International Journal of Neural Systems, Vol. 12, No. 6 (2002) 447–465

INCREMENTAL LEARNING IN BIOLOGICAL AND MACHINE LEARNING SYSTEMS

STEPHAN K. CHALUP
School of Electrical Engineering and Computer Science,
The University of Newcastle, Australia
chalup@cs.newcastle.edu.au

Received 9 May 2002
Revised 23 September 2002
Accepted 25 September 2002

Incremental learning concepts are reviewed in machine learning and neurobiology. They are identified in evolution, neurodevelopment and learning. A timeline of qualitative axon, neuron and synapse development summarizes the review on neurodevelopment. A discussion of experimental results on data incremental learning with recurrent artificial neural networks reveals that incremental learning often seems to be more efficient or powerful than standard learning but can produce unexpected side effects. A characterization of incremental learning is proposed which takes the elaborated biological and machine learning concepts into account.

Keywords: Neurodevelopment; critical periods; incremental learning.

1. Introduction

Incremental learning is a concept with several dimensions. It has been associated with learning processes where a standard learning mechanism is combined with or is influenced by stepwise adjustments during the learning process.[47,80] These adjustments can be changes in the structure or parameters of the learning system or changes in the presentation or constitution of the input signals. Adjustments of this kind are inherent to many biological learning mechanisms, too. Therefore, we regard incremental learning as a general learning concept and our review will cover biological and machine learning mechanisms with the aim to increase our understanding of both. This article addresses researchers who are interested in correspondences between the biological and machine learning aspects of incremental learning.

In the area of machine learning the term "incremental learning" has alternatively been used synonymously with "pattern learning" and "online learning" to describe the opposite of "batch learning". However, as we will see later in Sec. 5, this characterization is not precise enough.

Incremental learning methods have been proposed as candidates to model learning mechanisms of the *sensitive periods* of the developing brain.[80] These sensitive periods, sometimes also called *critical periods*, are within the first few years of childhood, between birth and puberty.[96] During the sensitive periods massive learning takes place and a large set of mechanisms and programs collaboratively and efficiently set up, tune and refine the developing central nervous system. Accordingly, corresponding incremental machine learning algorithms are expected to be sophisticated and powerful.

An almost classical example which supports this claim is the study of Elman[47] on language processing with artificial neural networks. With incremental learning Elman's neural networks were able to achieve results which could not be observed in the same setting without incremental learning. However, the experiments of Rohde and Plaut[119,120] drew Elman's claim of the superior abilities of incremental

learning into question. Therefore, it is apparent that the concept and algorithms of incremental learning still require more exploratory investigation.

To improve our understanding of biological and machine incremental learning, the present article reviews aspects of biological learning mechanisms which are incremental. The aim is to motivate and find a characterization for corresponding machine incremental learning that captures the essence of the biological counterpart as well.

The biological nervous system, and in particular neural connections in the brain, develop according to a well-organized plan.[110] The mechanisms of neurodevelopment build and set up the initial architecture, which then is tuned and refined by a variety of adaptive or learning mechanisms.[77] The fundamental organizational plan for all this activity is encoded in the genes and is a product of long-term phylogenetic adaptive processes of the genome. From the viewpoint of an individual, the biological structure and functionality of the central nervous organ (brain) has therefore been established or 'learned' in three time phases:

(I) Evolution: Evolution operates on the largest time scale of our model. Significant evolutionary changes in a population take many thousands of years, while every individual only lives for an almost negligibly short period of time. During the time phase of evolution the structure of the genome undergoes a process of *phylogenetic learning* which is based on evolutionary concepts such as selection and mutation.

(II a,b) Neurodevelopment: Within the prenatal (IIa) and the sensitive (postnatal) phases (IIb) of neurodevelopment the neural structure of an individual brain is established by massive constructive learning mechanisms which are typical for this time phase. Most of them do not reappear later in life and are critical for the acquisition of several sensory and cognitive abilities, including emotions and language.[96] Neurodevelopmental processes are controlled by genetic regulatory programs and shaped through learning in interaction with the environment. This relatively short period of time links ontogeny and phylogeny.

