Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 51 (2018) 14–23

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Book review

Musings on the utility and challenges of cognitive unification: Review of *Anatomy of the Mind*, Ron Sun. Rensselaer Polytechnic Institute (2016). 480 pp.

Action editor: George Kampis

Ronaldo Vigo

Consortium for the Advancement of Cognitive Science, Psychology Department, Ohio University, Athens, OH 45701, United States

Received 3 April 2018; accepted 20 April 2018
Available online 16 May 2018

1. Preamble

Some of the most intrepid and ambitious attempts at rigorously understanding the human mind can be traced back to the late 19th century when psychophysicists and perception scientists attempted to formulate principled and integrated formal accounts of psychological phenomena (Ebbinghaus, 1885; Fechner, 1860; von Helmholtz, 1885; Wundt, 1874). A biproduct of these early attempts was the realization that the mind and brain are highly complex systems and that much still had to be learned before making the kind of progress initially envisioned. Subsequently, this awareness fueled the mid-20th century development of theories that aimed to integrate the wide variety of capacities and mental experiences identified by these and other early pioneers as characteristic of the human species: namely, the ability to perceive, conceive (i.e., generalize), store and recall information, reason, solve problems, make decisions, and experience emotions and drives.

As is typically the case in the evolution of an academic field, these developments were not without context. Just over a century before, several brilliant minds wrote seminal philosophical works attempting to tackle the problem of unification, albeit from an epistemological and, debatably, more ambitious perspective. For example, Hume's *Treatise on Human Nature* (1738), Le Matrie's *Man a Machine* (1747), Kant's *Critique of Pure Reason* (1781), and much later, James' *The Principles of Psychology* (1890), and Carnap's *The Logical Structure of the World* (1921) may be regarded as early systematic and unificatory inquiries into the nature of the human mind from the standpoint of the limits of human knowledge. Hume, for instance, presents one of the most clear and sensible informal accounts of cognition ever proposed that is still relevant today. In contrast, Carnap attempted to reduce what is knowable to phenomenalistic terms. His formal system was grounded on set-theoretic relations and a similarity relation on qualia. Using a more humanistic conceptual framework, James covered a much larger spectrum of what are now widely regarded as "psychological phenomena".

Although one may still learn much from these pioneering works, unlike their late twentieth century counterparts, they lacked the benefit of nearly one hundred years of modern experimental psychology and, more importantly, the tools and methods borne out of a handful of developments from the 1920s to the middle of the 20th century that characterized the beginning of the computational age. It was during this period that, alongside the advent of more sophisticated and complex computing devices – made possible by the development of the vacuum tube, the transistor, integrated circuits, and finally, very large scale integrated circuits – came the gradual development of more user-friendly high-level computer languages. Likewise, with computational resources being increasingly more accessible, some of the most influential theories of the time thrived

E-mail address: vigo@ohio.edu
URL: http://www.scopelab.net

https://doi.org/10.1016/j.cogsys.2018.04.007

on the power of mathematical and computational modeling.

Among these theories, the early theories of computation (Gödel, 1931; Church, 1932; Turing, 1937) and information (Hartley, 1928; Shannon & Weaver, 1949), along with the theory of artificial neurons (Rosenblatt, 1958) were particularly influential in refining and improving methods for the systematic and integrative study of cognitive phenomena. For instance, Simon, Shaw and Newell (1959) were among the early pioneers to take advantage of the mentioned computational mindset in their attempt to integrate and generalize problem solving phenomena from a computational modeling perspective. Lessons learned from their failures were significant steps forward in the quest for unification. Others followed suit without much more success. Naturally, these accounts varied in their level of analysis, scope, and in the central constructs and principles (some implicit and some explicit) at their core. In my own work (*Mathematical Principles of Human Conceptual Behavior*; Vigo, 2015), for example, I have sought such cognitive integration via a primarily mathematical theory in contrast to the primarily computational approaches of the twentieth century and the beginning of the twenty-first century (more on this distinction later). Thus, before proceeding to review Ron Sun's book in the following section, the reader should be alerted that any perceived biases in my part come from this specific background.

2. Overview

To start with, by a computational approach to cognitive integration, I roughly mean an approach under which formal descriptions of processes, phenomena, and interactions between phenomena takes the form of a computer program (i.e., a collection of data structures and algorithms rendered in some computer language). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture* by Ron Sun (2016) aims at such computational integration via what is known in the literature as a "cognitive architecture" – a general ("generic", as he refers to it) computational and integrative framework for modeling cognitive capacities, processes, phenomena, and their interactions. Ron Sun is a professor of cognitive science at the Rensselaer Polytechnic Institute in New York and the principal builder of the Clarion cognitive architecture. From the outset, he informs the reader that the current state of Clarion is the product of over two decades of development.

