Cognitive Development 27 (2012) 326–348

Contents lists available at SciVerse ScienceDirect

Cognitive Development

The past, present, and future of computational models of cognitive development

Matthew Schlesinger[a,b,c,*], Bob McMurray[d,e,f]

[a] Department of Psychology, Southern Illinois University Carbondale, United States
[b] Department of Electrical and Computer Engineering, Southern Illinois University Carbondale, United States
[c] Center for Integrated Research in the Cognitive and Neural Sciences, Southern Illinois University Carbondale, United States
[d] Department of Psychology, University of Iowa, United States
[e] Department of Communication Sciences and Disorders, University of Iowa, United States
[f] Delta Center, University of Iowa, United States

ARTICLE INFO

Keywords:
Computational models
Cognitive development
History

ABSTRACT

Does modeling matter? We address this question by providing a broad survey of the computational models of cognitive development that have been proposed and studied over the last three decades. We begin by noting the advantages and limitations of computational models. We then describe four key dimensions across which models of development can be organized and classified. With this taxonomy in hand, we focus on how the modeling enterprise has evolved over time. In particular, we separate the timeline into three overlapping historical waves and highlight how each wave of models has not only been shaped by developmental theory and behavioral research, but in return also provided valuable insights and innovations to the study of cognitive development.

The year 1986 saw the Challenger shuttle disaster, the Chernobyl nuclear accident, the Iran-Contra affair, and the publication of Rumelhart and McClelland's 2-volume set, *Parallel distributed processing: Explorations in the microstructure of cognition*. While the last event would not make most people's top-10 list, the *PDP* volumes have exerted an enormous influence on developmental science and their impact continues nearly three decades later (Elman, 2005; Elman et al., 1996; Karmiloff-Smith, 1992; Mareschal & Thomas, 2007; McClelland, 1995; Munakata & McClelland, 2003; Plunkett & Sinha, 1992; Quinlan, 2003; Schlesinger & Parisi, 2004; Shultz, 2003; Spencer, Thomas, & McClelland, 2009).

* Corresponding author at: Department of Psychology, Southern Illinois University Carbondale, Carbondale, IL 62901, United States.
E-mail address: matthews@siu.edu (M. Schlesinger).

0885-2014/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogdev.2012.07.002

Although computational models of learning and development existed before 1986, the *PDP* "bible" not only made *connectionism* a household word (among psychologists), but also catalyzed the emergence of modern developmental science. It restored learning as a core topic in cognitive science by introducing more powerful learning rules that enabled the development of internal representations. This evolution cogently illustrated the concept of *emergence*: Connectionist networks could, in some ways, develop themselves (McClelland, 2010). Perhaps most importantly, the PDP volumes illustrated the power of computational modeling as a tool for inquiry into the mechanisms of development.

Historically, most work in developmental science has combined careful observation/description of children and their environments (in all their glorious complexity) with well-controlled experiments attempting to distill some of these factors. In this context, what can a handful of equations or a simple algorithm offer, beyond what these methods already provide? The 25th anniversary of *PDP* presents the developmental community with an opportune moment to address this question by raising a series of important issues about the computational modeling enterprise (and not just connectionism): Why do we construct models? How has this enterprise changed? What can they teach us? And what have they taught us?

Our goal is to address these questions by surveying the history of computational models of cognitive development to highlight three themes. First, the array of modeling approaches available to developmental scientists has grown to cover a broad landscape of timescales, domains, and theoretical perspectives. But the consequence is not simply better or more diverse computing devices; rather, this diversity reflects increasingly creative and subtle theoretical development. Second, computational models have not only provided tests for theoretical accounts, but also generated insights and predictions that transcend the models themselves. Third, and perhaps most importantly, the very nature of the computational enterprise has evolved, as researchers have migrated from computational models as a tool for description and implementation of theory to models as a tool for inquiry and experimentation about theory.

We begin by describing the modeling enterprise, offering principles for evaluating and interpreting models. These have evolved along with more specific modeling approaches, and it is useful to bear them in mind. We next present a taxonomy of modeling paradigms. Here, our objective is to illustrate the complementarity between particular approaches and specific developmental domains and timescales. We then shift to a chronological perspective, describing three major waves since the mid-1980s. We highlight not only how computational models have improved, but also how this refinement is the result of interaction between model-testing, empirical research and theoretical development. Finally, we return to our three themes, to reflect on the progress of the last 25 years and to consider the future of developmental models.

1. (Why) does modeling matter?

Oakes, Newcombe, and Plumert (2009) asked, "Are dynamic systems and connectionist approaches an alternative to Good Old-Fashioned Cognitive Development?" The outcome of their analysis is optimistic, but the question is reasonable, particularly as dynamic systems and connectionism are not just theories of development, but particular ways of using modeling to understand it. Although many who study development are open-minded about the use of models, they are not yet convinced that computation adds value to the research enterprise.

Computational modeling will never supplant conventional behavioral methods. Rather, it *complements* such methods by answering questions about mechanism that are not immediately accessible to experimentation. Braitenberg (1984) offers an example. He describes a series of mechanical "creatures" (see Fig. 1) and uses them to illustrate both the power of simple mechanisms working in concert over time and the unique contribution of computational models:

> It is pleasurable and easy to create little machines that do certain tricks. It is also quite easy to observe the full repertoire of behavior of these machines—even if it goes beyond what we had originally planned, as it often does. But it is much more difficult to start from the outside and try to guess internal structure just from the observation of behavior. It is actually impossible in theory to determine exactly what the hidden mechanism is without opening the box, since there

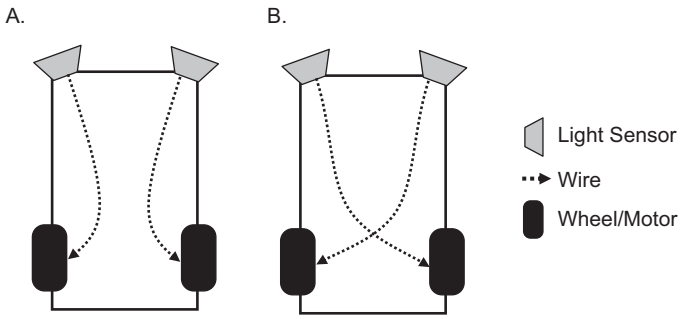

Fig. 1. Two examples of Braitenberg's Vehicles. When a light source is directly in front of them, both will approach it. If the light source is off to one side, however, (A) will turn away from it, while (B) will approach it. When allowed to move over time, such vehicles can show emergent behavior like light-loving, or light-fearing, with only minimal connections, receptors and actuators.

> are always many different mechanisms with identical behavior. . . .analysis is more difficult than invention in the sense in which, generally, induction takes more time to perform than deduction: in induction one has to search for the way, whereas in deduction one follows a straightforward path. (p. 20)

As an example of this process, imagine two simple creatures like those in Fig. 1. Assume the speed of each creature's wheels varies as a function of the distance from each sensor to a light source in the environment. In the case of the first creature (A) a light on the right causes faster spinning of the right wheel, which drives the creature *away from the light*. In the second case (B), where each sensor now drives the wheel on the *opposite* side, a light on the right causes faster spinning of the left wheel, which drives the creature *toward the light*. Thus, a small change in the connections between the sensors and motors leads to a qualitatively different behavior pattern. In a similar fashion, Braitenberg describes a progression of relatively minor modifications to his creatures, each one resulting in a new and often surprising behavior, like violence (attacking a light source), fear (retreating), restlessness and even love and attachment.

Thus, a fundamental motivation for computational modeling is that it offers relative freedom over the mechanisms built into the model (and the environment it experiences), coupled with the ability to observe the (sometimes nonobvious) consequences of such choices. This allows us to investigate the functioning of hypothesized mechanisms, working from the inside out, rather than the outside in. Of equal importance, modeling also offers the advantage of being able to start with a relatively simple model that becomes progressively elaborated through an iterative *design* ⇒ *test* ⇒ *refine* process, what Braitenberg calls "downhill invention."

To broaden this discussion beyond Braitenberg, there are several other motivations of modeling as a tool for investigating development. First, model-building forces researchers to explicitly define theoretical constructs (and the causal relations between them, i.e., developmental mechanisms). We do not always know the consequences of our theoretical commitments, particularly in the context of multiple interacting mechanisms. Thus, sometimes, the only way to understand their implications is to formalize them mathematically. However, models and behavioral experiments are mutually informative. Models implement theories, provide sufficiency proofs, generate predictions, and answer questions about theories. Behavioral studies evaluate these predictions and provide a metric for comparing models. Also, they can help us understand processes even when the theory is not well-specified enough to build a model.

Second, the most theoretically informative models can be thrown away. A successful model should offer an explanation in the language of developmental science, without relying on modeling jargon, specific mathematical algorithms, or cognitive architectures. For example, Munakata, McClelland, Johnson, and Siegler (1997) describe a recurrent neural network that is used to study the development of object permanence in infants. A key lesson from the model – independent of the specific architecture – is that relatively advanced skills such as reaching may require stronger internal representations than more basic skills such as visual tracking.

Similarly, models that fail to capture a phenomenon are also informative. Models that are implausible, non-parsimonious or perform poorly may reveal constraints on learning that are exploited over development. For example, the failure of a standard statistical learning model to learn speech categories has suggested the need to buttress statistical learning with online competition (McMurray, Aslin, & Toscano, 2009). Similarly, the failure of Elman's (1993) simple recurrent network (Wave 2 models) to learn simple grammars led to the idea that a limited memory capacity early in development may be beneficial.

Third, models can bridge levels of analysis. In particular, a core commitment of developmental systems theory (Gottlieb, 1997; Johnston & Edwards, 2002; Oyama, Griffiths, & Gray, 2001; Spencer, Blumberg, et al., 2009; Spencer, Thomas, et al., 2009) is that explanations of developmental change often span levels of analysis – genes-to-neurons, neurons-to-brains, brain-to-behavior and back again. Computational models can play a critical role in developing coherent explanations that span these levels.