(III a,b) Learning: The final individual neural system is able to learn, that is, to adapt its parameters according to both its current internal state and in interaction with the environment. It is essentially the fine-tuning and modulation of connections and other parameters established during neurodevelopment. The ability to learn has influence on the rate by which genes of a specific individual are passed to the next generation. This nests the phases of individual life and learning (ontogeny) into the tree of evolution (phylogeny). Learning is closely related to memory and involves short-term (IIIa) and long-term (IIIb) processes.[8,56]

Timephases (II) and (III) together can be interpreted as *ontogenetic learning*, that is, learning during the lifetime of an individual in contrast to phylogenetic learning. Phylogenetic and ontogenetic learning can be further distinguished from *sociogenetic learning* where the learned behavior is social culture which is accumulated within a group's lifetime.[4] The above model takes the viewpoint of an individual and distinguishes conceptionally and functionally different phases of learning in the establishment of an individual nervous system. The phases of neurodevelopment (II) and learning (III) operate on the same time scale and they have some mechanisms in common, for example, Hebbian Learning.[37,67] However, we treat them as different phases because there are essential differences in their underlying neurobiological and regulatory mechanisms. In particular, we see phase (II) more associated to the setup of the central nervous organ while phase (III) is more associated with its refinement.

The following sections review some of the most well-known facts about evolution (Sec. 2), neurodevelopment (Sec. 3) and learning (Sec. 4) with the aim of finding a biologically motivated characterization of (machine) incremental learning which will be proposed in Sec. 7. Before that, traditional machine learning concepts related to incremental learning are reviewed in Sec. 5 and an overview of key experimental work is given in Sec. 6. The article concludes with a summary in Sec. 8.

2. Evolution

The longest time phase of the above mentioned three time phases is evolution. The algorithms of evolutionary computation have been successfully applied as optimization algorithms for a large variety of

engineering tasks.[126–128] However, there have also been intentions to employ evolutionary algorithms as abstract models of biological evolution, see for example Refs. 69 and 70. This is reflected in the shared vocabulary of the biological theory of evolution or genetics and the terminology of evolutionary computation.[38,52] Atmar[4] does not agree with the widespread opposition to accepting simple evolutionary optimization algorithms as an interpretation of Darwinian[35] evolution. He claims this opposition is caused by a misinterpretation of evolution theory. Instead of focusing on the complexity of isolated structures built by evolutionary trial-and-error, such as genes, neurons or processes such as recombination or the representation of individuals, it would be more appropriate to emphasize those processes that optimize and evaluate the whole of the evolving structure. According to Atmar[4] the essential mechanisms of evolution can be stated in rather simple terms. To formalize the process of evolution of a population within a single generation he proposes a sequence of four mappings between a genotypic coding space G together with an input alphabet of environmental symbols I and a phenotypic behavioral space P:

$$I \times G \xrightarrow{e} P \xrightarrow{s} P \xrightarrow{r} G \xrightarrow{m} G\,.$$

The first map, epigenesis (e), is many-to-one and translates each genotype into a phenotype. The processes of selection within the local population are described by the map s. Mapping r describes genotypic representation within the population prior to reproduction. Mutation (m) includes random and directed coding alterations including repair and recombination. The process of evolution progresses through indefinite repetition of these four maps.[4,52]

2.1. *Evolution of nervous systems*

On the time scale of evolution all the chemical and basic hardware, as well as the overall concepts of the central nervous system, evolved over a long period of time. Evolution theory assumes that life on earth began several thousand million years ago and the cerebral cortex of mammals evolved from the primordial cortex of amphibians and reptiles about 300 thousand million years ago.[11,97] According to Roth and Wulliman[122] most of the basic elements of the central nervous system, such as neurotransmitters, neuropeptides, ion-channels, receptors and transport molecules are evolutionarily older than neurons and nervous systems. Neurons emerged to enhance the functionality of these previous substances and structures, by binding them functionally together. At a later stage central nervous special organs (brains) evolved by applying the same principle to neurons and glial cells. Two groups of nervous systems evolved, one with and the other without central nervous special organ. The class of vertebrate nervous systems is one of four subclasses of the class of nervous systems with a brain. In vertebrate nervous systems, the brain rostrally directly connects to a dorsal nervous tube.[122] The brain consists of a large number of organizational modules. Tooby and Cosmides[139] claim these evolved because they were appropriate for certain tasks essential for the survival of our ancestors. Primates have, compared with other mammals, relatively small limbic structures but a relatively larger isocortex.[30,50] This can be explained by the following interaction between the mechanisms of evolution and neurodevelopment which determines relative size of brain modules:

> " ... if a species gains extra cycles of neurogenesis across the course of evolution, the greatest relative enlargement occurs in the parts of the brain that develop relatively late ... "[10,32,51]

2.2. *Artificial models of the evolution of nervous systems*

One might propose that the evolution of nervous systems can be roughly modeled by hybrid systems which combine techniques from evolutionary computation with models from computational neuroscience such as artificial neural networks. To evaluate this type of proposal we briefly review some of the key studies and systems available in that area.