Cognitive architectures, of course, are not new to cognitive science. They have been around for quite some time. Among the best-known cognitive architectures are ACT-R (Anderson, 1983), Soar (Newell, 1990; Rosenbloom, Laird, & Newell, 1993), and Psi (Bach, 2009). They differ in the type of core representations employed, the algorithms that operate on these, the granularity of their modular structure, the principles that the architecture must abide by, and the size and nature of their parameter space, among other factors. Furthermore, they are supremely important because they give a much-needed backbone to our relatively young science of cognition. As I have noted in previous work (Vigo, 2015, p. xi): "In colloquial terms, cognitive science in its present state mostly amounts to a hodge-podge of theories, models, and empirical results that lack genuine predictive and explanatory power, and a common solid basis for unification." Thus, in my view, any well-grounded formal attempt at unification, such as Clarion, deserves serious attention. This is especially true because Clarion seems to be more eclectic than the mentioned cognitive architectures. More specifically, Clarion seems to encompass ideas from more subfields (e.g., evolutionary cognition, embodied cognition, affective science, goal-oriented problem solving, motivation, instrumental learning, etc.) than its competitors while invoking the seminal works of Watson, Gibson, Rosenblatt, Simon, Tversky, Kahneman and other luminaries for inspiration.

Furthermore, cognitive architectures are useful because through their comprehensive computational explanations based on a set of core representations, they can provide the context necessary for understanding and organizing the complex landscape of cognitive phenomena. David Marr (1982), in his landmark book *Vision*, illustrates the importance of such integration by proposing the need to analyze information processing systems from multiple levels. Such analysis can reveal the potential interactions (i.e., mechanisms) between the smaller subsystems that determine the outputs of the main system. However, at this early stage of development, cognitive architectures also have their share of problems and challenges. Throughout this review, I shall discuss these problems, not as a form of criticism, but in the interest of providing constructive context for evaluating their contributions. Indeed, any work as ambitious as a cognitive architecture should be preceded with a warning that given the relative infancy of cognitive and psychological science, we are what alchemists were to physics in its early stages of development.

With an eclectic spirit in mind, *Anatomy of the Mind* was designed to be the most comprehensive and accessible conceptual account of Clarion to date: in short, a compendium of Sun's previous work on Clarion and that of his associates over several years. The book is well-written and I suspect that it will be relatively easy to understand by non-experts. On the other hand, experts will have much to gain and may enjoy the way Sun incorporates ideas from many familiar studies in the cognitive literature into the Clarion framework and (as he refers to it) the Clarion "world view" (i.e., intuitive narratives regarding the way that information from the environment and the mind itself is hypothetically processed). In fact, it is the very fact that Sun uses a plethora of empirical and theoretical classic studies to build his case on behalf of Clarion that makes this book also suitable for a graduate introductory course to cognitive architectures, or even an introductory course to cognitive science. The book is divided into nine neat, but long chapters. The arrangement of each chapter into nested sections

makes it a pleasure to navigate through them. Next, I summarize the content of each chapter; this will be followed by a more detailed discussion of some of the topics that I found particularly interesting and/or debatable in each chapter.

To start with, the preface and first chapter offer a motivational perspective on cognitive architectures in general and on the Clarion architecture. Particularly, Sun proposes that one of the main goals of Clarion is to rigorously account for a wide variety of psychological phenomena within a unified framework and that computational modeling is essential in achieving this goal. In addition, he alerts the reader that the book provides a conceptual level exposition only and that the technical details of Clarion will be the subject of a forthcoming volume. Unfortunately, as many of us know, the "devil" is in the details, and a proper assessment of a computational theory involves perusing its technical details. Thus, the expert reader's assessment of *Anatomy of the Mind* must necessarily be limited to the conceptual narratives and explanations (aided by brief technical appendices on simple algorithms), and not on whether the models in Clarion function as described.

Chapter 2 introduces the essential dual constructs underlying Clarion, such as the implicit-explicit knowledge distinction, bottom-up and top-down learning, motivational and metacognitive control, and the four subsystems of Clarion (more on this later). Chapter 3 explains the action-centered subsystem and non-action centered subsystems (NACS) and describes some of the basic representations and algorithms operating within them. Similarly, Chapter 4 explains the motivational and metacognitive systems with their basic representations and algorithms, while Chapter 5 gives some welcomed details regarding how to simulate procedural and declarative processes in Clarion. Chapter 6 shows how to simulate motivational and metacognitive processes building on similar procedures introduced in Chapter 5. In Chapter 7, Sun discusses cognitive social simulations using representations based on the dual constructs introduced in previous chapters. Finally, Chapters 8 and 9 provide some needed relief in that they address important issues that perceptive readers will probably entertain after reading the first seven chapters. Specifically, Chapter 8 anticipates and answers a list of questions that readers may have about Clarion. Unfortunately, some important questions (discussed in this review) were not covered in the chapter. Regardless, the questions that were discussed will likely help the reader gain a deeper appreciation for the strengths and limits of Clarion. Chapter 9 wraps up the book with a summary of Clarion and a general discussion of important issues and future directions. Next, I shall discuss in greater detail the content of each chapter under the lens of particular challenges facing cognitive architectures.

3. Musings per Chapter

First, note that because the Preface and Chapter 1 summarize the entire book, some of my comments about these two sections extend to the entire book. In the preface and Chapter 1, Sun gives a general road map of the book after having defined in the preface his target readership rather broadly as cognitive scientists, experimental psychologists, researchers and lay readers who are interested in computational psychology. A large portion of Chapter 1 is devoted to a discussion of the character and nature of cognitive architectures and why computational models and theories are desirable in science. Sun regards mathematical models as subsets of computational models, where computational models are algorithmic descriptions while mathematical models are abstract relationships between variables expressed by mathematical equations. However, the reverse interpretation of computational models being subsets of mathematical models may be construed as equally valid.