Finally, models do not have to capture life-as-it-is. Unlike their human counterparts, computational models (and simulated organisms) are open to a much wider range of experiences. They can also be exposed to unnatural inputs, or "broken" or manipulated in ways that cannot be done with humans. This makes them an ideal tool for investigating phenomena such as critical periods, sensory deprivation, atypical development, and so on – variation that cannot always be studied empirically.

All this is not to say that modeling is without pitfalls. Beyond technical skills, modeling is a complex theoretical skill. Designing a useful model requires constraining model development with more abstract commitments (e.g., parallel processing, incremental learning), and simultaneously using the model to answer theoretical questions. It is tempting to "rig" the model, as modelers may sometimes explicitly select the parameters of a model to match a target behavior and fail to fully explore the parameter-space (although see Apfelbaum & McMurray, 2011; Pitt, Kim, Navarro, & Myung, 2006). Models can also be reified when the model and the real-world phenomenon are treated as isomorphic. And it can be tempting to construct more and more sophisticated models that can do more and more things, without really answering any new questions.

Ultimately, however, it is what we do with a model that is most crucial for advancing our science. A particularly compelling approach is to compare alternative models. For example, Shultz (2003) describes a series of developmental tasks (e.g., the balance scale, conservation of number) and presents systematic comparisons or *bakeoffs* between alternative modeling approaches on each of these tasks. One can also compare competing versions of the same model. For example, McMurray, Samelson, Lee, and Tomblin (2010) compare about a dozen versions of the TRACE model of speech perception (McClelland & Elman, 1986) to data on word recognition in language impairment to determine how best to characterize their deficits. Unfortunately, this is done only rarely, as model-builders typically focus on a preferred paradigm (perhaps due to the complexities of mastering two), and even when models are compared explicitly, it is often not clear what can be learned (other than which model is better). The single-model studies that are the norm can still be valuable for the aforementioned reason. But when they are highly abstract and complex, opaque to non-specialists, and difficult to disseminate outside the modeling community, their impact can be muted.

2. A taxonomy of developmental models in four dimensions

Most computational models start from a set of core theoretical commitments, things like parallel processing, neuron-like units and so forth, and use these principles to constrain model construction. In this way, one of the most common questions asked by a model is whether some phenomena can emerge from, or be accounted for, by such principles. Thus, understanding a specific model is often as much a matter of understanding the paradigm as it is the model. Here we highlight four dimensions that describe modeling paradigms: the type of representation, the way change is described, the mechanism of learning, and the time-scale in which change occurs. While the first two dimensions broadly distinguish most of the modeling paradigms in use (see Fig. 2), the last two are crucial for understanding specific models.

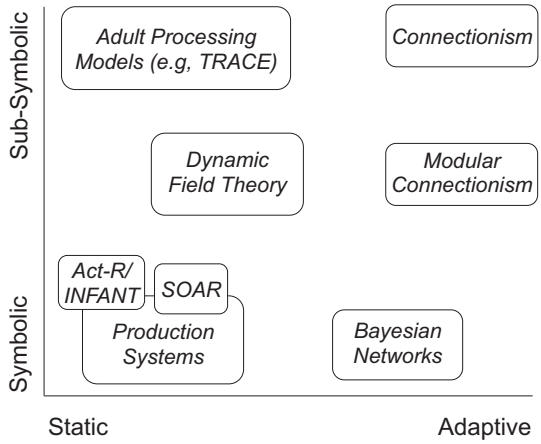

Fig. 2. A compression of our four-dimensional taxonomy of models into a two-dimensional space of model architectures. The figure illustrates where a number of common architectures are positioned with respect to the symbolic/sub-symbolic distinction and with respect to the static/adaptive distinction.

2.1. Representation: symbolic vs. sub-symbolic

A critical issue is how knowledge (sensory data, experience, memory, beliefs, plans) is represented. Historically, two approaches have addressed this issue, symbolic (Klahr & MacWhinney, 1998) and sub-symbolic (Rumelhart & McClelland, 1986b) models. Symbolic models treat knowledge as discrete units of information that are encoded in a language-like format. A well-known symbolic model is the *production-system model* (Klahr, Langley, & Neches, 1984). Productions are *if-then* or *condition-action rules* that guide decision-making, reasoning, and planning. For example, Klahr and Siegler's (1978) model of Piaget's balance-scale task uses rules like "if the weights on each side of the scale are unequal, then the side with the larger weight will tip." Other symbolic models of cognition include SOAR (Newell, 1990) and ACT-R (Anderson, 1993).

As symbolic models represent knowledge propositionally, they are well-suited to phenomena involving verbal reasoning (e.g., Piaget's conservation tasks). However, it is not always clear how developmental change occurs in such a framework. In response to this limitation, a relatively new class of symbolic models are Bayesian networks (Gopnik & Tenenbaum, 2007; Perfors, Tenenbaum, Griffiths, & Xu, 2011; Xu & Tenenbaum, 2007), which acquire explicit knowledge using the laws of rational probabilistic inference (Fig. 2). We return to these models in the next section, where we describe their application to language acquisition.

The alternative to symbolic representations is sub-symbolic or distributed representations. Artificial neural networks are an example. In these networks, representations emerge as a pattern of activity distributed over a set of computing elements (i.e., neuron-like units) and their connections. Crucially, these activity patterns and/or connections are not usually discrete and do not always correspond to rational symbol systems. In addition, instead of devoting a particular unit to a specific representation, each unit typically participates (to some degree) in representations for many inputs.

For example, we describe several connectionist models of Piaget's balance-scale task (McClelland & Jenkins, 1991; Shultz, Mareschal, & Schmidt, 1994). Each of these models receives as input a set of continuous values that encodes the positions and amounts of weights placed on each side of the scale. These values are propagated to the next layer in the network, creating an internal pattern of activity that serves as the representation for that particular input pattern. Crucially, these internal representations are highly overlapping with every node used (in part) to represent every input. This makes it difficult to point to any single part of the internal representation as representing a certain thing; rather representation is distributed across the network in a sort of graded way. While network models like these do not explicitly encode knowledge in a rule-like format, they are nevertheless

able to capture both conscious, rule-based behavior and also sensorimotor, implicit, or procedural knowledge (e.g., Shultz, 2003), and often in the same set of pathways (Seidenberg, 2005).

2.2. Adaptive vs. static models

Another important issue is whether the internal structure of the model (parameters or architectural properties) changes over time. Models that change are adaptive, as they improve or shift their response patterns over encounters with a set of inputs. In network models there are two ways to achieve this – changes in the strength of the connections between units, and changes in the architecture of the network, e.g., the appearance of new units (Quartz and Sejnowski, 1998; Shultz, Schmidt, Buckingham, & Mareschal, 1995). Adaptive approaches are not limited to connectionism, however. For example, Halford et al. (1995) propose a production-system model of transitive inference in which productions adjust their probability of "firing" as a function of performance on the task; and Bayesian approaches acquire statistics over the input to select the appropriate knowledge structure.

The alternative to adaptive models are static models. They maintain fixed internal parameter values and structure to capture behavior at a particular developmental time point (i.e., a specific age on a particular task). An example of this is INFANT (Simon, 1998), a modified version of ACT-R that simulates infants' perception of addition and subtraction events (Simon, Hespos, & Rochat, 1995; Wynn, 1992). While using static models to capture development seems counterintuitive, such models make the valuable point that just because one is studying children not all explanations need be developmental. Children, like adults, engage in online processes such as attention, selection, and competition. To the extent that such processes can account for a complex pattern of data without developmental change, they serve as a powerful reminder that we must integrate online and developmental timescales (McMurray, Horst, & Samuelson, in press; McMurray, Horst, Toscano, & Samuelson, 2009; McMurray et al., 2010).

Our use of the term *adaptive* is intended to highlight models that capture change on the developmental timescale. Nevertheless, as Fig. 2 illustrates, the adaptive-static dimension is a continuum rather than a dichotomy. Models that are capable of self-modification on the developmental timescale are situated toward the adaptive end of the continuum, while those that have fixed internal parameters and/or structures fall toward the static end. A number of models can be situated between these two extremes. Several models focus on behavioral change at the microgenetic or real-time level. Although these models do not develop or learn in the same sense that adaptive models do, their internal states do vary dynamically over time, in response not only to short-term changes in input, but also to memory "traces" of the internal states themselves (McClelland & Elman, 1986; Thelen, Schöner, Scheier, & Smith, 2001).

A related factor is how parameters within the model are modified over simulated developmental time. At the adaptive end of the continuum, the model-builder plays a minimal role in this process, by specifying the model architecture and learning algorithm and then allowing the model's output and performance to dictate the developmental trajectory (via incremental updates to the internal parameters or structure of the model). An intermediate case includes models whose internal parameters are modified over time, but in a relatively rigid way that does not depend systematically on the model's experience or performance level. In this case, such models are referred to as *hand-tuned*. A well-known example of a hand-tuned model is Elman's "starting small" connectionist network (1993), which illustrates the benefit of gradually increasing the model's short-term memory on its ability to perform a sentence-processing task. Another example is provided by Schlesinger, Amso, and Johnson (2007), who simulate the performance of young infants on a visual-search task. In this case, hand-tuning of model parameters is used as a proxy for the growth of corresponding neural components in early visual processing (Schutte, Spencer, & Schöner, 2003).

2.3. Supervised vs. unsupervised learning

Within the category of adaptive models are two strategies for simulating learning or development. Supervised-learning models learn through interaction with a "teacher" that provides explicit feedback. The backpropagation-of-error (or "backprop") rule (Rumelhart, Hinton, & Williams, 1986) is an

example of such a learning rule from connectionism. Backprop works by comparing the output pattern produced by the network with a desired or target pattern, and using the difference between output and target to modify connection weights within the network (moving the network's output closer to the target). Essentially, the model guesses the right response, gets a signal as to what it actually is, and adjusts itself to try to align the two. Such learning is very powerful as its ability to organize its own internal representations, and the detailed information it receives about the "error" at each component of the output allows it to learn complex input–output relationships. As a result, the backprop algorithm is used in the vast majority of connectionist models (Elman, 1990; Seidenberg & McClelland, 1989).