Several reviews show that evolutionary artificial neural networks have been investigated and applied successfully in recent years.[104,125,142,146] According to Yao[146] evolution in artificial neural networks can include adaption of connection weights (an alternative to standard network training, see Ref. 92 and 121), evolution of architectures (topology and activation functions) and evolution of learning rules (adaption of learning parameters). The evolution can take place on all three levels simultaneously. For example,

the evolution of the learning rule can interact with the evolution of the architecture.

An early example of combining artificial neural networks and evolutionary algorithms is the study "Designing Neural Networks using Genetic Algorithms" by Miller.[93] Neural networks were represented as connection matrices and every network (phenotype), corresponding to one of the connection matrices, was trained on the xor–task to produce a fitness determined by the error on the test set. Based on this fitness the population of matrices (genotypes) was then modified using the operators of a genetic algorithm.

EPNet and *ENZO* are evolutionary systems for evolving the topology and the connection weights of artificial neural networks. *EPNet* was proposed and described by Refs. 147 and 148 and information about different versions of *ENZO* can be found in Refs. 17–19. The name *EPNet* stems from the technique *E*volutionary *P*rogramming which is an evolutionary algorithm without crossover operator.[53] *ENZO* is an acronym for the German name *E*volutionärer *N*et*z*werk *O*ptimierer. Both systems use a form of *Lamarckian learning* where the trained connection weights are passed to the new generation where they are employed as initial weights. Braun *et al.*[18] report that Lamarckian learning could improve training times by 1–2 orders of magnitude and in some cases the Lamarckian approach was even necessary to solve the learning task. In *EPNet* networks are trained by a hybrid method including modified backpropagation learning; that is, standard backpropagation with adaptive stepsize and simulated annealing. *ENZO* employs resilient backpropagation learning (*RProp*).[118] Comparisons of *EPNet* and *ENZO* and larger studies using these systems are not known and would be difficult to obtain due to very long training times which rapidly increase with growing network size.

An alternative approach to model the interaction of evolution and learning is the study of Batali.[9] He focuses on training recurrent networks and evolves the set of initial weights. The learning ability of recurrent nets is typically very sensitive to the selection of their initial weights. Batali argues that for successful training it is necessary to start with initial weights which have an "innate bias" that gives them enough flexibility and the ability to learn. He draws parallels with the learning ability of biological neural networks during their sensitive periods and points out that the initial learning ability of recurrent nets can degrade after a longer period of training. In Batali's approach the initial weights are reset to their initial values at the end of training before they are passed to the new generation. Therefore, his approach is different from Lamarckian learning as it was employed for reasons of efficiency by *EPNet* and *ENZO*.

From a practical point of view, attempts to obtain a large general purpose evolutionary learning system, with systems such as *EPNet* or *ENZO*, have led to a dead end, due to training times which are too long. Modular systems could provide a solution, given that separately trained network modules already exist, and only a smaller number of connections remain to be evolved.

It is well-known that modularity is a common concept in biological nervous organs too. For example, Brodmann[20] found that the isocortex is divided into cytoarchitectonic regions, the now so called *Brodmann areas*, which are related to specialized cortical functions. Mountcastle[94] proposed that distributed activity in selected sets of brain modules implements higher cognitive function. Comparisons across species show that, for reasons of efficiency, biological brains become more modular, with increasing size of the isocortex.[75]

There are at least two approaches to modular systems in artificial neural networks: modular neural networks and neural ensemble networks. *Modular neural networks*[5,130] consist of autonomous modules which are smaller neural networks with specialist abilities. The artificial modules exchange information only by way of their inputs and outputs. The concept of modular neural networks is more general than that of *neural network ensembles* which are sets of neural networks that are all trained on the same task, either in sequence or simultaneously,[87] while a suitable combination of the outputs of the networks is taken to obtain better results than by taking the output of one network alone.[130]