More specifically, Sun suggests that computational modeling offers the expressive power that mathematical modeling cannot match because mathematical models are subsets of computational models. Although in spirit I agree that computational modeling holds certain advantages over mathematical modeling, I give four reasons why the converse is also true. First, given their high level of abstraction, mathematical models are, by their very nature, better suited for developing high-level descriptions of information processing systems than computational models. Second, mathematical models may be unwrapped (or interpreted) in algorithmic terms, if need be, anyway. In other words, collections of mathematical models that form the backbone of a theory may be broken down and interpreted in terms of a theoretically meaningful step-by-step mechanistic qualitative description that is consistent with their outputs; in turn, computational models may be described in terms of a collection of mathematical models as formal relations (i.e., nested mathematical functions, relations, and equations). Take, for example, neural network models: these may be construed as simply sets of nested mathematical functions and relations. This fact appears more compelling when comparing theories of computation that are equivalent in character but Turing complete, such as recursive number theory (Gödel, 1931), the calculus of lambda conversion (Church, 1932, 1936), and combinatory logic (Curry, 1930; Schönfinkel, 1924). These theories of computation and computability have different mathematical constructs at their core but all characterize the nature of computation in a manner that is logically equivalent to a Turing machine (Turing, 1937), albeit in more symbolic, abstract, and algebraic terms.

Third, mathematical models are pithy descriptions of phenomena. One may say that what computational modeling is to prose, mathematical modeling is to poetry. Accordingly, the laws of science are typically expressed as mathematical models, not algorithms, because of the conceptual tractability and parsimony of this form of expression. Fourth, because of their economy of means, mathematical models make relatively fewer assumptions, and these are usually well delineated by the axioms from which they are derived. Accordingly, they tend to relate

in abstract terms a set of inputs to a set of outputs with a relatively small set of parameters. To summarize these four points, mathematical models, in a very fundamental sense, exhibit clear advantages over computational models. I shall give some specific examples of this assertion when discussing Chapter 5.

Furthermore, in this chapter, Sun extols the virtues of computational models and cognitive architectures. These include precision, testability, and integration. However, although he also alludes to some of the limits of computational models and architectures, I found that key fundamental challenges that could have been addressed were omitted. Among these is the idea that cognitive architectures are only as good as their building blocks: namely, their core representations and associated processing rules. If, for example, the core of a cognitive architecture consists of a certain type of neural network architecture, comprised of certain types of neurons, and implementing certain types of rules and learning algorithms, then the cognitive architecture will inherit the limitations intrinsic to these formal representations. Furthermore, the computational framework is often biased by the style of program or encoding of the algorithms. In view of these factors, if one of the building blocks of the architecture is flawed, the entire edifice may collapse: in short, a cognitive architecture is only as good as its weakest link. If one then resorts to one of several available implementational alternatives for the connectionist component, one in effect is altering the details of the original assumptions held by the architecture to a sufficient extent as to change its nature. In a nutshell, where does this tradeoff end?

Second, as mentioned under my characterization of computational frameworks above, unlike mathematical theories, computational frameworks often feature very large parameter spaces which require many assumptions and offer many degrees of freedom in making predictions and generating explanations. In this regard, Luce (1995, p. 10) states: "The [computational] approach is very flexible, which is both a virtue and a fault. It can be exceedingly difficult to be sure what about a particular processing model is correct. This is especially true when the processes are entirely hypothetical ..." This situation may be aggravated when the computational framework is very large in scale as is the case with cognitive architectures. Notwithstanding these issues, cognitive architectures do have one clear advantage over mathematical theories: they provide tools for facilitating connectivity between the various chosen formal representations of cognitive functions, albeit at the expense of conceptual tractability.

Sun points at how Clarion is grounded on empirical results. This is certainly of primordial importance and, in fact, the Clarion architecture is heavily informed by well-known empirical results. Some of these studies are more persuasive than others in their empirical rigor and are chosen by Sun to make his case in support of the constructs and basic structure of the architecture. However, I could not help noticing that these results were treated as facts, with the veracity of their replicability, robustness, competing results and theories, unquestioned. Clearly, in a book of this scope not everything can be covered nor addressed, but an admission of such adversity possibly arising would have been sufficient. Perhaps the admission was there, but I missed it. Note that the lack of such admission can act as a "double-edged sword". The problem lies in the inherent flaws of past research and the necessary credo that only results that meet the standard of an empirical law should be the basis of such grounding. Unfortunately, there are not many results yet in the psychological sciences that reach this standard. Indeed, it is hard enough to find results that are even replicable (Open Science Collaboration, 2015). To make matters worse, the publish or perish culture of today's academic landscape, where quick science plagued by sloppy methodology and small effect sizes seems to be the norm rather than the exception, has not helped to alleviate the situation. Thus, cognitive architectures should be grounded in laws and not in common sense.

