I will elaborate on these advantages in later chapters when I revisit these points.

1.7.4. Dynamic Interaction

In a cognitive architecture, various modules (in the previously discussed sense) have to work with each other to accomplish psychological functioning. Modules of different kinds and sizes (e.g., subsystems and components within each subsystem) interact with each other dynamically.

At the highest level, the interaction among subsystems may include metacognitive monitoring and regulation of other processes (i.e., the interaction between the metacognitive subsystem and the other subsystems). The interaction among subsystems may also involve motivated action decision making (i.e., the interaction between the motivational subsystem and the action-centered subsystem). Within each subsystem, many component modules exist and they also interact with each other, necessary for accomplishing cognitive-psychological functioning.

Note that these characteristics may not have been sufficiently captured by most existing cognitive-psychological models (including cognitive architectures). Compared with these other models, Clarion is unique in terms of containing (well-developed, built-in) motivational constructs and (well-developed, built-in) metacognitive

constructs. These are not commonly found in existing cognitive architectures. Nevertheless, I believe that these features are crucial to a cognitive architecture because they capture important or indispensable elements of the human mind, necessary in the interaction between an individual and his or her physical and social world (Sun, 2009). Details will be presented in subsequent chapters.

1.8. Concluding Remarks

So far I have covered only some preliminary ideas, which are necessary background regarding cognitive architectures. The questions that have been addressed include: Why should one use computational modeling for studying cognition-psychology? Why should one use cognitive architectures among other computational models? Why should one use the Clarion cognitive architecture, among other possible cognitive architectures? And other questions and issues.

More importantly, the basic "philosophy" in regard to a number of fundamental issues has been outlined. In particular, the principles of modularity, multiplicity of representation, and dynamic interaction (include that among motivation, cognition, and metacognition) are of fundamental importance to Clarion.

The rest of the book is divided into eight chapters. They include three chapters for presenting various theoretical, conceptual, and technical aspects of Clarion, three chapters on various simulations using Clarion, and additional materials in the remaining two chapters.

Finally, a note for the interested reader: for general surveys, discussions, and comparisons of computational cognitive architectures in the context of cognitive-psychological modeling, covering other well-known cognitive architectures (such as ACT-R and Soar), see Pew and Mavor (1998), Ritter et al. (2003), Sun (2006), Chong, Tan, and Ng (2007), Taatgen and Anderson (2008), Langley et al. (2009), Thórisson and Helgasson (2012), Helie and Sun (2014b), among other existing publications (see also Chapter 9).

2

Essential Structures of the Mind

In this chapter, I introduce the basic framework (i.e., the relatively abstract conceptual-level theory) of Clarion, and discuss the justifications for this framework.

In a way, this chapter presents a worldview—an essential, overarching framework for understanding the mind. One should view it as the more abstract general theory of Clarion, as opposed to the more detailed computational theory of Clarion, which will be presented in chapters 3 and 4, or as opposed to the specific computational simulation models derived from Clarion, which will be presented in chapters 5, 6, and 7.

Below I will first review and justify the essential desiderata that have been driving the development of Clarion. Then, on that basis, the overall structure of Clarion will be sketched.

2.1. Essential Desiderata

As has been characterized earlier, Clarion is a computational cognitive architecture: it is a generic and comprehensive model of cognitive-psychological structures, mechanisms, processes, functionalities,

and so on, specified and implemented computationally. As such, it needs substantial justifications.

Clarion has indeed been justified extensively on the basis of empirical data (see, e.g., Sun, 2002, 2003; see also Sun, Merrill, & Peterson, 2001; Sun, Slusarz, & Terry, 2005; Helie & Sun, 2010), as well as theoretical (fundamental, philosophical) considerations. In particular, a number of essential (philosophical and psychological) desiderata have been central to the conception of the framework. These essential desiderata include those described below (along with others described elsewhere, e.g., in Sun, 2002, 2004, 2012). Together, they present a situated/embodied view of the mind in a generalized sense (Sun, 2013b), consistent with the ecological-functional perspective discussed in Chapter 1, in addition to the other considerations outlined there (e.g., representational multiplicity, modularity, and dynamic interaction).