2.3. *Discussion*

The above review on evolution or phylogenetic learning in biology and machine learning has raised questions about the emergence of modular structure

and relative module sizes. This is of relevance for incremental learning in large systems. In cognitive science and evolution biology the concept of general purpose learning systems for modeling the computations of the brain in under discussion. For example, Gallistel[58] proposes to replace the idea of a general-purpose learning system with adaptively specialized learning modules. The modules would only have some elementary computational processes in common. These would be required for manipulating neural signals in accord with the laws of arithmetic and logic, and for storing and retrieving the values of variables. Gould and Marler[64] express the view that these modules have a task specific structure that we see in instinctive behaviors. To understand and detect the modules in the human learning system Tooby and Cosmides[139] suggest that investigators reverse engineer the brain to find why the brain was built and which specific families of computations the brain evolved to accomplish. This approach is based on knowledge of evolutionary biology and takes into account ancestral activities, selection pressures and the environments where our ancestors evolved. On the contrary Finlay *et al.*[51] and Clancy *et al.*[31,32] argue in favor of the concept of general purpose-learning and suggest that "structure leads function":

> " ... the form of these sensory, motor and cognitive systems are the result of competitive recruitment of processing resources from a super-abundant pool of cortical neurons made available more or less at the same time. ... "[51]

3. Neurodevelopmental Phases

Rakic[111] proposes that understanding the principles and mechanisms controlling the production of cells destined for the cerebral cortex may be the key to understanding human intelligence. The developmental time phase, where large constructive processes establish the structure of the central nervous system, consists of a prenatal part (IIa) and a postnatal part (IIb). Several incremental learning processes occur in parallel and these are still not completely understood. Some of the processes are genetically predetermined[140]; however, interaction with the environment is a crucial part of the developmental phase. This must therefore be regarded as a phase of massive learning. Connections are established during the developmental process via different forms of axon guidance (cf. Chap. 4 of Price and Willshaw[106] or Chap. 9 of Brown *et al.*[22]) and other more local activity dependent mechanisms which refine the precision of the connection.[2,106] Connections in this context include projections, which connect large regions or modules of one brain region to another through thousands of fibers (e.g. thalamocortical information processing[25,39,90]).

Some of the information the organism receives in the developmental phases, for example, via mother-child interaction, is essential for its later life. If certain stimuli and information are not properly received during these *sensitive periods*, the organism will lack important abilities and may not be able to survive. The sensitive periods are unique in that they do not reappear in later stages of life. If certain aspects of vision[71,76,143] and language[73,85,100,102] are not acquired during the sensitive periods, they cannot be learned anymore at later stages of life. The collection of Birdsong[12] contains more details on the sensitive periods' influence on first and second language acquisition. Sensitive periods are not restricted to humans but have been similarly observed in animal studies, see for example Refs. 13, 81 and 88.

We propose one of the keys for understanding biological incremental learning, as it appears during the sensitive periods, is to analyze the time structure of the main mechanisms of the neurodevelopmental set-up process. For that purpose we collected a number of studies that shed light on the timeline of cortical neurogenesis. As an outcome, we summarize the qualitative time structure of the development of different neuron components in a roughly interpolated timeline in Fig. 1.

Most of the results and experimental data that were used for the present review were reported by Rakic and colleagues from experimental studies with rhesus monkeys (e.g. Refs. 110 and 111). Similar results were predicted by the regression model that was employed by Clancy, Darlington, Finlay and colleagues.[30,34,50] In their study about 40% of the data was backed up by experimental results. Another review which summarized neurodevelopmental concepts was given by Quinlan.[108] Rakic[110] reports that the prenatal genesis of cortical neurons in primates lasts approximately 60 days starting from the fortieth embryonic