Sun seems to recognize this conundrum and does his best to appeal to such statements in Chapter 5 (more on this later). The need to discover these empirical and mathematical laws (via mathematical models, not computational) is essential before we can arrive at architectures that, based upon them, are correct. Mathematical theories, whether fully axiomatized or not, provide the conceptual framework and rigor to be able to proceed forward to a lower level analyses of the type found in computational theories. I for one cannot envision the implementation of an electronic system, such as a radio or a cell phone, without having knowledge of the high-level mathematical laws that describe the relationship between current, voltage and resistance. The very parts that are used to build such a device are manufactured in a manner consistent with the dictum of these fundamental laws. Lastly, while I am on the subject, it is worth noting that attempts at cognitive integration are greatly informed by our most successful science, physics. One of the ultimate challenges of modern physics has been to specify the quantitative and qualitative relationships that exist between forces at different levels of physical reality. These attempts at unification began early: for example, Maxwell managed to unify magnetic and electrical phenomena in his theory of electrodynamics with the help of the empirical work of Faraday. On the other hand, Helmholtz, albeit unsuccessfully, attempted to show the connection between energy fields in his *Über die Erhaltung der Kraft* (*On the Conservation of Force*, 1847). Other examples of theories abound, but the point is that even our most mature science, when compared to cognitive science, has not achieved its goal of ultimate unification. This fact makes cognitive architectures perhaps unreasonably ambitious in character.

Chapter 1 also contains a list of Clarion attributes that give it an edge over other architectures such as ACT-R and Soar. These include: comprehensiveness in the inclusion of motivational, affective, and social components, along with empirical tenability. However, readers will have to wait

until the final Chapter 9 to see a detailed treatment of this topic. This makes sense because readers should become acquainted first with the goals and constructs underlying Clarion in order to appreciate its full scope of contributions relative to the other considered architectures. The chapter concludes with a set of meta assumptions underlying Clarion: specifically, (1) an ecological perspective where interactions with the environment play a primordial role, (2) modularity of mind in the sense prescribed by Fodor (1983), (3) the possible coexistence of multiple representations corresponding to different types of representational content (e.g., procedural, declarative, metacognitive and motivational), and (4) dynamic interaction between motivation, cognition, and metacognitive components. In sum, Chapter 1 contains a clear specification of Clarion's general characteristics and goals.

In Chapter 2 Sun introduces the basic framework of Clarion. The chapter begins with a list of desired properties ("desiderata") for a cognitive architecture. The list of properties includes trial and error adaptation, synergistic interaction, implicit and explicit processing, bottom-up and top-down learning, motivational and metacognitive control, and procedural and declarative processing. Sun provides what he takes to be evidence in support of each of the desired properties. Note the dual nature of the last four properties. I found this dual character to be an interesting and overarching theme in Clarion. In fact, these dichotomies manifest themselves throughout the four basic subsystems of Clarion: namely, the action centered subsystem (ACS; where actions are controlled via procedural knowledge whether mental or motor), the non-action-centered subsystem (NACS; where general (declarative) knowledge is maintained for retrieval of appropriate information in the service of action-decision-making by the ACS), the motivational subsystem (MS; which provides underlying motivations for perception, action, and cognition) and the metacognitive subsystem (MCS; which monitors and regulates operations of the other subsystems). For instance, there is a top-level and bottom-level representational structure for each of the subsystems. In other words, each subsystem consists of a high-level and low-level representation that, as the terms imply, indicate two different types of knowledge encoding.

In Clarion, the top-level representation encodes explicit knowledge while the bottom-level encodes implicit knowledge. Unfortunately, like that of so many other constructs in the psychological landscape, I find the definition of "implicit" and "explicit" somewhat lacking in character and rigor. Admittedly, these are terms that are often used in the literature, but in my view, are not very well-defined. Intuitively, Sun equates implicit knowledge as relatively phenomenologically inaccessible in nature and cites Reber (1989) and Seger (1994) where phenomenological accessibility refers to the direct and immediate availability of mental content. In contrast, explicit knowledge is more accessible and more manipulatable. Correspondingly, Sun hypothesizes that the relatively inaccessible nature of implicit knowledge may be formally captured by subsymbolic, distributed representations. He explains that this is because distributed representational units (such as those that partake in hidden layers of backpropagation artificial neural networks) can carry out computations but are subsymbolic and generally not individually meaningful. On the other hand, explicit knowledge at the top level of each subsystem may be captured in computational modeling by symbolic or localist representations where each unit is easily interpretable and has a clearer conceptual meaning. The rest of the chapter is devoted to explaining the working of each subsystem from the perspective of the implicit vs. explicit knowledge dichotomy.

Chapter 3 focuses on the ACS (where action/procedural knowledge resides) and NACS (where conceptual/semantic knowledge resides). It is in this chapter that we first see a summary of the computational algorithms underlying the representations and operations in Clarion. In my view, this is the stuff that makes an architecture interesting. The reader will also see for the first time the specific mathematical expressions describing the algorithms and basic principles motivating them. Non-experts need not worry, for Sun does a fine job in making the material easy to comprehend. The core formal representation in Clarion is referred to as a "chunk". Chunks are represented by a chunk node (at the top level) connected to a set of microfeature nodes (at the bottom level). Thus, each chunk node at the top level of a chunk helps to identify and label each set of "dimensional features as a whole". The chunk nodes at the top level of the ACS can be connected to form conditional action rules of the form "condition-chunk-node → action-chunk node". For example, "if there is an obstacle in front, then jump over it". The microfeatures for each chunk node together constitutes the condition for activating the rule.