Models of unsupervised learning do not include explicit feedback. One variant of unsupervised learning is *Hebbian learning*. Hebb's rule states that connections between units in a network are gradually strengthened when they are active at the same time (Hebb, 1949). (Connections typically decay when one or both is inactive.) Hebbian learning can associate two inputs (or features of the same input), as seen in work by Sirois, Buckingham, and Shultz (2000), who used Hebbian learning to simulate habituation in young infants, and work by Apfelbaum and McMurray (2011), who showed how association principles can account for complex interactions between speech category learning and early word learning. Hebbian learning can also be used to associate inputs with some internal representation as in *self-organizing networks*. Self-organizing networks are typically used to cluster a set of input patterns into groups. For example, Mareschal, Plunkett, and Harris (1999) show how such models can learn to recognize and track moving objects; McMurray, Horst, et al. (2009) show how they can be applied to speech category learning (Guenther & Gjaja, 1996).

Between supervised and unsupervised learning is reinforcement learning (RL; Sutton & Barto, 1998). Like supervised learning, RL models generate an output and receive feedback. However, instead of receiving information about which specific elements are mismatched between actual and ideal outputs, in RL models, the feedback is a more general reinforcement "score," reflecting overall match. An appealing feature of RL is that it corresponds well to principles of learning in psychology (e.g., primary and secondary reinforcement, extinction, stimulus generalization) and to learning systems in the brain. Examples include work by Schlesinger (2003; Schlesinger & Parisi, 2001), who used RL to simulate the development of oculomotor control and the ability to track visible and occluded objects, and by Berthier, Rosenstein, and Barto (2005), who used RL to explore how variability influences real-time reaching dynamics over development.

2.4. Timescales of change

The final dimension is the timescale over which the model operates. As in behavioral studies, models span (at least) three timescales relevant to development (Table 1): real-time (seconds and minutes), microgenetic time (hours, days), and ontogenetic time (months and years). In addition, many models ignore time in a continuous sense all together, and instead focus on what we call "trial-time" – the mapping of an input to an output (no matter how long it takes) – as the minimum unit. Microgenetic and ontogenetic time are perhaps most often studied in development. In microgenesis, a common strategy is to construct a model that corresponds to a child at a given age and then to train the model in a way that is analogous to experience during an experimental session. For example, Sirois et al. (2000) use this strategy to simulate visual habituation/dishabituation, and Simon's (1998) INFANT model simulates the experience of an infant in Wynn's (1992) small-number arithmetic experiment.

There are two strategies for simulating ontogenesis. One is to employ a microgenetic model that learns over a relatively short period of time (or even a static model). To then simulate development, the model-builder systematically changes specific parameter values, usually different parameters than those involved in learning (i.e., "hand-tuning"). Jones, Ritter, and Wood (2000), for example, implemented an ACT-R model to solve a tower-building task. It begins with a set of general rules (*if-then* productions) and generates new, specific rules as progress is made through the task. Differences between age groups are then simulated by modifying parameters that constrain processing speed and capacity.

A second strategy is to directly simulate development in a model over longer timescales (months or years) using learning rules. Munakata's (1998) model of the A-not-B task simulates the

Table 1
Popular classes of models and the timescales they emphasize.

Model/approach	Real-time	Trial-time	Microgenetic	Ontogenetic	Model of change
Adult Processing Models (e.g., Dell, 1986; McClelland & Elman, 1986)	√	√			Static
Production systems (Klahr & Siegler, 1978)		√			Static
Act-R/INFANT (Simon, Hespos, & Rochat, 1995)		√	√		Static
Connectionist models of habituation (Sirois, Buckingham, & Shultz, 2000)		√	√		Static
Connectionist Models (Rumelhart & McClelland, 1986a)		√		√	Adaptive
Simple Recurrent Networks (Elman, 1990)		√		√	Adaptive
Bayesian Networks (Xu & Tenenbaum, 2007)		√		√	In between
Growth Models (e.g., McMurray, 2007; van Geert, 1998)				√	Static
Dynamic Field Theory (Thelen, Schöner, Scheier, & Smith, 2001)	√	√		√	Mostly static
Hebbian Normalized Recurrence (McMurray, Horst, et al., 2009; McMurray et al., in press)	√	√		√	Adaptive
Agent-Based Models (Schlesinger, Parisi, & Langer, 2000)	√	√		√	Adaptive

development of gaze and reaching in 6–12-month-olds. Similarly, Mareschal and Johnson (2002) simulate the development of perceptual completion (perceiving partially occluded objects as integrated wholes) in 2–4-month-olds. While these models simplify the environment relative to what a real infant experiences, they also generate a rich source of data that approximates a longitudinal study with frequent assessments. Model parameters can be "frozen" (held constant) at any time, while performance is tested, allowing the model to be tested periodically without influencing subsequent learning.

In between real-time and microgenetic time, a very common simplifying assumption of many connectionist and production-system models is to treat time as a series of discrete "trials". Here, the minimal unit is a single presentation of an input, like a word or visual scene (Elman, 1990), assuming a somewhat instantaneous mapping to the output. Doing so bypasses the need to worry about the complexity of real-time behavior or to make controversial assumptions about how many seconds elapse for each unit of time and offers a common unit of time that is related to the functional mapping the model is performing. However, it may also miss the contribution of real-time processes that children may engage as they map from input to output behavior.

Developmentalists rarely work with models that simulate real-time phenomena. However, this is common in adult cognition work. Examples include models like the TRACE model of speech perception (McClelland & Elman, 1986); normalized recurrence, which has been applied to problems like categorization, visual search and speech perception (Spivey, 2007); and interactive activation models of phenomena like speech production (Dell, 1986) and comparison (Goldstone & Medin, 1994). These models accurately capture timescales of process at the level of milliseconds. They can offer profound insight into developmental phenomena, but have rarely been considered in this light (but see McMurray et al., 2010).

One of the few classes of models that simultaneously examine real-time behavior and development is the Dynamic Field Theory (DFT) framework. DFT models are complex, nonlinear neural networks that capture real-time interactions between and among layers of networks. These have been used to explain several aspects of sensorimotor activity, including perceptual processing, motor planning, and

the integration of short-term and long-term memory. For example, Thelen et al. (2001) propose a DFT model of the A-not-B task. An important feature of their model is a "cooperativity" parameter, which modulates the relative influences of perceptual input (seeing an object hidden at location B) and motor history (prior reaches). By hand-tuning this parameter, Thelen et al. (2001) capture the development of perseverative reaching in young infants, as well as fluctuations in behavior that occur in real-time.

3. The recent history of developmental models: three (overlapping) waves

Given these core dimensions, we now shift to a chronological perspective, examining developmental models from the last 25–30 years. We describe three major waves of modeling work, in the sense of Siegler's (1996) "overlapping waves". As in Siegler's metaphor, successive waves reflect a trend from simple models toward more elaborate ones. In addition, the waves overlap; subsequent waves have not entirely replaced or eliminated earlier ones. Indeed, for some phenomena, a less-elaborate model may be more informative. We have identified these waves as marking a change in the zeitgeist of the developmental community, a new set of computational tools, a new vocabulary for expressing emerging ideas like distributed representation, embodied knowledge, and self-organization, or in many cases, all three.

The scope of our survey spans roughly from the publication of the *PDP* volumes in 1986 to the present. However, we begin by briefly describing the important ideas that preceded this time. We refer to these ideas as Wave 0 (1950–1985) as they span a wide historical period and represent an accumulation of models and events that created a foundation for subsequent work. During Wave 1 (1985–1995) many researchers focused on a narrow set of traditional neural networks as a core platform. Wave 2 (1995–2000) was a period of rapid growth in which the relatively homogeneous connectionist models of Wave 1 became elaborated. Finally, Wave 3 (2000–2010) simultaneously reflects the maturation of conventional neural networks and a substantial diversification, including the emergence of several important alternative paradigms.

3.1. Wave 0 (1950–1985): the prehistory of developmental models

At the advent of the cognitive revolution, older theories of learning as the basis of cognition were being discarded, largely because they did not admit internal representations (but see Hebb, 1960). In response, many researchers adopted an algorithmic view of the mind (the mind as computer), and symbolic models of development began to be established. Simon and Newell (1962; Newell & Simon, 1972) proposed a comprehensive framework that provided the foundation for modern information processing theory, and simultaneously designed the General Problem Solver (GPS), a computer program capable of reasoning and planning. The GPS is a performance (i.e., static) model, that is, it represents problem-solving behavior as a process of search through options and alternatives. While it was not specifically intended to model learning, Simon (1962) noted its potential as a model of cognitive development; subsequent work led to production systems, the SOAR model, and eventually, developmental models of planning, reasoning, and problem solving.

At the same time, classic learning approaches (Fig. 3A) were being extended beyond small sets of discrete stimuli to multiple banks of inputs/outputs with the introduction of the *perceptron*, a two-layer neural network proposed by Rosenblatt (1958) (Fig. 3B). Perceptrons enjoyed initial success but ten years later Minsky and Papert (1969) published a highly critical analysis demonstrating that perceptrons are unable to compute the exclusive-OR (XOR) rule, a relatively simple logical function. In particular, perceptrons are limited to learning input–output mappings that can be discriminated by a linear function; XOR, in contrast, requires a non-linear mapping in order to be computed. Minsky and Papert's finding not only confirmed for many the insufficiency of basic learning principles, but it also slowed research in artificial neural networks for nearly two decades, until Rumelhart et al. (1986) were able to solve this fundamental problem with the back-propagation algorithm, setting the stage for the connectionist revolution and Wave 1.