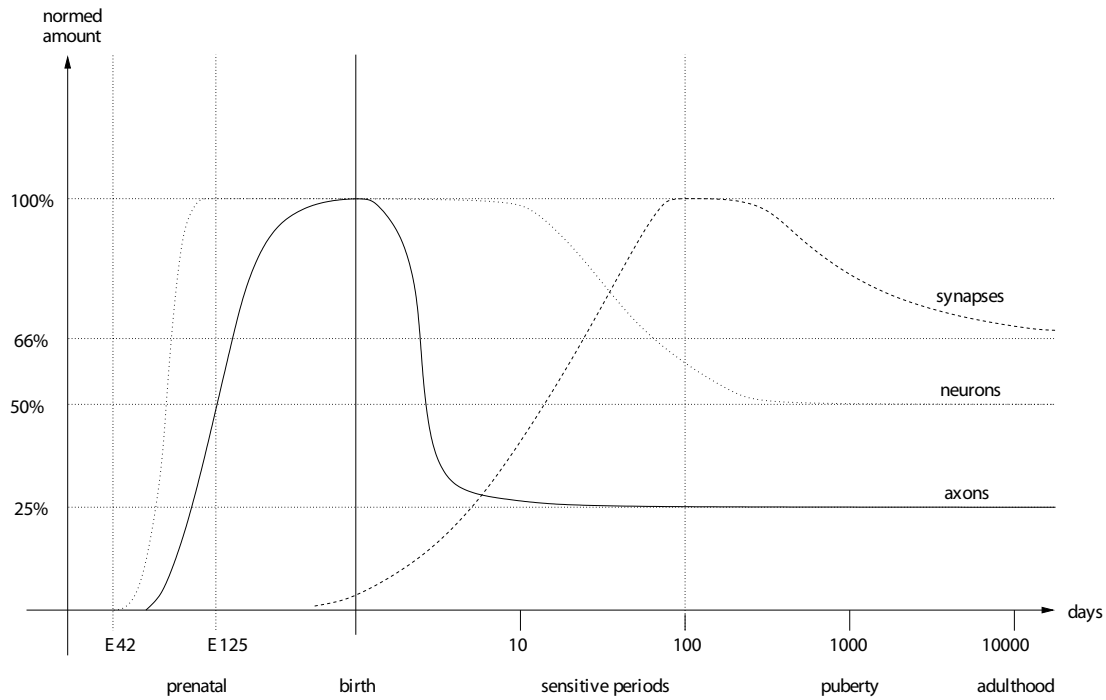


Fig. 1. We have derived a hypothetical timeline of quantitative neuron (dotted line), axon (continuous line) and synapse (dashed line) development in the cortex. The experimental data was reported by Rakic[110] in the case of rhesus monkeys. The number of neurons and axons grows fast, while the number of synapses slowly starts to increase later. The amount of synapses continues to increase even when the number of axons and neurons is already decreasing during the "sensitive periods".

day (E40). Primates acquire all their neurons within that period before birth and almost no isocortical neuron is generated at any time later.[109] Neurogenesis involves strategies of nerve cell differentiation. These depend on the regularization of transcription proteins by signals within the cell itself (control by cell lineage) or by signals received from neighboring cells (control by induction); see Ref. 77. Neurons migrate after their birth at the ventricular side of the cortex, along scaffolds of glial cells, radially outwards into one of the six cortex layers (cf. Table 1).

Axons often grow along a complicated path where they meet other neurons which they ignore, connecting only to the target neuron.[23] They form synapses with up to 1000 other neurons and are finally precisely connected with specific neurons or parts of them. The period of fast increase of the number of axons follows shortly after the period of growth of neurons and lasts until birth. During the course of development, and in particular at birth, most brain structures in higher vertebrates possess many more neurons and axons than in adulthood. For example the newborn monkey has about 200×10^6 calossal axons and the adult has only 50×10^6. Soon after birth a period of fast axon loss begins. The loss rate within the first three weeks is extremely high (50 axons per second) and then slows down (to 5 axons per second) until puberty.[83,84] Axon loss occurs after topographical and columnar organization and when precise connections have been achieved. It is involved mainly in synaptic reorganization on a local rather than global level.

After a period of fast axon loss, a period of programmed cell death of neurons (*apoptosis*) starts.[99,101] It causes a significant decrease in the number of neurons over a short period. Depending on the brain area, there are about 35%–60% more neurons in the fetal monkey than in the adult monkey; that is, nearly half of all neurons will die.[110] Apoptose of neurons is a process of genetically programmed cell death. For example, in different individuals of the nematode worm *Caenorhabditis elegans*, the same cells always die. Some of them die even before they start neural processes.[134]

Table 1. The six cortex layers counted from surface I to inside VI. Each layer has its characteristic type of neurons and fibers.[77]

Layer	Cell types	Functionality and connectivity
I	Apical dendrites of the pyramidal cells and axons of stellar cells, only a few cells	Local connections of neighboring neurons
II	Small pyramidal neurons	Long-distance cortico-cortical connections
III	Small/medium pyramidal neurons and interneurons	Short-distance cortico-cortical connections
IVa	*Small stellar neurons*, small pyramidal neurons	Cortical interneurons (inhibitory and excitatory), *thalamo-cortical afferences (vertical entry)*
IVb	*Stellar neurons*	Cortical interneurons
V	*Large pyramidal neurons*	Projection to subthalamic structures (basal ganglia, brain stem), short and long-distance cortico-cortical fibres
VI	Small (fusiform) pyramidal neurons	*Cortico-thalamic efferences*

Brown *et al.*[22] give a general overview about how developing neurons can die. This overview includes target-dependent neuronal death, death of newly generated neurons, developmentally programmed neuronal death and death removing whole populations of neurons.