Chunks are useful representations because they afford the flexibility of formally representing the implicit and explicit knowledge dichotomy described in Chapter 2. The chapter continues by specifying decisions rules for the ACS and a set of bottom-level (using mostly backpropagation neural nets) and top-level learning algorithms for the ACS and NACS. Most of the algorithms introduced are standard or well-known in the literature; however, some (specially for top-level processing), seem to have been developed exclusively for Clarion. There is also an unfortunately superficial discussion of how one may characterize new conceptual knowledge by the introduction of new attractors in a Hopfield-type network. However, no attempt to connect the long and rigorous modeling tradition of concept learning and categorization research to the Clarion framework is made in this chapter (more on this later in my discussion of Chapter 5 below). Finally, I should mention that this chapter (as well as a few other chapters) includes a set of short technical appendices aimed to offer greater details on some of the technical topics discussed in the chapter. Unfortunately, given their brief character, their content would be best appreciated by advanced readers.

Chapter 4 explains how Clarion incorporates two forms of control: the control of actions and the control of the decision-making process to act. To accomplish the latter, motivational and metacognitive mechanisms are incorporated in Clarion. Sun argues that for an individual to survive and function well in the world, its behavior must conform with certain necessary characteristics such as sustainability (attending to essential needs), purposefulness (choose actions to enhance sustainability), focus (sustained actions with respect to purpose), and adaptivity (ability to learn). Each of these characteristics have been discussed extensively in the literature by prominent researchers cited by Sun. To account for these, motivational and metacognitive processes are incorporated into Clarion. Note that, as with so many other instances throughout the book that support the highly eclectic nature of Clarion, Sun makes it clear (particularly in his acknowledgment of the work of others) that he is merely bringing together ideas, not inventing them. This spirit characterizes the book throughout. Accordingly, he borrows from Maslow's hierarchy of motivations/drives (Maslow, 1943) and from Flavell's definition of metacognition as "one's knowledge concerning one's own cognitive processes and products" (Flavell, 1976). The remainder of the chapter is devoted to the motivational (MS) and the metacognitive subsystem (MCS) of Clarion. Particularly, the essential motivational representations, mechanisms and processes are discussed and described within the general framework of top-level and bottom-level processing. The algorithmic technical details (e.g., learning algorithm in the MCS) are left again to a technical appendix that experienced readers will greatly appreciate.

I found Chapter 5 somewhat controversial. This is not because of its underlying goals; on the contrary, this is the first time that we get to see evidence for how effective Clarion performs in simulating procedural and declarative processes. Rather, more specific details about Clarion's modus operandi began to emerge that suggested potentially great limitations. In the chapter, results from Clarion simulations are compared to data from a handful of experiments. Unfortunately, when closely examining the reported data, I was left with questions about how well the performance of cognitive architectures in general may be evaluated. Because the presented simulations come from published articles, the reader with similar questions should examine the details of the simulations and the appropriateness and rigor of the test methods used by perusing said articles. As already alluded to, one of the issues that confronts cognitive architectures in general, and not just Clarion, is that, given their parametric complexity and the vastness of their parameter spaces, they are endowed with enormous power and flexibility in simulating results. Thus, issues of experimental validity and experimental error should be taken particularly seriously when assessing their performance. Robust validation methods such as large scale cross validation (or bootstrapping) are particularly pertinent in these situations.

Under cursory examination, the simulations seem to perform well; however, I wish Sun would have presented simulations that, based on a set of wrong assumptions (according to Clarion), failed in Clarion. Again, sometimes modeling frameworks, due to the overzealous use of supervised learning and the gratuitous use of free parameters, can give the impression of being too powerful more than being correct. This issue is brought up by Sun himself as one of several questions about Clarion in Chapter 8. Essentially, Sun believes that the number of free parameters in Clarion is not exceedingly large when compared to other computational models and frameworks. Although this may be the case, it misses the essential point about the degrees of freedom introduced by free parameters. Overall, parametrically sparse mathematical models (especially those of the non-parametric variety) seem to not be plagued by this problem. Regardless, if cognitive architectures are to ever achieve the kind of respect commanded by successful physical theories, the development of appropriate benchmarks for evaluating their effectiveness must be established. Especially because the principles on which they are based, their intrinsic representations (connectionist, symbolic, or hybrid) and corresponding algorithms, as well as the granularity of their levels of description and processing, can all vary significantly among them.

After the simulations and models are considered, two subsections of this chapter, one on concepts and one on psychological "laws", were particularly interesting to me because they touch upon areas of personal familiarity. Categorization and concept learning research have a long and rich history of formal modeling approaches and robust results. Sadly, many of the more significant advancements in these areas were left out of the discussion (see Vigo, 2015 for examples). Specifically, researchers have known for some time now that depending on the nature of the dimensions of stimuli, humans will learn structurally equivalent concepts with varying degrees of subjective difficulty (Nosofsky & Palmeri, 1996; Vigo, Doan, & Zhao, 2018). However, these results have proven difficult to predict unless one utilizes models with many free parameters (e.g. connectionist models such as ALCOVE; Kruschke, 1992). The problem is that such models, because of the many degrees of freedom afforded by their multiple free parameter structure, are too powerful. Fortunately, a recent mathematical and computational model referred to as the generalized invariance structure theory model (GISTM) or the invariance law (Vigo, 2013, 2015) can make accurate predictions about concept learnability across thousands of instantiations of many concept structures (i.e., concept types) without free parameters. Note that this level of performance exists regardless of the type of dimensions (e.g., integral vs. separable) involved (Vigo et al., 2018). In addition, models by Feldman (2000, 2006) and Kurtz (2007) have attempted to make predictions about concept learnability in ways that cut down on the number of potential free parameters. Because of the fundamental nature of the mentioned empirical and