Rumelhart et al.'s solution included three essential innovations (Fig. 3C). First, they introduced a middle layer of units, which is called a *hidden layer* and is situated between the input layer (the sensory surface) and the output layer (the response system). While the input and output layers make

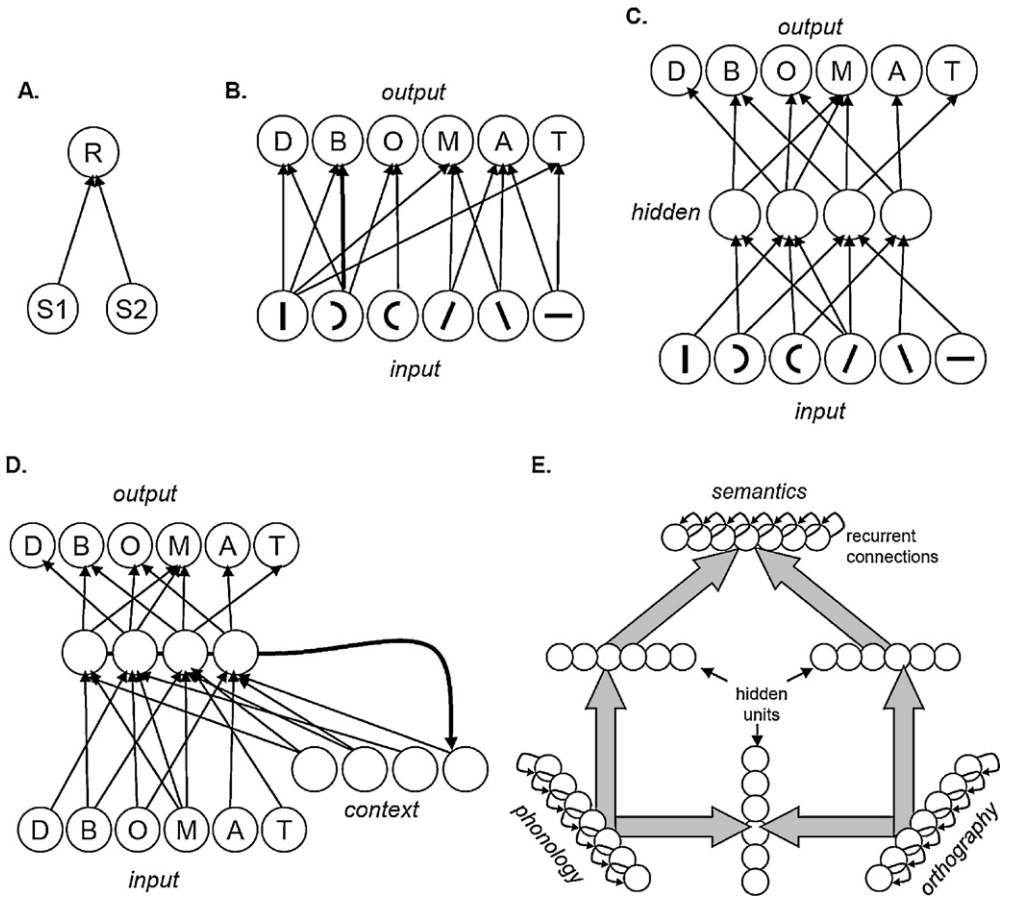

Fig. 3. The development of connectionist models of development. (A) Basic learning theory linked stimulus to response using simple, formal learning rules. (B) From such principles grew the perceptron, like this network mapping visual features onto letters. Perceptrons used many more inputs and outputs, and had a much more "cognitive" flavor, often dispensing with purely observable inputs and outputs. (C) Perceptrons failed to learn some mappings, leading Rumelhart and colleagues to develop three-layer networks in which hidden layers could be trained via back-propagation. (D) Limitations in the way that three-layer networks approached time led Elman to add context units and recurrence to give the network a short-term memory as in this network that learns sequences of letters by predicting the next letter. (E) Finally modern connectionist networks, such as Seidenberg and colleague's Triangle model often incorporate multiple banks of input and output units, multiple hidden units, and recurrence within layers to master multi-pathway systems like written word recognition.

contact with the physical world, the middle layer is buffered on each side and is therefore "hidden." Second, Rumelhart et al. modified the *activation function* used by the network. Perceptrons employ a linear activation function such that increasing the level of input into a unit linearly increases the output. In contrast, Rumelhart et al. proposed a non-linear, *sigmoidal* activation function, in which units have a maximum input threshold; when input is beyond this threshold, the firing rate asymptotes. Finally, and most importantly, they also proposed the *generalized delta rule*. This learning rule is based mathematically on an older learning-theoretic rule (Rescorla & Wagner, 1972). However, crucially, Rumelhart's back-propagation algorithm allowed this rule to distribute the training signal or error-feedback provided by the "teacher" throughout all of the connections in the network – including connections to the hidden layer – and across units with non-linear response functions.

Two important consequences followed from these improvements. First, Rumelhart et al. (1986) demonstrated that a 3-layer network could learn the XOR rule. Here, the hidden layer provides a

critical advantage over 2-layer perceptrons. By "recoding" the input, the hidden layer transforms the XOR decision task from a non-linear into a linear classifying task. Second, and more generally, it was subsequently demonstrated that a 3-layer network can learn any arbitrary (well-defined) non-linear function (Cybenko, 1989), solving a major critique of basic learning processes and offering an example of how virtually any complex behavior can emerge from basic learning principles distributed over many elements.

3.2. Wave 1 (1985–1995): connectionist nets

The first wave of developmental models took shape in the middle 1980s, largely inspired by the PDP volumes and the revolutionary power of back-propagation. The second volume, which focused on neural network models of specific tasks, included an influential chapter on learning the past tense of English verbs (Rumelhart & McClelland, 1986a), which may be one of the first truly *developmental* connectionist models. Rumelhart and McClelland described a network that was presented with the "stem" of English verbs (using a phonological representation), and was trained to produce the corresponding past-tense verb. Notably, the model not only learned to produce the past tense for regular verbs (help ⇒ helped), but it also learned the past tense forms for irregular verbs (go ⇒ went) using a single set of processing pathways.

This led to two insights. First, it challenged the long-held assumption that language knowledge must be represented by a set of rules and that development involves moving from one set of rules to another (Pinker, 1991). The model provided an existence proof that rule-like behavior – including transitions between putative rules – could occur in a system with only distributed, non-symbolic representations. Second, this model captured a key developmental pattern during learning – the model overregularized irregular verb forms that it initially learned to inflect correctly (e.g., "feel" was inflected as "feeled"). Like human children, the model produced a U-shaped pattern of development, even though such behavior was not explicitly built in. Subsequent work by Plunkett and Marchman (1991; see also McClelland & Patterson, 2002) demonstrated that this U-shaped pattern emerged "for free" from the model and was not the result of bias or systematic variation in the training sample. Perhaps most important, the model also generated a key developmental insight that transcended the model: *Overregularization* may not be a discrete developmental stage; rather it may be a micro-phenomenon that occurs repeatedly with the acquisition of new verbs in the lexicon. A subsequent analysis of corpus data supported this prediction (Marcus et al., 1992).

The past-tense model highlights a number of features typical of most first wave developmental models. In particular, they featured a 3-layer feedforward architecture and used backprop as the learning algorithm (Mareschal & French, 1997; McClelland & Jenkins, 1991; Quinn & Johnson, 1997). Like many other Wave 1 models, they also emphasized an ontogenetic timescale (e.g., early childhood).

McClelland and Jenkins (1991) offer a second example of a typical Wave 1 model. They used a 3-layer network to simulate the development of children's performance on Piaget's balance-scale task. Like the verb-learning model, this model also demonstrated that rule-like behavior can emerge from a system that encodes experience in a graded and distributed pattern. However, it provided two additional insights about development. First, McClelland and Jenkins adopted the set of four balance-scale rules identified by Siegler (1981), and asked whether model performance conformed to these rules. Rule 1, for example, states, "If the weights on each side of the scale are unequal, the side with more weight will tip." Model performance was consistent with these rules, providing further support for the idea that rule-based behavior can emerge from a system with distributed representations. Second, the model progressed through the rules in the same sequence as did children. This finding was novel and important, as it not only demonstrated a stage-like pattern of development in the model (despite continuous, quantitative learning rules), but also illustrated that a consistent series of stages can develop without explicitly training the stages in a particular order driven by the statistics of the problem-space (Raijmakers, van Koten, & Molenaar, 1996; Shultz et al., 1994).

The first wave of developmental models received a number of criticisms. Backprop was criticized in terms of biological plausibility as well as its assumption of error feedback during learning (Sejnowski, Koch, & Churchland, 1988; Shultz et al., 1995). A related concern focused on the relatively simple structure of 3-layer networks, which often performed sub-optimally in comparison to networks with

more elaborate structure. These issues provided an important influence that helped to shape the second wave of developmental models.

3.3. Wave 2 (1995–2000): differentiation and elaboration

By the mid-1990s, the basic 3-layer architecture was widely seen to be insufficient for solving some learning problems and for representing things like sequences. The models that emerged during the second wave – which largely remained within the sub-symbolic, connectionist class – addressed these concerns by dramatically expanding and elaborating on the existing approach to include new learning algorithms and architectures.

A key innovation was the use of recurrent networks (Elman, 1990, 1993). This innovation was motivated by a core property of language (and many other behaviors): time. In standard connectionist architectures, there is no clear way to represent time, and modelers adopted practices like duplicating the network at different times slices (McClelland & Elman, 1986) or converting time into space (e.g., having a copy of the input layer that is tuned to different times). This was unsatisfying in many ways (e.g., how many copies does one need?), and made it difficult to develop effective models of sequential behaviors.

Motivated by earlier work by Jordan (1986), Elman (1990) developed what is now known as the Simple Recurrent Network (SRN; Fig. 3D). SRNs are like conventional feedforward networks, with one modification: "Backward" connections are added to the network, so that activity from one layer in the network is routed either back (a) into the same layer (i.e., self-activation), or (b) to an earlier layer. Architecturally, this alteration was modest. However, functionally, this gave connectionist networks short-term or working memory for the first time.

In an SRN, the units that receive feedback from the hidden layer are called "context units" (Fig. 3D). While input units are normally activated by stimulation from outside the network, context units are stimulated by the internal state of the network on the last time-step, effectively maintaining an internal representation over time and offering a computational mechanism for working memory. A second innovation of the SRN was that rather than training them to achieve some pre-determined output, Elman trained the SRN to predict the next input pattern. This addressed a pervasive criticism of backprop networks – that the feedback these networks receive is far more detailed and precise than the kind of feedback human children experience. Elman demonstrated that SRNs could learn to solve a number of tasks that involved processing a series of input patterns over time (e.g., segmenting speech, parsing sentences).