Sequentiality. Everyday activities are sequential: they are often carried out one step at a time. Temporal structures are essential to such activities and form the basis of behaviors in different circumstances (Sun, 2002).

Routineness. Everyday activities are largely made up of reactive routines (skills), or habitual sequences of behavioral responses (on a moment-to-moment basis mostly). They are, generally speaking, gradually formed and subject to continuous modification (with the possible exception of some innate routines or instincts). Therefore, human everyday activities may be viewed as comprised of forming, adapting, and following routines (or skills; Sun, 2002; Tinbergen, 1951; Timberlake and Lucas, 1989).

Trial-and-error adaptation. Learning of reactive routines (and other behaviors) is often a trial-and-error process. Such learning has been variously studied under the rubric of law of effect, classical conditioning, instrumental conditioning, probability learning, and implicit learning (Reber, 1989). Such learning is essential to human everyday activities (Sun, 2002).

Implicit versus explicit processes. Reactive routines are mostly implicit. Implicit processes are (relatively) inaccessible and "holistic," while explicit processes are more accessible and more precise (e.g., Reber, 1989). These two types interact with each other. This dichotomy is related to some other well-known dichotomies: the conscious versus the unconscious, the conceptual versus the subconceptual, and so on (Evans & Frankish, 2009; Sun, 2002).

Synergistic interaction. It was hypothesized that one reason for having these two types of processes, implicit and explicit, was that these processes worked together synergistically, supplementing and complementing each other in a variety of ways (Sun, Slusarz, & Terry, 2005). These two types have qualitatively different characteristics, thus often generating better overall results when they interact (Sun, 2002).

Bottom-up and top-down learning. The interaction between implicit and explicit processes allows for a gradual transfer of knowledge (memory) from one type to the other (besides separate, standalone learning within each type). Learning resulting from the implicit-explicit interaction includes top-down learning (explicit learning first and implicit learning on that basis) and bottom-up learning (implicit learning first and explicit learning on that basis; Sun, 2002).

Procedural versus declarative processes. Procedural processes are specifically concerned with actions in various circumstances (i.e., how to do things). Declarative processes are not specifically concerned with actions but are more about objects, persons, events, and so on, in generic terms. This distinction has provided useful insight in interpreting a wide range of psychological data in the past (Proctor & Dutta, 1995). Furthermore, the procedural-declarative distinction is orthogonal to the implicit-explicit distinction (based on empirical evidence as summarized in Sun, 2012).

Motivational control. A full account of behavior must address why one does what one does.[1] Hence motivational processes need to be understood (Sun, 2009). An individual's essential motivations (needs) arise prior to deliberative cognition (Sun, 2009) and are the foundation of cognition and action. In a way, cognition has evolved to serve the essential needs (motives) of an individual, and bridges the needs (motives) of an individual and his or her environments.

Metacognitive control. Metacognition regulates cognition. For need fulfillment, metacognitive monitoring and regulation are necessary. They help to set goals, to assess progresses, and to adopt or change various parameters and strategies (large or small) for goal achievement. The importance of metacognition has been well established (see, e.g., Reder, 1996, and Sun & Mathews, 2012).

1. Simply saying that one chooses actions to maximize rewards or reinforcement is not sufficient. It leaves open the question of what determines them.

Table 2.1. Fundamental issues relevant to Clarion (see Chapter 1 for details).

Ecological-functional perspective
Modularity
Multiplicity of representation
Dynamic interaction

Table 2.2. Some essential desiderata for Clarion (see text for details).

Sequentiality
Routineness
Trial-and-error adaptation
Implicit versus explicit processes
Synergistic interaction
Bottom-up and top-down learning
Procedural versus declarative processes
Motivational and metacognitive control

For justifying these desiderata (see tables 2.1 and 2.2), more supporting arguments and evidence are needed. But before that, an example that illustrates how these desiderata might be tied together is in order. The example, in a way, also justifies the desiderata above.