theoretical research, and because of the traditional path taken by Sun via Clarion to address the fundamental issues of concept formation, I regard efforts in this area lacking. Indeed, Clarion structures seem to rely on the traditional (and inadequate), often parametrically enriched, similarity-based paradigms (see Vigo, 2013, 2015 for a discussion) of concept formation. Some of the most fundamental and robust experiments on concept formation should be accounted for in a manner that convinces us that their accounting is not due to overpowering efforts in part of gratuitous supervised learning and/or the inclusion of multiple free parameters in the architecture. Disappointedly, this is not done in *Anatomy of the Mind*.

Given this concern and the "weakest link" concern mentioned earlier in this review, it seems as if only cognitive architectures that are consistent (and/or based) on empirical and mathematical laws can ever be correct. Sun alludes to this under a section in this chapter entitled "Capturing Psychological 'Laws'" in attempting to validate Clarion. Note the use of quotes around the term "laws" which is meant to preempt the section with a bit of wise and reserved judgment. In previous work, I have discussed the problem of characterizing and discovering laws in cognitive science by contrasting its status to that of physics and chemistry as follows: "... unlike physics and chemistry, cognitive science does not have at its core a set of standards in the form of empirically validated measures and mathematical principles from which to systematically derive potential laws and well-defined constructs." (Vigo, 2015. P. xi). Most of the "laws" discussed by Sun, claimed to be derivable in Clarion, are laws only in a weak sense of the construct. Specifically, most are weak informal or qualitative statements such as: "people are risk averse", "people's decisions are influenced by the availability heuristic", and so on. These statements are not formal statements that make precise predictions about phenomena. At best, most make qualitative predictions as to whether observer A will judge a situation as more probable or less probable than another under some constraint.

Furthermore, even when they are of a quantitative nature, where the main concern is "how much and when", formal statements such as Tversky's similarity measure (Tversky, 2007) are, at best, models laden with free parameters that have been proven ineffective in accounting for most phenomena that characterize the capacity, or they are built into the architecture as an assumption in the first place. One of the special features, among many others, of laws (in the strong sense observed in the physical sciences) is a property mentioned by Luce (1995) when criticizing Newell's Soar architecture (1990). His criticism refers to the often-large amount of supervised learning on some dataset necessary for the estimation of the parameter values of a model in psychology, and the lack of generalizability of the estimated parameter values in terms of these being applicable to other datasets. In physical models (and especially in physical mathematical laws) parameter values generalize from one set to any other data set. Luce (1995, p. 13) writes:

> ``Little comparable [parametric] invariance has evolved in psychology. It is moderately rare to find a psychologist who, when confronted by a new set of data, invokes already known mechanisms with parameters estimated from different situations. Newell (1990) claimed to do so in his computer-based, unified theory of cognition called SOAR, but I am not persuaded by the claim. When each model is unique to a particular experimental situation, all of the model's free parameters must be estimated from the data being explained. Frequently the resulting numbers of parameters outrun the degrees of freedom in the data. This reflects a failure of the science to be cumulative, an unfortunate feature of psychology and social science that is widely criticized by natural scientists. I view it as one of the greatest weaknesses of modeling (and other theory) in our science."

Thus, with a sufficiently large number of free parameters many proposed architectures may account for the same laws. Furthermore, if one builds into the system the "law" that is being predicted (as perhaps in some of the neural net learning algorithms and threshold rules), and not merely the components, indeed the law will be predicted.

To avoid these problems, cognitive architectures should be based on laws in the first place: in other words, fundamental empirical and mathematical laws (in the strong sense) should precede any architecture to avoid the pitfalls of the "weakest link" effect mentioned earlier. As discussed earlier, consider building an electronic device without knowledge of the basic laws of electricity. The implementation of the logic of the information flow that does the work in the device would not be attainable because it is through knowledge of these physical laws that one can characterize the performance limits and the exact quantitative and qualitative operations (transformations) achievable by devices that are the fundamental components of the system (i.e., the resistors, transistors, diodes, inverters, capacitors, etc.). In turn, architectures that are built from the start on a few very flexible set of learning algorithms can be misguiding for the reasons already given.

Chapter 6 is analogous to five, where a handful of simulations are used to demonstrate the usefulness and effectiveness of Clarion in modeling metacognitive and motivational processes. There are sections on modeling personality, moral judgements (and their relationship to motivation) and emotions (as rooted in motivation). Each topic is interpreted in terms of the groundwork established by the specification of the MS and the MCS in Chapter 4.