SRNs have since been widely used as a sort of generic sequence learner, allowing a number of important theoretical advances that go well beyond the architecture. For example, Elman (1993) trained an SRN to predict the next word in a series of sentences. Interestingly, the network initially failed. However, when the context layer was modulated, so that its working memory capacity gradually increased over training, it succeeded. Essentially, the SRN started with a relatively short memory span and first learned to exploit short-range (within-phrase) grammatical dependencies. Then, as working memory increased, the network discovered longer-range dependencies, and consequently it learned to predict sequences of words in longer, more complex sentences. Elman used these simulations to highlight the "importance of starting small" in development (see also Newport, 1990). In particular, the model offered a persuasive demonstration that limitations in information-processing capacity could facilitate early learning (i.e., developmental immaturity; Bjorklund, 1997). In the next section, we provide a related example from motor learning that demonstrates the same principle (Schlesinger et al., 2000). Crucially, this shows how simple, well-used models like the SRN can help illustrate much broader theoretical points, and subsequent work in this vein has used them to examine issues like the developmental origin of working memory differences (MacDonald & Christiansen, 2002) and the interaction between evolution, language change, and development (Reali & Christiansen, 2009).

A second elaboration during Wave 2 was the use of modular networks (Jacobs, Jordan, & Barto, 1991; Rueckl, Cave, & Kosslyn, 1989). Like SRNs, the innovation offered by modular networks was architectural. The conventional 3-layer network was divided into two or more specialized processing "pathways," often with additional layers introduced. While the modular networks of the second wave

were still relatively far from biologically plausible, they were an important step toward designing and testing models that incorporated known anatomical structures and functions of the brain.

Two examples of modular networks are the object-processing models of Munakata et al. (1997), and of Mareschal et al. (1999). Both address an intriguing question: Why does object permanence appear to develop earlier when measured by gaze behavior than by reaching (Mareschal, 2000)? To investigate this question, both models use modular networks, in which one sub-network of the model controls gaze while another sub-network controls reaching, while a sub-network common to both contains object representations. In the Munakata model, the gaze sub-network is initially trained first, as a proxy for earlier development in gaze compared to reaching. As expected, the model anticipates the outcomes of occluded-object events using gaze first. While this is not surprising, the model provides a useful demonstration of how graded internal representations can be strong enough to guide gaze, but not enough to guide reaching.

The Mareschal model goes further by incorporating dual-stream visual processing (i.e., dorsal–ventral or "what" and "where" pathways; Milner & Goodale, 1995; Mishkin, Ungerleider, & Macko, 1983). The ventral sub-network simulates feature-based processing in the temporal cortex, while the dorsal sub-network simulates spatiotemporal processing in the parietal cortex. Unlike the Munakata model, the entire network is trained simultaneously. A key finding from this model is that the dorsal sub-network learns more rapidly than the ventral one. Mareschal et al. propose that while gaze recruits the "where" or dorsal system, reaching requires integrating both systems. Consequently, the developmental advantage of gaze over reaching behavior is due to the earlier emergence of the "where" system.

Finally, a third innovation during the second wave challenged the fixed architecture of 3-layer models with the use of cascade-correlation (Fahlman & Lebiere, 1990). The key feature of cascade-correlation networks is their ability to "grow" new hidden units over time. Thus, in contrast to conventional networks, cascade-correlation networks have a dynamic topology that adapts to the processing demands of training. Because cascade-correlation networks develop more elaborate internal encodings or representations over time, they can increase their representational power (Mareschal & Shultz, 1996; Shultz et al., 1995). Indeed, cascade-correlation networks have been successfully applied to a wide range of phenomena, including the balance-scale task, seriation, and conservation of number (Mareschal, 1992; Mareschal & Shultz, 1999; Shultz, 1998; Shultz et al., 1994), suggesting a new dimension on which connectionist networks can learn and suggesting ways to account for qualitative developmental change.

Perhaps the capstone of Wave 2 from a developmental perspective, however, was the publication of *Rethinking Innateness* (Elman et al., 1996), which built on the mounting complexity of connectionist networks and their increasing power to account for development. This volume used the logic of connectionism (and many simulations) to articulate a clear theoretical view of development, challenging whether constructs like modularity need be innate, putting forth a clearly emergentist view of development, and asking whether species-typical behavior need derive from innate domain-specific knowledge. While this book did not introduce new technical advances (e.g., types of modeling), its ability to translate connectionism to basic developmental theory (with a healthy dose of neuroscience) was a transitional moment in the development of modeling. In many ways, the purely theoretical view it articulated enabled researchers to adopt a connectionist perspective on development without running any models.

3.4. *Wave 3 (2000–2010): diversification*

While the second wave of developmental models emerged as a response to the homogeneity of the early connectionist networks, in the third wave entirely new theoretical and computational paradigms emerged, producing a wide diversity of learning algorithms, architectures, and modeling frameworks, each tied to new and different conceptualizations of development and cognition. Connectionism, of course, was not yet finished and significant development was seen. However, three new theoretical influences became important at this time.

First, there was a growing interest in conceptualizing cognition as embodied and situated (e.g., Bechtel, 1997; Schlesinger, 2001; Thelen et al., 2001). This led to dynamical-systems and agent-based

approaches. Second, the success of connectionist modeling coupled to empirical demonstrations of statistical learning (Gómez, 2002; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996; Yu & Smith, 2007) became a problem for symbolic accounts. This led to Bayesian formulations as a way of retaining the transparency and knowledge-based accounts of symbolic models while incorporating sensitivity to input statistics. Finally, theoretical ideas like statistical learning that were grounded in computational concepts (but not associated with modeling paradigms), together with the new diversity of approaches for doing computational modeling, led some researchers to distill core theoretical concepts into computational terms, but to implement them in ways that were somewhat independent of any particular modeling paradigm.

Across paradigms, however, there was also convergence. Prior models tended to represent both the child and the environment in a somewhat artificial manner. Now, model-builders designed models that corresponded more closely to real-world situations and to the scale of real-world development and tested their models in ways that aligned more closely with the experimental paradigms used to study real children (Christiansen & Chater, 2001; McMurray et al., in press; Perone, Spencer, & Schöner, 2007; Schlesinger, 2003; Xu & Tenenbaum, 2007).

During Wave 3, connectionism was (and is still) strong. However, two new trends have emerged. First, many researchers shifted from the development of new modeling architectures to deeper explorations of the mathematics and theoretical implication of existing ones (Elman, 2009). Second, a number of researchers began scaling up their models substantially. For example, the Triangle model of reading (Seidenberg & McClelland, 1989) was initially conceptualized as involving three bi-directional mappings between orthography, phonology, and semantics. However, the initial implementation only modeled the mapping from phonology to orthography in a single pass. More recent implementations (Fig. 3E; Harm & Seidenberg, 2004) have implemented all three pathways, using recurrence to model real-time decision making and demonstrating how multiple sources of information influence word reading. Similarly, Sibley, Kello, Plaut, and Elman (2008) taught a modified SRN nearly 75,000 word forms and discovered that such scale can be a benefit in enhancing generalization.

Beyond connectionism, dynamical systems began to offer competing computational models. Dynamic systems were a theoretical approach to development introduced by Thelen and Smith (1996) that emphasized non-linear dynamics, soft assembly between multiple coupled systems, embodiment of action and cognition, and the real-time incremental nature of development. Dynamic Field Theory, or DFT (Schutte et al., 2003; Spencer & Schöner, 2003; Thelen et al., 2001) became one of its most successful computational models, by building these concepts into a dynamical system based on neural population dynamics.

DFT was developed to investigate the idea that embodiment and real-time interaction with the environment could explain many developmental phenomena. DFT is implemented in a mathematical framework that captures real-time interactions among populations of neurons. Thus one of its strengths is its ability to capture unfolding behavior in real-time. Crucially, the DFT's commitment to embodiment means that it represents properties of the behavioral context (e.g., spatial location, object color) in a reasonably natural way that parallels the experience of a child engaged in a comparable task. For example, Schutte et al. (2003) use a DFT model to simulate children's performance on a "sandbox" version of the A-not-B search task. In this task, objects are hidden in A and B locations of a sandbox, creating a more continuous version of this classic Piagetian task. Uniquely, DFT encodes both the location of objects to be hidden and the rough geometry of the task space in the same perceptual and working memory fields. Schutte et al. (2003) show how the task space influences memory in this task and made (and confirmed) the startling prediction that in this more continuous version of the task, even four-year-olds show some error, searching on the correct (A) side of the sand-box but in a location that drifts to the other (B) location.

To model development, Schutte et al. (2003) hand-tuned parameters controlling the excitatory and inhibitory connections that influence the precision of representations in memory. In particular, spatial processing in young children was represented by large or broad excitatory connections; development in older children was represented by relatively narrower excitatory connections, with a corresponding increase in inhibitory connections. This successfully captured the developmental pattern of children's search performance between 2 and 6 years. Interestingly, linearly incrementing this same set of parameters can also give rise to qualitative differences in children's midline bias

in memory, showing drifts toward midline when young, and away when older (Schutte & Spencer, 2009).

Closely aligned with the DFT's goal of simulating embodied real-time interactions and constraints in the task-space is a class of models that simulates the child as an autonomous agent (Schlesinger, 2003). Rather than focusing on a specific cognitive architecture or learning algorithm, this approach instead highlights organism–environment interaction and emphasizes the importance of simulating the child's cognitive processing as well as the body and the physical environment. The notions of "body" and "environment" are flexible and may vary widely across models, such as an eye and arm that learn to coordinate their movements to reach a nearby object (Schlesinger et al., 2000), or an eye that learns to visually track objects that move in and out of view (Schlesinger & Parisi, 2001), or arm/joint dynamics that are well specified in terms of weight, torsion (Berthier et al., 2005). While much of this work reflects bodies and environments that remain stable over time, we describe in the final section the emerging field of developmental robotics, which is capable of modeling not only physical growth but also real-world, dynamic environments.