During development the periods of axon loss and apoptose coincide with a period of rapid synaptic production, rather than synaptic elimination. Synaptic density in monkey brains increases quickly during the first 2–3 months of life, until it reaches a 30%–40% higher level than in the adult; see Refs. 15, 110 and 112. A "plateau" phase of high synaptic density lasts for about two years.[15,16] It is followed by a 2–3.5 year period of loss of approximately 2,500–5,000 synapses per second.[14] Synaptogenesis with initial overproduction and subsequent elimination appears synchronously for synapses and neurotransmitter receptors.[86] It occurs following a well determined schedule in several brain areas and is correlated to the emergence of specific cognitive functions.[30,72]

Connections developed prenatally are refined postnatally and to a lesser degree already in the prenatal phase, by activity-dependent mechanisms which work in response to sensory input.[78,106] At the intersection of time phase (IIb) and (IIIa) two different processes overlap and occur in parallel: *competitive synaptic elimination* and *learning*. Both involve the same principle which is related to Hebb's hypothesis.[a] We have a quantitative and a qualitative change of synapses in the neural network. It involves neurotransmitters, second messengers and differential gene expression, that is, an interaction of short-term, long-term and very long-term processes. More details about these processes can be found for example in the book of Kandel *et al.*[77]

Apart from the soma, the axon and the synapses, dendrites are another important part of the biological neuron.[91] Quartz and Sejnowski[107] pointed out that dendrites play a more active role than traditionally believed; these are dynamic structures whose growth depends on many factors. A neuron with active dendritic segments can perform tasks similar to artificial hidden layer networks. Dendrites grow slowly and act on a longer time scale than axons. Since dendrites are the cell parts which have most of the synaptic contacts with other neurons, synaptic changes are closely related to dendritic changes. It has been proposed that dendrite morphology is also responsible for long-term memory.[145] Therefore, in our model, we count dendritic changes as synaptic changes, rather than as axonal changes.

If we summarize the above reviewed data and roughly interpolate it, we obtain a hypothetical, schematic timeline for neurogenesis, apoptose of neurons, axon growth, axon loss, synaptogenesis and synapse elimination. The result is depicted in Fig. 1 (cf. Ref. 27). The section of the timeline between birth and puberty consists of a speculation

[a]Hebb's hypothesis[67] postulates that synapses between neurons are strengthened if pre- and postsynaptic activation occurs synchronously. In artificial neural networks often a modification of Hebb's original rule is realized: Synapses between synchronously active neurons are strengthened while other synapses are weakened. Hebbian learning in the brain is realized by *NMDA* receptors, for example in the hippocampus, cerebellum and visual system.[37]

which, to our current knowledge, is not experimentally confirmed. The time scale of the timeline covers a 165 day period of gestation, followed by a postnatal period of about three years to sexual maturation. This corresponds to data observed in the case of rhesus monkeys (cf. Ref. 110 and 111). There are quantitative changes between different brain regions and across species. However, recent statistical analyses in the studies of Clancy, Finlay and colleagues[30–51] indicate that the qualitative picture is virtually invariant and can almost certainly be extrapolated to humans. For humans, the picture would therefore be qualitatively identical, but stretched, due to a 280 day period of gestation and a much longer period until sexual maturation.

The review of the relative time structure of the development of neuron components identifies a "biological incremental learning algorithm" which works as follows:

(i) Fast growth of the number of neurons.
(ii) Fast growth of the number of axons.
(iii) Slow synaptogenesis which continues.
(iv) Very fast axon loss.
(v) Slow apoptose of neurons.
(vi) Slow synapse elimination.

The essential principles of this algorithm are *timing* and *initial overproduction and subsequent elimination.* It is partially determined by genetic regulatory programs and partially by activity dependent interaction with the environment.

4. Learning

Learning during the lifetime of an individual is generally referred to as *ontogenetic learning.* From the viewpoint of cellular and molecular neurobiology learning is essentially the fine-tuning of developed synapses or dendrites through experience.[22,77,145] It begins approximately at the end of the first stage of neurodevelopment (i.e., stage IIa) after initial synapse formation, controlled by genetic regulatory and early developmental processes, has been completed. Several processes are involved: for example, long-term potentiation, long-term depression, and changes to the synapse as associated with the *NMDA* receptor. The outcomes of learning are different kinds of short-term or long-term memory which can be either implicit and unconscious for motor skills or explicit and conscious for remembering things. In humans, three stages of memory can be distinguished[7,132,145]:

- *Sensory memory*: Information from multimodal sensory fiber collaterals is integrated by neurons in the brainstem and triggers acetylcholine to be released into specific regions of the cerebral cortex and hippocampus. This control mechanism of ongoing sensory processing and the attention level are updated every 150–700 ms.
- *Short-term memory*: Prolonged acetylcholine elevation at specific cortical sites for stimuli that have been allocated a greater degree of attention.
- *Long-term memory*: A variety of long-term changes at synapses and dendrites triggered by recurring intervals of enhanced acetylcholine release.