Like in Chapters 5 and 6, Chapter 7 contains a handful of simulations involving cognitive social situations. The general strategy in these simulations is to construct agents based on the subsystems of Clarion. For example, in one simulation referred to as the "Tribal Society Survival Task", a tribal society was simulated in which the interaction between individual cognition and social (and environmental) factors were explored using Clarion-based agent

models. Sun points out that these types of simulations have been attempted but not with the "cognitive realism" afforded by Clarion. Different metrics were used in the simulation to determine the success of individuals in the society (e.g., life expectancy). The simulation showed that certain societal strategies, such as the presence of a central store (i.e., a mechanism to which all individuals in a group transferred part of their resources) as opposed to a strategy in which each individual uses only its own resources, promotes survival. Another Clarion-based simulation introduced in this chapter pitted motivational factors against survival. These simulations are followed by a section in the chapter devoted to simulations in organizational decision making. There is even a section devoted to simulations of academic publishing.

Chapter 8 commanded my interest because it addresses several questions that perceptive readers may have regarding Clarion. The questions are the types of questions that reviewers of journal articles, conference attendees, students, and colleagues may have asked Sun and collaborators during their twenty years developing Clarion. The questions vary significantly in character and the answers are of varying degrees of tenability. Some examples follow: What is Clarion about? Is it a psychological theory? Can it be disproved? What distinguishes humans from other primates according to the Clarion framework? When are implicit processes used and when are explicit processes used? Is Clarion just a collection of old computational techniques? A set of "computational questions" are also addressed. Why was a purely connectionist or purely symbolic approach not attempted? Are there too many free parameters in Clarion? Why were recurrent neural networks not used in the ACS? Some of the issues that I brought up in previous chapters, in my opinion, were not adequately addressed by the answers to these questions. Nonetheless, many of the answers shed insight into Sun's thought processes and motivation in creating Clarion. The chapter is a clever and important component of the book in its ability to shed additional light on potentially difficult ideas introduced in earlier chapters.

Finally, Chapter 9 is quite interesting because it contains the material that it had been previously lacking on the limits of Clarion and how it compares to other cognitive architectures. Here Sun compares Clarion, albeit in brief terms to other cognitive architectures and admits that the comparisons are not detailed. First, Clarion is compared to some existing approaches. Before doing so, he reminds the reader that Clarion goes beyond narrower conceptions of situated/embodied cognition in a few ways: in summary, although no general purpose centralized symbolic processor is posited, Clarion addresses the existence of symbolic processes. In addition, Clarion addresses the emergence of symbolic processes from subsymbolic ones and the grounding of symbolic representations in subsymbolic processes and in ongoing interactions with the world. Clarion also addresses the following points: the innate distinction between implicit and explicit processes, the dual representation approach toward capturing this distinction, the importance of symbolic processes in the resulting system and the importance of bottom-up learning in the resulting system. However, in my view, these factors are too general to capture with precision the true nature of cognitive phenomena. That goal will only be achieved with the correct specific representations and structures and not with general guidelines as to their basic nature. Sun then highlights how Clarion is different from the well-known cognitive architecture ACT-R (Anderson, 1993). A central difference between the two is the lack of a distinction between implicit and explicit processes. However, ACT-R does capture the distinction between procedural and declarative knowledge analogous to the ACS and the NACS in Clarion. Other differences are highlighted, such as the lack of a psychologically realistic built-in modeling and explanation of motivation beyond simple goals. Soar (Rosenbloom et al., 1993), a cognitive architecture based on state spaces and operators for searching state spaces is also compared to Clarion.

Like ACT-R, Soar makes no distinction between implicit and explicit processing; furthermore, it does not distinguish between procedural and declarative processes. It also requires a large amount of initial (apriori) knowledge about states and operators unlike Clarion which is capable of autonomous bottom-up learning. He also compares Clarion to Psi (Bach, 2009). Like Clarion, Psi addresses autonomous learning and regulation of behaviors. It also addresses motivation via the construct of "demands" (external, cognitive, and social), but it does not include the implicit-explicit processes distinction nor does it address bottom-up learning and top-down learning. Also, Sun asserts that Psi does not have as much empirical grounding as Clarion. Unfortunately, in making these comparisons, Sun does not comment on the extent to which these various alternative architectures can account for the classic, robust experimental findings in the cognitive literature. Thus, the reader may be left with a sense of unresolved conflict regarding which architecture achieves the most with respect to the results that matter the most.

The chapter concludes with a few pages on future directions. Sun points out that much more work needs to be conducted in the area of cognitive social simulations. Here he calls for the modeling of individuals, not in isolation, but in terms of their relation to society: i.e., in terms of some social context. With respect to general directions for cognitive architectures, he calls for more comprehensive, more psychologically realistic, and more algorithmically sophisticated ones – a call that, undeniably, cognitive scientists will find agreeable.

4. Concluding remarks

In this review, to give constructive context to *Anatomy of the Mind*, I have touched upon seven challenges that face all cognitive architectures, not just Clarion. These were: modeling limits, degrees of interpretative freedom and

overfitting based on the overuse and misuse of free-parameters and supervised learning, the need of "real" laws as a foundation, empirical robustness, theoretical robustness, tractability, and testability. These are not meant to be criticisms but points of reflection in developing future architectures and refining existing ones.