2.2. An Illustration of the Desiderata

According to the framework of Clarion, when an individual is born into the world, that is, when an agent is instantiated into the system, little information, skill, or knowledge is readily available. For instance, the individual comes with no explicit knowledge, either about the self or about the world. But the individual does come with evolutionarily hard-wired instincts (e.g., reflexes). Moreover, the individual has needs, such as hunger and thirst, which constitute innate motives driving actions and reactions. The individual certainly has no explicit knowledge of how to meet these needs but does have hard-wired instinctual responses, including primitive behavioral routines, which may be applied in attempts to satisfy the needs.

The individual is endowed with sensory inputs regarding environmental states and internal states. Whenever there is a growing physiological deficit, an internal change may lead to heightened activation of a motive (need). It may therefore lead to a goal to address the need (i.e., to reduce

the deficit), which may then lead to corresponding actions (based on innate behaviors initially). In the process, even the perception of the individual might be modulated somewhat so that, for example, it focuses more on the perceptual features that are relevant to the pressing needs.

Similar processes happen when there is a growing "deficit" in terms of a socially oriented need, such as the need for interaction with others (the need for affiliation and belongingness). In such a case, the individual may similarly generate a corresponding goal, which in turn leads to corresponding actions (initially based on whatever primitive behavioral repertoire that is available, for example, by crying).

Gradually, with trial and error, the individual learns more and more how to meet various needs, in part based on successes or failures in attending to these needs. The individual learns what actions to perform in what situations, in order to fulfill an outstanding need. When an outstanding need is fulfilled to some extent, pleasure is felt—a positive reinforcement. Based on such reinforcement, the individual learns to associate needs with concrete goals and in turn also learns to associate goals with actions that best accomplish the goals. Through the trial-and-error process, the individual increases competence (developing more effective and more complex routines or skills), which helps to deal with similar or more difficult situations in the future.

In this process, the individual may experience a variety of affect states, which facilitate learning and performance of actions: *elation* when goals are accomplished (needs are met and positive reinforcement is received), *frustration* when unable to accomplish goals despite efforts, and *anxiety* when negative consequences (thus negative reinforcement) are expected, and so on.

Moreover, gradually, the individual starts to develop explicit (symbolic) knowledge regarding actions (i.e., explicit procedural knowledge), beyond implicit associations acquired through trial and error (that is, implicit procedural knowledge discussed above). Explicit procedural knowledge may be extracted on the basis of already acquired implicit procedural knowledge (through "bottom-up learning"). Explicit knowledge in turn enables the individual to reflect on the knowledge and the situations, to plan ahead, to communicate the knowledge to others, and so on. Thus, implicit and explicit procedural knowledge together may lead to more effective coping with the world (i.e., a synergy effect).

Furthermore, even general knowledge that is not directly tied to actions (namely, declarative knowledge) may be generated over time. It

may be generated on the basis of acquired procedural knowledge (which may involve bottom-up learning) or from information provided by others (which may involve top-down learning). Such declarative knowledge adds more capabilities to the individual.

So, drawing lessons from this scenario, according to the Clarion framework, an individual starts small: there are only minimum initial structures. Some of these initial structures have to do with behavioral predispositions (e.g., evolutionarily pre-wired instincts and reflexes); some others have to do with learning capabilities; yet some others have to do with motivation. Together they constitute the genetic and biological pre-endowment.

Most of the mental contents within an individual have to be "constructed" (learned) during the course of individual ontogenesis. Development occurs through interacting with the world (physical and sociocultural). It leads to the formation of various implicit, reactive behavioral routines (skills), which in turn lead to explicit (symbolic) representation. The generation of explicit representation is, to a significant extent, determined by implicit mental contents within an individual. Of course, there is also another source: sociocultural influence, including through symbols employed in a culture.

Overall, the mind of an individual is mostly activity-based, action-oriented, and embedded in the world. An individual often interacts with the world in a rather direct and immediate way (Heidegger, 1927; Dreyfus, 1992), although more explicit, more contemplative, less direct ways may develop within the individual.

In Chapter 3, another example will pick up from here, continuing the learning processes discussed thus far, adding more details. But now, I will explore further the desiderata that were identified and illustrated above by examining the relevant empirical literature.