Agent-based models offer several features not typically exploited by other models. First, because agent-based models can simulate both single-agent and multi-agent scenarios, they are well-suited to investigating questions in which social interaction plays a role. Second, in addition to capturing changes over ontogenetic time, agent-based models can also simulate the evolutionary or phylogenetic timescale (Schlesinger, 2004; Schlesinger & Parisi, 2007). For example, Schlesinger et al. (2000) use an agent-based model to simulate the development of hand–eye coordination in young infants. Like infants, the model develops reaching skill in a proximodistal pattern: At both the muscular and joint levels, control is established over central body segments before peripheral segments (e.g., shoulder joint vs. elbow joint). This finding complements Elman's (1993) starting-small principle, but it emerges autonomously, rather than through hand-tuning. Thus, in a complex task environment with multiple degrees of freedom, the proximodistal "bias" need not be built in: If children are trying reach in an increasingly optimal way, this constraint may be an obligatory way station through the search space.

In contrast to these sub-symbolic approaches loosely aligned with connectionism, there were also advances among symbolic models. The growing understanding of statistical learning combined with the assumptions of symbolic models (e.g., innate structure, rule-based performance) led to a novel hybrid: Bayesian models (Frank, Goodman, & Tenenbaum, 2009; Kemp, Perfors, & Tenenbaum, 2007; Xu & Tenenbaum, 2007). Bayesian models focus on an informational (rather than a mechanistic) account of cognition. Like other symbolic models, they operate at a higher level of abstraction than connectionist networks (Jones & Love, 2011). However, they share with connectionism the notion that statistical dependencies between elements in the input (e.g., visual features, linguistic labels) are fundamental for learning. Consequently, the central task for Bayesian learners is to establish probability estimates for the occurrence of particular events and to update these estimates with experience. In Bayesian systems, probabilities are acquired as knowledge used to represent the world in general. However, the implicit assumption is made that development is geared toward acquiring knowledge about the true state of the world, as opposed to developing knowledge that is needed for action and implementing it in these same action systems (as in connectionism).

As a concrete example, consider the problem of learning to label a set of novel objects that vary in shape, size, and color. When presented with this task, children not only learn to label the objects rapidly based on only a few examples; they also consistently rely on shape as a basis for grouping objects together with the same label – the shape bias (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). In a Bayesian model of this word-learning task, Kemp et al. (2007) show that across a diverse set of objects, presenting two objects with the same shape and label (e.g., the round, green, smooth object and the round, orange, fuzzy object are both "blickets") can induce a shape bias to new unlabeled objects, explaining generalization as a lawful inference over the data thus far encountered. Thus, the Bayesian approach not only highlights children's sensitivity to statistical regularities; it also provides an account of how these regularities provide a structure for interpreting new experiences.

Finally, most of the modeling we have described so far is strongly paradigmatic. A set of principles such as those of connectionism, DFT, or Bayesian approaches constrain what can be built into the model, and in many ways the model itself is an answer to the question, "Can some complex developmental phenomena derive from such principles?" In contrast to this paradigmatic approach, during

Wave 3 a number of researchers began implementing simple computational tools to test implications of more qualitative developmental theory. This was in some ways a response to the growing complexity of more paradigmatic models and the difficulty in answering basic theoretical questions when it is unclear which components of the model are responsible.

For example, Werker et al. (2006) reasoned that if infants learn speech categories using the statistical distribution of acoustic cues, they could model the ideal learner using logistic regression (see also McMurray & Jongman, 2011). This allowed them to ask the simple question of whether the statistics of infant directed speech are sufficient to distinguish vowel categories in the limit. In some ways this represents a sort of ideal learner, as logistic regression employs supervised learning, while infants more likely engage in some sort of unsupervised process (Maye et al., 2002). Nonetheless, they found that the model was able to learn to classify vowels and to weight language-specific cues appropriately, lending strong credence to statistical learning approaches to speech development.

Similarly, McMurray, Aslin, et al. (2009) examined the mechanism of such learning. They argued that infants must learn using (1) unsupervised learning and (2) iterative learning in which learning occurs after each presentation of a stimulus. (Most Bayesian learning systems learn in "batch mode" from all the data at once, as opposed to trial-by-trial.)

They started from a Bayesian framework (a mixture of Gaussians), but found it insufficient. However, adding a connectionist-style competition to the network allowed it to succeed and model the time course of development. Critically, a range of connectionist architectures applied to the same problem also requires competition between categories (McMurray, Aslin, et al., 2009; McMurray, Horst, et al., 2009). That is, learning systems must do more than just accumulate statistics over inputs; on each input it must attempt to identify a category, suppress competing categories and then learn. This need for competition (or some real-time or trial-time mechanism for suppressing alternative interpretations) may be a general requirement of unsupervised categorization.

A number of models have also used this more paradigm-neutral approach to isolate and investigate ontogenetic timescales. van Geert (1998), for example, implemented the Piagetian constructs of assimilation and accommodation as simple differential equations. By implementing many simultaneous assimilation and accommodation events and iterating over time, he showed how both qualitative and quantitative change could emerge from the same quantitative system. Similarly, McMurray (2007; Mitchell & McMurray, 2009) reasoned that all word learning models are essentially accumulators, accumulating evidence for many words in parallel. They implemented and analyzed a series of models in which each word simply acquires a "point" on each trial until it reaches some threshold for learning. This demonstrated that as long as multiple words are acquired in parallel, and words vary in difficulty, the sudden increase in the rate of word learning during the second year (the vocabulary spurt) is mathematically guaranteed and thus is not evidence for any specialized change on the part of the child.

4. The past, present, and future of developmental models

4.1. Covering the theoretical/empirical landscape

Our taxonomy of modeling paradigms not only offers a classification scheme for organizing a rapidly expanding field of modeling architecture; it also illustrates something of greater value – think of a research question and the existing array of modeling approaches can likely handle it. But beyond the descriptions of how a model works (e.g., timescale, static/adaptive), the space of computational models of development has expanded such that several features of the theoretical landscape are now well-covered.

Modeling frameworks correspond to each of the major theoretical orientations – nativism, constructivism, and empiricism. Connectionist models align well with both empiricist and constructivist views, and Bayesian models are an appropriate tool for examining the influence of a priori (innate) knowledge on learning. Existing modeling approaches also span the timescales of behavioral change, from milliseconds and minutes up to days, months, and years.

As our examples illustrate, computational models cut across the full range of behaviors that fall within the domain of cognitive development, including sensorimotor skill, object perception and

categorization, speech perception and language acquisition, as well as models of planning, problem-solving, decision-making. Of course, as Braitenberg (1984) notes, no model is a model of all these behaviors. Instead, most models remain tractable (and offers clearer theoretical points) by focusing on a specific developmental domain, time scale, skill, or behavior. While there is clear value in bridging across time scales and domains (McMurray, Aslin, et al., 2009; McMurray, Horst, et al., 2009), even here an element of simplicity, or distillation is often necessary to make theoretical advances.

4.2. What do models add?

While developmental models have become progressively more elaborate (and sometimes, more complex), an open question is whether this trajectory reflects a path toward greater explanatory power. In the historical context we have outlined, each wave of models is not only a response to the limitations or shortcomings of previous models, but also a reflection of the goals and priorities of the developmental community itself. For example, an increase in the use of eye-tracking measures over the last several years has also increased the value of computational models that simulate eye-movements in real-time (Schlesinger et al., 2007).

Developmental models can have a crucial role as part of the scientific-deductive process. In particular, several recurring themes can be seen across the examples here. First, each model is an implementation of a specific theory, and when successful, provides a demonstration proof that the theoretical account is plausible or possible. That on its own is important. Second, there were occasions where models failed or were less capable of capturing a particular developmental phenomenon than an alternative model (e.g., models of past-tense, grammar, phoneme categorization and the balance-scale task). Thus, simulation data provide a key metric for comparing models, and indeed, model-builders use these data to refine and improve their models. Third, several models have provided a new perspective on the developmental phenomenon under study, sometimes suggesting a novel methodological manipulation or a counterintuitive hypothesis later confirmed experimentally (e.g., the sandbox search task investigated by Schutte et al., 2003). Perhaps most important, these models have allowed us to answer questions about developmental theory itself and occasionally revealed unexpected consequences of our theoretical assumptions (e.g., the vocabulary explosion as an unintended consequence of parallel learning).

Across just the limited examples here, we see examples of clear theoretical advances that lie well beyond a specific modeling framework – concepts and theories that could be implemented in multiple ways and are of value to the non-modeling community (see Table 2). In fact, if there were a theme to emerge across multiple models and multiple paradigms, it is this: Development often derives from non-obvious causes – hidden factors, unexpected non-linearities and subtle interactions between organism and environment. Computational modeling can often illuminate sources of non-obvious causation. Indeed, it was the surprising success of many connectionist models to capture small sub-regularities of the input that led to statistical learning, a theoretical construct that goes well beyond connectionism. Many of these models illustrate how incremental, unidirectional change in a critical model parameter can give rise to qualitative changes in the child – U-shaped curves, acceleration and the like. This is not something that can be observed from empirical data alone and represents a powerful constraint on interpretation.

4.3. Where are developmental models headed?

A number of important trends will help shape the direction of the field over the next decade. One is the emergence of developmental robotics (Cangelosi et al., 2010; Lungarella, Metta, Pfeifer, & Sandini, 2003; Schlesinger, 2009; Vernon, Metta, & Sandini, 2007; Weng et al., 2001). In contrast to classical AI systems, which are engineered to solve a particular problem and are given the requisite skill and knowledge in advance, the goal of developmental robotics is to design autonomous agents that have as little knowledge in advance as possible and must learn and develop skill through exploration and interaction with the environment. Therefore, developmental robotics is an extension of the agent-based approach described earlier, including not only a focus on embodiment, organism–environment interaction, and emergent knowledge, but also the use of both simulated and real-world robotic

Table 2
Summary of theoretical advances made from the models discussed in this article.