As we have already pointed out, "learning" and its connection to memory constitute a large topic which is treated by many disciplines and from many different perspectives.

In machine learning, a traditional distinction between three main classes of learning paradigms is made:

- *Unsupervised learning*: Learning without external signal or control.
- *Reinforcement learning*: Only a scalar reinforcement signal is provided.
- *Supervised learning:* A target output or teacher signal must be specified.

In all three cases, given an input vector $x(t) \in \mathbf{R}^n$ and an output vector $y(t) \in \mathbf{R}^m$ at time step $t \in \mathbf{N}$, a mapping $F : \mathbf{R}^n \supset X \to Y \subset \mathbf{R}^m$, $x \mapsto y$ is learned. In many situations the mapping F is identified with a parameter array, such as a neural network weight matrix W.

Unsupervised learning is a common learning principle of the cortex and is related to associative memory. In machine learning it is realized for example by *Kohonen networks* which are also called *self-organizing maps.*[82] Self-organizing maps learn a topographic map of the input patterns where the location of the neurons, given by their coordinates in a lattice, indicates statistical features of the input map.[65] Learning employs a Hebbian-type rule. Many other examples of unsupervised learning paradigms have been developed.[68]

In the supervised paradigm a parameter update is based on an error, given by the difference between the actual output $y(t) = F(x(t))$ and the target output $y'(t)$. The target output is sometimes also called teacher signal. A typical supervised learning algorithm for neural networks is backpropagation, which is described for example in Ref. 65.

It could be argued that reinforcement learning is the most general concept which includes both unsupervised and supervised learning as extreme forms. Using this interpretation, in unsupervised learning a zero reinforcement learning signal is provided constantly and in supervised learning the error information could be interpreted as a form of "rich" reinforcement learning signal.

K. Doya[42,43] proposed four correspondences between the effects of four parameters of a temporal difference algorithm, and four neuromodulatory systems of the brain. Each of them involved one of four neurotransmitters: dopamine, serotonin, acetylcholine, or noradrenaline. They project from the brainstem to the cortex, the cerebellum and to the basal ganglia.[79] The mechanisms associated with the basal ganglia were proposed to perform a type of reinforcement learning[41]; while the mechanisms of the cerebellum could be described in terms of supervised learning; and the calculations of the cortex could be modeled by an unsupervised paradigm.

5. Machine Incremental Learning

In machine learning literature, the term *incremental learning* is used inconsistently for several different forms of learning or training. Before we propose a characterization of incremental learning in Sec. 7, we will discuss a selection of commonly used classification schemes for supervised learning in artificial neural networks. Some of them have been discussed in Refs. 65, 123. These classification schemes apply to many machine learning algorithms including unsupervised and reinforcement learning for artificial neural networks.

5.1. *Structure modifying learning methods*

Constructive and pruning methods can be motivated by theoretical and experimental results, which show that training speed and generalization performance is affected by the size and architecture of the neural network; see, for example, Refs. 1, 62 and 141. Because pruning modifies the network structure or topology, it is closely related to constructive learning. This conforms to the biological perspective, that growing mechanisms are strongly related to pruning mechanisms. *Constructivism* and *selectionism* — two concepts from cognitive neuroscience reflecting the ideas of constructive and pruning methods — are discussed by Quartz and Sejnowski.[107]

During *constructive learning*, single neurons and links, or whole layers or subnetworks, can be added or deleted from the network architecture. Most well-known are probably the algorithms which incrementally add hidden units to a neural network, such as Cascade-Correlation[48] or Tower[57] algorithms. A large amount of work has been done to investigate these type of algorithms, see for example Refs. 95 and 105. Hayward[66] introduced the GenTower algorithm which, inspired by the Tower algorithm, adds small subnetworks to the neural network during training.

Another group of algorithms which start with a small network and successively add new units has been introduced by Platt.[103] Building on the same idea are Fritzke's[55] Growing Cell structures, where units are added by evaluating local statistical measures gathered during previous adaption steps and the network dimensions are preserved. In Growing Neural Gas[54] the network topology is generated incrementally by competitive Hebbian learning.[89] The dimensionality depends on the input data and can be locally different. All these methods employ radial basis functions. Growing Cell Structures are good for unsupervised learning and neural gas is good for supervised learning, where it leads to small networks with strong generalization ability.