2.3. Justifying the Desiderata

Here I will not attempt to address all of the desiderata enumerated earlier, but instead will focus on some more controversial ones. Some points such as sequentiality, routineness, and trial-and-error adaptation have been thoroughly discussed in Sun (2002), and they seem almost self-evident by now. These will not be discussed again here.

2.3.1. Implicit-Explicit Distinction and Synergistic Interaction

To justify the Clarion worldview, I will start by examining in detail the distinction between implicit and explicit processes, which is the foundation of the Clarion framework. The theoretical distinction between implicit and explicit processes, as well as its ecological-functional significance, has been argued in the past in many psychological theories. See, for example, Reber (1989), Seger (1994), and Sun (1994, 2002).

First, the distinction of implicit and explicit processes has been empirically demonstrated in the implicit memory literature (e.g., Roediger, 1990; Schacter, 1987). The early work on amnesic patients showed that these patients might have intact implicit memory while their explicit memory was severely impaired. Warrington and Weiskrantz (1970), for example, demonstrated that when using "implicit measures," amnesic patients' memory was as good as normal subjects; but when using "explicit measures," their memory was far worse than normal subjects. The explicit measure used included free recall and recognition, while the implicit measures used included word-fragment naming and word completion. It has been argued that the implicit measures reflected unconscious (implicit) processes because amnesic patients were usually unaware that they knew the materials (Warrington & Weiskrantz, 1970). Such results demonstrating dissociations between implicit and explicit measures have been replicated in a variety of circumstances.

Second, Jacoby (e.g., Jacoby, 1983) demonstrated that implicit and explicit measures might be dissociated among normal subjects as well. Three study conditions were used: generation of a word from a context, reading aloud a word in a meaningful context, and reading aloud a word out of context. The explicit measure used was recognition (from a list of words), while the implicit measure was perceptual identification (from fast presentations of words). The results showed that, using the explicit measure, generated words were remembered the best and words read out of context were remembered the least. However, using the implicit measure, the exact opposite pattern was found. Other dissociations were also found from other manipulations (see, e.g., Roediger, 1990; Schacter, 1987). Toth, Reingold, and Jacoby (1994) devised an inclusion-exclusion procedure for assessing implicit and explicit contributions, which also provided strong indications of dissociation.

Third, the distinction of implicit and explicit processes has also been empirically demonstrated in the implicit learning literature (Reber, 1989;

Seger, 1994; Cleeremans et al., 1998). For example, serial reaction time tasks involve learning of a repeating sequence, and it was found that there was a significant reduction in response time to repeating sequences (compared to random sequences). However, subjects were often unaware that a repeating sequence was involved (e.g., Lewicki, Czyzewska, & Hoffman, 1987). Similarly, dynamic process control tasks involve learning of a relation between the input and the output variables of a controllable system, through interacting with the system. Although subjects often did not recognize the underlying relations explicitly, they nevertheless reached a certain level of performance in these tasks (e.g., Berry & Broadbent, 1988). In artificial grammar learning tasks, subjects were presented with strings of letters that were generated in accordance with a finite state grammar. After memorization, subjects recognized new strings that conformed to the artificial grammar, although subjects might not be explicitly aware of the underlying grammar (except for some fragmentary knowledge; Reber, 1989). In all, these tasks shared the characteristic of implicit learning processes being involved to a significant extent.

Generally speaking, explicit processing may be described mechanistically as being based on rules in some way, while implicit processing is more associative (Sun, 2002). Explicit processing may involve the manipulation of symbols, while implicit processing involves more instantiated knowledge that is more holistically associated (Sun, 1994, 2002; Reber, 1989). While explicit processes require attention, implicit processes often do not (Reber, 1989). Explicit processes may compete more for resources than implicit processes. Empirical evidence in support of these differences can be found in, for example, Reber (1989), Seger (1994), and Sun (2002).