Paper	Architecture	Finding
General		
Rumelhart et al. (1986)	Three-layer backprop	Internal representations are required for some simple problems.
Language development		
Rumelhart and McClelland (1986a)	Three-layer backprop	Existence proof that rule-like behavior could occur in a system with only distributed, non-symbolic representations. U-shaped development need not represent a change in the learning mechanism or a switch in representational formats; it may derive from systematicities in the statistics of the task space.
Elman (1993)	Simple recurrent network	Limited memory capacity can be beneficial for finding structure.
McMurray (2007)	Non paradigmatic: accumulator	Accelerating learning is a necessary by-product of parallel learning and not a sign of any qualitative developmental change.
McMurray, Aslin, et al. (2009) and McMurray, Horst, et al. (2009)	Mixture of Gaussians (statistical learning)	Online competition may be needed for statistical learning to find categories.
Visual cognition		
Munakata et al. (1997)	Modular neural network	A graded object concept can account for differences between tasks and ages in object permanence tasks.
Mareschal et al. (1999)	Modular neural network	Differential development of what/where pathway can account for differences across tasks/ages in object permanence tasks.
Schutte et al. (2003)	Dynamic Field Theory	Linear changes in precision of neural fields can lead to qualitative differences in spatial memory.
Motor development		
Schlesinger et al. (2000)	Agent-based	Constraints on early motor control need not be built in but can arise as a more optimal response to a complex task-space.
Berthier et al. (2005)	Cerebellar network model with arm dynamics and RL	Variability (noise) and spontaneous movements may help infants explore the space of motor movements and lead to better learning.

platforms (e.g., humanoid robots, wheeled robots, "insect" robots) for designing and testing computational models of development.

A second influence is computational developmental neuroscience (Thivierge, 2010). This influence reflects continuing pressure toward biologically plausible models of cognition and development and a trend toward representing neural processes in a way that better reflects known properties of the mammalian brain. Recent examples include work on the role of dopamine and reward-learning during the development of gaze-following (Triesch, Teuscher, Deák, & Carlson, 2006), as well as the influence of growth in the posterior parietal cortex on the development of visual selective attention (Schlesinger et al., 2007) and models of the cerebellum and joint mechanics being applied to motor control (Berthier et al., 2005).

Finally, a third and equally important influence on developmental models is the study of individual differences. Thus far the majority of computational models of development have focused on a normative perspective. In contrast, relatively few models highlight the diversity of developmental pathways, and in particular, the biological and environmental factors that give rise to developmental pathology (Harm, McCandliss, & Seidenberg, 2003; Harm & Seidenberg, 1999; McMurray et al., 2010; Thomas & Karmiloff-Smith, 2002). We anticipate that the developmental community will see increased attention in this area, not only because it is a comparatively understudied area, but also because researchers in robotics have begun to investigate developmental psychopathology (Dautenhahn & Werry, 2004; Kozima & Yasuda, 2007).

4.4. Closing the gap between theory and research

In tracking the arc of *design* ⇒ *test* ⇒ *refine* characteristic of the modeling enterprise, we note a subtle but significant shift in emphasis over the last 25–30 years. In the first wave, developmental models were largely descriptive, enabling researchers to state and test theories in an explicitly mechanistic language, and, where successful, demonstrate that the theory in principle works. The *test* and *refine* steps were focused on getting the model to reproduce the target phenomenon. During the second and third waves, the focus has gradually shifted away from developmental models as confirmatory tools, and has emphasized their use as tools for inquiry and experimentation. Models now provide a bridge in both directions between theory and research. Models not only validate theories, but also provide feedback that helps to revise theories. At the same time, models not only reproduce core findings from critical experiments; they also generate new predictions and suggest novel tests.

To illustrate the closed loop between developmental theory, behavioral research, and computational modeling, we highlight two examples. The Munakata et al. (1997) model, recall, which simulates gazing and reaching behavior toward objects, learns to produce an internal representation of the target object, which gradually decays while the target is occluded. Although this occluded-object representation is too weak to accurately guide reaching behavior, it is strong enough to guide gaze behavior. This model started with a familiar idea – internal representations are graded – but also broke new ground. It provided a compelling demonstration of how graded representations could be implemented and studied in a modular neural network. It also demonstrated how a developmental gap between reaching and gazing could emerge as a function of the relative strength of occluded-object representations. More important, these simulation findings helped reshape the ongoing debate on early object knowledge, by highlighting the role of infants' graded representations (Baillargeon, 2000; Haith, 1998; Munakata, 2000; Smith, 1999; Spelke, 1998).

A second example is the paradigm shift that occurred in language learning in the 1990s. A lasting generalization from the dozens (or hundreds) of connectionist models in the first and second waves of modeling was the powerful role that very subtle statistics can play in giving rise to useful language learning. Statistics like the transition probabilities between words, the distribution of phonetic cues across utterances, or the quasi-regular mapping between elements (e.g., letters and phonemes) were the grist on which such models learned, and these early models taught us how powerful these statistics could be in the hands of such sensitive learning devices. At the time, work in language was largely ignoring such processes based on the input, in favor of constructs like parameter setting and innate constraints.

In contrast, connectionist models as a whole revealed how powerful such learning could be. This influence reached a critical point when Saffran et al. (1996) developed the artificial language learning paradigms, which demonstrated that human infants and adults (Saffran, Newport, Richard, Tunick, & Barrueco, 1997) could learn via such statistics. This has turned into an incredibly fruitful enterprise with dozens of empirical studies (Saffran & Thiessen, 2007, for a review) and numerous theoretical accounts based on these ideas. However, the theoretical development did not stop there. Researchers' struggle with these results led to the view that statistical learning needed to be constrained to be effective (Newport & Aslin, 2004; but see Spencer, Blumberg, et al. 2009). This theoretical notion of combining statistical learning to the constraints and structure offered by information processing led in no small part to the Bayesian models now important in language study. Thus, we see at multiple levels how individual models make specific predictions that can refine and test both model and theory, and how whole modeling paradigms shape theory and empirical work, which in turn can shape approaches to modeling.

We close by noting, once again, that computational modeling cannot replace theory in developmental science. In fact, as the foregoing discussion makes clear, the most important contribution of models is to push theory in new directions and to examine the non-obvious consequences of our theories. As our models become increasingly powerful and complex, it is equally important that they become increasingly accessible and understood, in order to contribute to the interplay between model and theory. As Braitenberg (1984) noted, modeling is at its best when it allows us to understand the range of possible mechanisms that can give rise to a behavior and why. In this light, modeling can allow us to ask questions about our theories themselves and discover new consequences to sometimes very

old theoretical views. A number of models we have surveyed capture the notion that variation is essential for learning (Apfelbaum & McMurray, 2011; Berthier et al., 2005; Schlesinger, 2004), an idea that harkens back to basic learning theory (Bourne & Restle, 1959; Bush & Mosteller, 1951) – variability in the irrelevant elements can highlight the invariant relationships. The same is clearly true with respect to computational models – only by examining and comparing many types of models will we learn what is irrelevant and what is relevant for understanding development.

References

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Apfelbaum, K., & McMurray, B. (2011). Successes and failures in early word learning: An emergent property of basic learning principles. *Cognitive Science*, *35*(6), 1105–1138.

Baillargeon, R. (2000). Reply to Bogartz, Shinskey, and Schilling; Schilling; and Cashon and Cohen. *Infancy*, *1*, 447–462.

Bechtel, W. (1997). Embodied connectionism. In D. M. Johnson, & C. E. Erneling (Eds.), *The future of the cognitive revolution* (pp. 187–208). New York: Oxford University Press.

Berthier, N., Rosenstein, M. T., & Barto, A. G. (2005). Approximate optimal control as a model for motor learning. *Psychological Review*, *112*(2), 329–346.

Bjorklund, D. F. (1997). The role of immaturity in human development. *Psychological Bulletin*, *122*, 153–169.

Bourne, L. E., & Restle, F. (1959). Mathematical theory of concept identification. *Psychological Review*, *66*, 278–296.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological review*, *58*(6), 413–423.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., et al. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, *2*, 167–195.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, *5*, 82–88.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy. *Pragmatics and Cognition*, *121*, 1–35.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences*, *9*, 111–117.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 1–36.

Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness. A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (pp. 524–532). Los Altos, CA: Moran Kaufmann.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Goldstone, R. L., & Medin, D. L. (1994). The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 29–50.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*, 281–287.

Gottlieb, G. (1997). *Synthesizing nature-nurture: Prenatal roots of instinctive behavior*. Mahwah: Lawrence Erlbaum Associates.

Guenther, F., & Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*, 1111–1112.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development*, *21*, 167–179.

Halford, G. S., Smith, S. B., Dickson, J. C., Mayberry, M. T., Kelly, M. E., Bain, J. D., et al. (1995). Modeling the development of reasoning strategies: The roles of analogy, knowledge, and capacity. In T. Simon, & G. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 77–156). Hillsdale, NJ: Lawrence Erlbaum Associates.

Harm, M., McCandliss, B., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading*, *7*(2), 155–182.

Harm, M., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hebb, D. O. (1960). The American Revolution. *American Psychologist*, *15*(12), 735–745.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, *15*, 219–250.

Johnston, T. D., & Edwards, L. (2002). Genes, interactions, and the development of behavior. *Psychological Review*, *109*, 26–34.