These points aside, I do not know whether the Clarion architecture will be able to account for future and past robust neurophysiological and psychological findings, nor whether it accounts for the data of experiments featured in the book as proclaimed. What I do know is that, regardless, the utility of Clarion is apparent. Cognitive architectures act as organizational frameworks that help us organize the diverse results of a complex field. Thus, to a significant extent, they prove to be useful if their users are perceptive enough to understand the dangers of construing the mental world in specific ways without giving enough attention to alternative, perhaps more valid accounts. Also, the same users should recognize that the amount of empirical and theoretical work necessary to prove the validity of any grand architecture can be enormous. Yet again, these architectures, including Clarion, represent ambitious early attempts at the scientific integration of cognitive phenomena and, at the very least, can be useful tools in guiding us as to which paths we should not pursue in our search for scientific truth.

It is under this mindset that I have formulated my final assessment of *Anatomy of the Mind*. I know of no other book on human cognition that attempts to integrate and account for such a wide range of psychological phenomena from so many different research areas via a cognitive architecture. Remarkably, Clarion attempts this integration via a basic dual representation and a handful of accompanying dual constructs. Hence, for this valiant and indefatigable effort, I regard the book as an important contribution, not only to cognitive architectures, but to cognitive and psychological science in general. On a more practical side, because the book is so multifaceted, it can easily assume a pedagogic role beyond its role as a scholarly book. In fact, it may serve as a good generic introduction to cognitive architectures or as an integrative introduction to cognitive science. I would not hesitate to use it as a textbook in a graduate course on either subject. However, because of its clear focus, it would best serve its purpose as intended: a comprehensive and gentle introduction to Clarion. As such, I highly recommend it.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cogsys.2018.04.007.

References

Anderson, J. R. (1983). *The architecture of cognition*. London, UK: Psychology Press; Taylor & Francis.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*, 35–44.

Bach, J. (2009). *Principles of synthetic intelligence PSI: An architecture of motivated cognition*. Oxford, UK: Oxford University Press.

Carnap, R. (1921). *The logical structure of the world*. Berkeley, CA: University of California Press.

Church, A. (1932). A set of postulates for the foundation of logic. *Annals of Mathematics. Series 2, 33*(2), 346–366.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics, 58*, 345–363.

Curry, H. B. (1930). Grundlagen der Kombinatorischen Logik [Foundations of combinatorial logic]. *American Journal of Mathematics. The Johns Hopkins University Press, 52*(3), 509–536 (in German).

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York: Dover.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel, 2, p. 559 (Reprinted, Bristol: Thoemmes Press, 1999).

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*, 339–368.

Flavell, J. (1976). Metacognitive aspects of problem solving. In L. Resnick (Ed.), *The nature of intelligence* (pp. 231–236). Hillsdale, NJ: Erlbaum.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I", Monatshefte für Mathematik und Physik, v. 38 n. 1, pp. 173–198, in Solomon Feferman, ed., 1986. Kurt Gödel Collected works, Vol. I. Oxford University Press, pp. 144–195.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal, 7*(3), 535–563.

Hume, D. (1738). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. Oxford, UK: Clarendon Press.

James, W. (1890). *The principles of psychology*. Mineola, NY: Dover Publications.

Kant, I. (1781). *Critique of pure reason*. Cambridge, UK: Cambridge University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review, 14*(4), 560–576.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46*, 1–26.

Marr, D. C. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*, 370–396.

de La Mettrie, J. O. (1747). *Man a machine*. Cambridge, UK: Cambridge University Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*(3), 151–166.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3*(2), 222–226.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. http://doi.org/10.1126/science.aac4716pmid:26315443.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology, 118*(3), 219–235.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Rosenbloom, P. S., Laird, J., & Newell, A. (1993). *The SOAR papers: Research on integrated intelligence*. Cambridge, MA: MIT Press.

Schönfinkel, M. (1924). Über die Bausteine der mathematischen Logik. Mathematische Annalen. 92: 305–316. Translated by Stefan Bauer-Mengelberg as "On the building blocks of mathematical logic" in Jean van Heijenoort, 1967. A Source Book in Mathematical Logic, 1879–1931. Harvard Univ. Press: 355–66.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin, 115*, 163–196.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sun, R. (2016). *Anatomy of the mind: Exploring psychological mechanisms and processes with the Clarion cognitive architecture*. Oxford, UK: Oxford University Press.

Turing, A. M. (1937). Computability and λ-definability. *The Journal of Symbolic Logic, 2*(4), 153–163.

Tversky, A. (2007). Features of similarity. *Psychological Review, 84*, 327–352.

Vigo, R. (2013). The GIST of concepts. *Cognition, 129*(1), 138–162.

Vigo, R. (2015). *Mathematical principles of human conceptual behaviour*. New York, NY: Routledge.

Vigo, R., Doan, C., & Zhao, L. (2018). *A serioinformative and parainformative investigation into the learnability of categories defined over integral vs. separable dimensions*. Unpublished manuscript.

von Helmholtz, H. (1847). *Über die Erhaltung der Kraft, Scientific memoirs, selected from the transactions of foreign academies of science, and from foreign journals: Natural philosophy (1853)* (John Tyndall, Trans.), p. 114. Google Books.

von Helmholtz, H. (1885). *On the sensations of tone as a physiological basis for the theory of music* (2nd English Ed.) (A. J. Ellis, Trans.). London: Longmans, Green, and Co.

Wundt, W. M. (1874). *Grundzuge tier physiologischen Psychologie*. Leipzig: Engelmann.