Similar distinctions have been proposed by other researchers, based on similar or different empirical or theoretical considerations (Grossberg, 1982; Milner & Goodale, 1995; McClelland, McNaughton, & O'Reilly, 1995; Erickson & Kruschke, 1998). There have also been many other tasks that may be used for demonstrating implicit processes, such as various concept learning, reasoning, automatization, and instrumental conditioning tasks (for a review, see Sun, 2002). In particular, it is worth noting that in social psychology, there have been a number of dual-process models that are roughly based on the coexistence of implicit and explicit processes (see, e.g., Chaiken & Trope, 1999). Evans and Frankish (2009) included a collection of theories and models based on this kind of distinction. Taken together, the distinction between explicit and implicit

processes may be supported in many ways, although details of some of these proposals might be different (or even contradictory to each other in some way). Although some researchers disputed the existence of implicit processes based on the imperfection and incompleteness of tests for explicit knowledge (e.g., Shanks & St. John, 1994), there is an overwhelming amount of evidence in support of the distinction (see Sun, 2002 for further arguments).

Now the question is whether these different types of processes reside in separate memory stores (memory modules or systems) or not. There have been debates in this regard, and differing views exist (Roediger, 1990). Squire (1987) proposed that memory be divided into declarative and procedural memory, with the former further divided into episodic and semantic memory and the latter into skills, priming, classical conditioning, and so on. According to Squire (1987), declarative memory was explicit while procedural memory was implicit. Tulving and Schacter (1990) incorporated some features of the one-system view on memory while preserving the separation of explicit and implicit memory. They proposed that there should be multiple priming systems in the implicit memory so that dissociations among different implicit measures could be accounted for. This proposal addressed some objections raised by the proponents of the one-system view. Sun, Slusarz, and Terry (2005) provided a theoretical interpretation of a variety of learning data (related to process control, serial reaction time, and other tasks, as mentioned earlier), based on the multiple memory stores view.

Work in neuroscience shows some evidence for the existence of distinct brain circuits for implicit and explicit processes (i.e., separate memory stores). For instance, the work of Schacter (1990), Buckner et al. (1995), Posner, DiGirolamo, and Fernandez-Duque (1997), Goel, Bruchel, Frith, and Dolan (2000), Lieberman (2009), and so on provided some such indications. There have also been arguments that implicit memory represents a phylogenetically older system. This system may be more primitive but yet powerful on behavior.

However, as pointed out by Hintzman (1990), "once the model has been spelled out, it makes little difference whether its components are called systems, modules, processes, or something else; the explanatory burden is carried by the nature of the proposed mechanisms and their interactions, not by what they are called" (p.121). The debates regarding whether dissociations and distinctions of various kinds mentioned above

point to different processes or difference systems should be seen in this light. Sun (2012) provided further arguments in this regard.

In relation to the ecological-functional perspective articulated before, it should be noted that there have been some indications that explicit processes are evolutionarily newer than implicit processes (Reber, 1989). But the juxtaposition of the two is functional. It is functional and thus evolutionarily advantageous, especially because the interaction between the two types of processes may lead to synergy in the form of better, more accurate, and/or faster performance in a variety of circumstances (as I have extensively argued in prior work). Further discussions of synergy from the interaction, both in an empirical sense and in a computational sense, can be found in Section 2.5, as well as in Sun (2002), Sun, Slusarz, and Terry (2005), Helie and Sun (2010), and so on. Synergy, although not universal (i.e., not present in all circumstances), has been amply demonstrated in a wide variety of situations. Therefore, the division of implicit and explicit processes may conceivably be favored by natural selection. In addition, the separation of the two types of information, knowledge, mechanisms, and processes enables the adoption of each type as appropriate for different types of situations. For example, highly complex situations may be better handled by implicit processes, while explicit processes operating in a more precise way may be better for more clear-cut situations (Sun, 2002; Sun & Mathews, 2005; Lane, Mathews, Sallas, Prattini, & Sun, 2008). Furthermore, the division also enables parallel applications of the two types for different purposes simultaneously. So, putting everything together, the separation and the interaction of these two types of processes are psychologically advantageous.

2.3.2. Separation of the Implicit-Explicit and the Procedural-Declarative Distinction

I now turn to the distinction between procedural and declarative processes (i.e., action-centered and non-action-centered processes in the action-centered and the non-action-centered subsystem, respectively, as will be explained later) and its orthogonality with the implicit-explicit distinction (which might be a more controversial point).

The distinction between procedural and declarative processes has been advocated by Anderson (1983), Squire (1987), and many others (although some details vary across different proposals). Procedural