Jones, G., Ritter, F. E., & Wood, D. J. (2000). Using a cognitive architecture to examine what develops. *Psychological Science*, *11*, 93–100.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach*. Institute for Cognitive Science Report 8604. University of California, San Diego.
Karmiloff-Smith, A. (1992). Nature, nurture, and PDP: *P*reposterous *D*evelopmental *P*ostulates? *Connection Science*, *4*, 253–269.
Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.
Klahr, D., Langley, P., & Neches, P. (1984). *Production system models of learning and development*. Cambridge, MA: MIT Press.
Klahr, D., & MacWhinney, B. (1998). Information processing. In D. Kuhn, & R. S. Siegler (Eds.), *Handbook of child psychology* (pp. 631–678). New York: Wiley and Sons.
Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese, & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 61–116). New York: Academic Press.
Kozima, C., & Yasuda, Y. (2007). Children–robot interaction: A pilot study in autism therapy. *Progress in Brain Research*, *164*, 385–400.
Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: A survey. *Connection Science*, *15*, 151–190.
MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54.
Marcus, G. F., Ullman, M., Pinker, S., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*.
Mareschal, D. (1992). *A connectionist model of the development of children's seriation abilities*. Unpublished masters thesis. Montreal: Department of Psychology, McGill University.
Mareschal, D. (2000). Object knowledge in infancy: Current controversies and approaches. *Trends in Cognitive Sciences*, *4*, 408–416.
Mareschal, D., & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. In M. G. Shafto, & P. Langley (Eds.), *Proceedings of the nineteenth annual conference of the Cognitive Science Society* (pp. 484–489). London: Erlbaum.
Mareschal, D., & Johnson, S. P. (2002). Learning to perceive object unity: A connectionist account. *Developmental Science*, *5*, 151–172.
Mareschal, D., Plunkett, K., & Harris, P. (1999). A computational and neuropsychological account of object-oriented behaviours in infancy. *Developmental Science*, *2*, 306–317.
Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, *11*, 571–603.
Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science*, *11*, 149–186.
Mareschal, D., & Thomas, M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, *11*, 137–150.
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 101–111.
McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In T. J. Simon, & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 157–204). Hillsdale, NJ: Lawrence Erlbaum.
McClelland, J. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*, 751–770.
McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.
McClelland, J. L., & Jenkins, E. (1991). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. van Lehn (Ed.), *Architectures for intelligence* (pp. 41–73). Hillsdale, NJ: Lawrence Erlbaum Associates.
McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences*, *6*, 465–472.
McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631.
McMurray, B., Aslin, R. N., & Toscano, J. (2009). Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science*, *12*(3), 369–379.
McMurray, B., Horst, J., & Samuelson, L.S. (in press). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*.
McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (2009). Towards an integration of connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. New York: Oxford University Press.
McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–246.
McMurray, B., Samelson, V., Lee, S., & Tomblin, J. B. (2010). Eye-movements reveal the time-course of online spoken word recognition language impaired and normal adolescents. *Cognitive Psychology*, *60*, 1–39.
Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. New York: Oxford University Press.
Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two central pathways. *Trends in Neuroscience*, *6*, 414–417.
Mitchell, C., & McMurray, B. (2009). On leveraged learning in lexical acquisition and its relationship to acceleration. *Cognitive Science*, *33*(8), 1503–1523.
Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the AnotB task. *Developmental Science*, *1*, 161–211.
Munakata, Y. (2000). Challenges to the violation-of-expectation paradigm: Throwing the conceptual baby out with the perceptual processing bathwater? *Infancy*, *1*, 471–477.
Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*, 413–429.
Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713.

Newell, A. (1990). *A unified theory of cognition*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.
Newport, E., & Aslin, R. N. (2004). Learning at a distance. I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.
Oakes, L. M., Newcombe, N. S., & Plumert, J. M. (2009). Are dynamic systems and connectionist approaches an alternative to Good Old Fashioned Cognitive Development? In J. P. Spencer, M. Thomas, & J. McClelland (Eds.), *Dynamic systems and connectionist approaches to development* (pp. 268–284). New York: Oxford University Press.
Oyama, S., Griffiths, P. E., & Gray, R. D. (Eds.). (2001). *Cycles of contingency: Developmental systems and evolution.*. Cambridge: MIT Press.
Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*, 302–321.
Perone, S., Spencer, J. P., & Schöner, G. (2007). A dynamic field theory of visual recognition in infant looking tasks. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the twenty-ninth annual Cognitive Science Society* (pp. 1391–1396). Nashville, TN: Cognitive Science Society.
Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.
Pitt, M., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83.
Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multilayered perceptron: Implications for child language acquisition. *Cognition*, *38*, 1–60.
Plunkett, K., & Sinha, C. G. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, *10*, 209–254.
Quartz, S., & Sejnowski, T. J. (1998). The neural basis of cognitive development: A constructivist manifesto. *Behavioral & Brain Sciences*, *20*, 537–596.
Quinlan, P. T. (Ed.). (2003). *Connectionist models of development: Developmental processes in real and artificial neural networks.*. London: Taylor and Francis.
Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist account. *Journal of Experimental Child Psychology*, *66*, 236–263.
Raijmakers, M. E. J., van Koten, S., & Molenaar, P. C. M. (1996). On the validity of simulating stagewise development by means of PDP networks: Application of catastrophe analysis and an experimental test of rule-like network performance. *Cognitive Science*, *20*, 101–136.
Reali, F., & Christiansen, M. H. (2009). On the necessity of an interdisciplinary approach to language universals. In M. H. Christiansen, C. Collins, & S. Edelman (Eds.), *Language universals* (pp. 266–296). New York: Oxford University Press.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.
Rueckl, J. G., Cave, K. R., & Kosslyn, S. M. (1989). Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience*, *1*, 171–186.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the Microstructure of Cognition no. 1: Foundations* (pp. 318–362). Cambridge: MIT Press.
Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the Microstructure of Cognition no. 1: Foundations*. Cambridge, MA: MIT Press.
Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, & P. D. P. Research Group (Eds.), *Parallel distributed processing: Explorations in the Microstructure of Cognition no. 2: Psychological and biological models* (pp. 216–231). Cambridge: MIT Press.
Saffran, J. R., Aslin, R. N., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
Saffran, J. R., Newport, E. A., Richard, N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105.
Saffran, J. R., & Thiessen, E. D. (2007). Domain-general learning capacities. In E. Hoff, & M. Shatz (Eds.), *Handbook of language development* (pp. 68–86). Cambridge, UK: Blackwell.
Schlesinger, M. (2001). Building a better baby: Embodied models of infant cognition. *Trends in Cognitive Sciences*, *5*, 139.
Schlesinger, M. (2003). A lesson from robotics: Modeling infants as autonomous agents. *Adaptive Behavior*, *11*, 97–107.
Schlesinger, M. (2004). Evolving agents as a metaphor for the developing child. *Developmental Science*, *7*, 158–164.
Schlesinger, M. (2009). The robot as a new frontier for connectionism and dynamic systems theory. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered* (pp. 182–199). New York: Oxford University Press.
Schlesinger, M., Amso, D., & Johnson, S. P. (2007). The neural basis for visual selective attention in young infants: A computational account. *Adaptive Behavior*, *15*, 135–148.
Schlesinger, M., & Parisi, D. (2001). The agent-based approach: A new direction for computational models of development. *Developmental Review*, *21*, 121–146.
Schlesinger, M., & Parisi, D. (2004). Beyond backprop: Emerging trends in connectionist models of development. *Developmental Science*, *7*, 131–132.
Schlesinger, M., & Parisi, D. (2007). Connectionism in an Artificial Life perspective: Simulating motor, cognitive, and language development. In D. Mareschal, S. Sirois, G. Westermann, & M. H. Johnson (Eds.), *Neuroconstructivism: Vol. 2. Perspectives and prospects* (pp. 129–158). Oxford, UK: Oxford University Press.
Schlesinger, M., Parisi, D., & Langer, J. (2000). Learning to reach by constraining the movement search space. *Developmental Science*, *3*, 67–80.

Schutte, A. R., & Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1698–1725.
Schutte, A. R., Spencer, J. P., & Schöner, G. (2003). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development*, *74*, 1393–1417.
Seidenberg, M. S. (2005). Connectionist models of word reading. *Current Directions in Psychological Science*, *14*, 238.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of visual word recognition and naming. *Psychological Review*, *96*, 523–568.
Sejnowski, T., Koch, K., & Churchland, P. (1988). Computational neuroscience. *Science*, *241*, 1299–1306.
Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science*, *1*, 103–126.
Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.
Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, *16*, 57–86.
Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a connectionist algorithm. In T. J. Simon, & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 205–261). Hillsdale, NJ: Lawrence Erlbaum.
Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2008). Large-scale modeling of wordform learning and representation. *Cognitive Science*, *32*, 741–754.
Siegler, R. S. (1981). Developmental sequences with and between concepts. *Monographs of the Society for Research in Child Development*, 46.
Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
Simon, H. A. (1962). An information processing theory of intellectual development. *Monographs of the Society for Research in Child Development*, *27*, 150–161.
Simon, H. A., & Newell, A. (1962). Computer simulation of human thinking and problem solving. *Monographs of the Society for Research in Child Development*, *27*, 137–150.
Simon, T. J. (1998). Computational evidence for the foundations of numerical competence. *Developmental Science*, *1*, 71–78.
Simon, T. J., Hespos, S. J., & Rochat, P. (1995). Do infants understand simple arithmetic? A replication of Wynn (1992). *Cognitive Development*, *10*, 253–269.
Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An auto-associator perspective. *Developmental Science*, *4*, 442–456.
Smith, L. B. (1999). Do infants possess innate knowledge structures? The con side. *Developmental Science*, *2*, 133–144.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13–19.
Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, *21*, 181–200.
Spencer, J., Blumberg, M., McMurray, B., Robinson, S., Samuelson, L., & Tomblin, J. B. (2009). Short arms and talking eggs: Why we should no longer abide the nativist-empiricist debate. *Child Development Perspectives*, *3*(2), 79–87.
Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*, 392–412.
Spencer, J., Thomas, M., & McClelland, J. (Eds.). (2009). *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. New York: Oxford University Press.
Spivey, M. (2007). *The continuity of mind*. New York: Oxford University Press.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, *24*, 1–86.
Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: The MIT Press.
Thivierge, J. P. (2010). Computational developmental neuroscience: Capturing developmental trajectories from genes to cognition. *IEEE Transactions on Autonomous Mental Development*, *2*, 51–58.
Thomas, M. S. C., & Karmiloff-Smith, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modeling. *Behavioral and Brain Sciences*, *25*, 727–788.
Triesch, J., Teuscher, C., Deák, G., & Carlson, E. (2006). Gaze following: Why (not) learn it? *Developmental Science*, *9*, 125–157.
van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review*, *105*, 634–677.
Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, *11*, 151–180.
Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Artificial intelligence: Autonomous mental development by robots and animals. *Science*, *291*, 599–600.
Werker, Janet, F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., et al. (2006). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*, 147–162.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749–750.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.
Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.