Constructive methods are generally regarded as being more powerful and sophisticated than fixed structure training methods. Most of the studies on constructive methods emphasize the engineering aspect of building a more or less complicated system, which somehow optimizes the network structure, learning parameters and hopefully generalization or other performance aspects, see for example Ref. 131.

Pruning is the process where links or units are removed from the network during training. These techniques are realized in various algorithms such as Optimal Brain Surgeon, Optimal Brain Damage and others; see Ref. 116 for an overview.

Weight decay can be regarded as a selectionist method, like pruning. However, instead of removing connections, it first restricts the growth of their weights by giving them a tendency to decay to zero; that is, the connection would disappear unless reinforced.[68]

Selective Learning with Flexible Neural Architectures (*SELF*) was introduced by Ref. 150. It works on both the data and the network structure. Starting with a small training set and a small feed forward network, the training set is increased incrementally after training. If training does not converge, hidden units are then added to the network to increase its capacity.

5.2. *Adaptive or parameter learning*

Adjustments of parameters of the learning rule, such as the learning rate during training, are called *adaptive learning* or *parameter learning*, and can be regarded as one form of incremental learning. An advantage of adaptive techniques is their robustness with respect to the choice of the initial parameters. According to Riedmiller,[117] adaptive methods can be divided into two categories, local and global adaptive learning. *Local adaptive techniques* rely only on local information, such as changes of a single weight, and reflect the principle of parallel processing, which is a characteristic of neural network learning. Among typical examples are *Quickprop*[49] or *RProp*.[118] *Global adaptive techniques*, such as the conjugate gradients method, use information about the state of the entire neural network, for example, the direction of the overall weight-update vector.[117] The conjugate gradients method requires more computation, but it converges faster when compared with standard backpropagation learning.

5.3. *Variations of data presentation*

In *pattern learning* weights are updated after each presentation of a single training pattern. In *epoch learning* weights are updated after presentation of the whole training set. For large training sets with a lot of redundant information, pattern learning is the preferred method. The term "pattern learning" is often used synonymously with "incremental learning" and "epoch learning" is often called *batch learning*. Here, the *batch* is the training set. It is possible to update weights after presentation of parts of an epoch. This is called *mini-batch learning*. By varying the size of the mini-batch during training the method can be altered from pattern learning to epoch learning, or vice versa. Many algorithms are applied as batch learning methods, because the accumulated error information at the end of an epoch is a more reliable basis from which to decide about the next step, than using only the error obtained from a single pattern evaluation. Evolutionary hill climbing, conjugate gradients or *RProp*[118] are typically used as batch techniques.

In *on-line learning* new data is generated constantly and each pattern is discharged after presentation to the network. In contrast *off-line learning* uses a fixed training set. The same data is reused and presented repeatedly to the network. While off-line learning can use the information of all the training data to tune its training strategy, on-line learning can encounter unexpected surprises arising in the stream of new data. Therefore off-line learning is, in most situations, more reliable than its on-line counterpart.[123] On-line learning is often used synonymously with incremental learning and pattern learning. But this can cause confusion, because pattern learning does not have to be on-line. Pattern learning can reuse data which has already been seen, whereas on-line learning does not.[123] Batch or epoch learning is typically off-line because it reuses the same data several times. But an epoch learning method, which uses each batch a number of times, until it replaces the batch with new data, could be regarded as an "on-line epoch learning method".

In *staged learning*, training examples are classified in a series of classes graded by degrees of difficulty. An example classification could have three classes: easy, moderately difficult and difficult. The staged training process is conducted in stages corresponding to the difficulty classes. We regard staged learning as a form of incremental learning. A similar approach is known within the framework of reinforcement learning[135] where it is called *shaping*. Shaping has many different aspects. For example, in Refs. 113, 114, 129, 136 and 137 shaping was applied by first solving a sequence of physically simpler reinforcement learning tasks before finally the real task was approached.

6. Experimental Observations

While the structure modifying type of incremental learning discussed in Sec. 5.1 has several theoretical

and experimental results available (e.g. Refs. 1, 62, 108 and 141) the essence of data incremental methods is still not well enough understood. Sometimes data incremental learning is inherent to a structure incremental approach, for example if the input layer or internal processing capacity of a neural network is initially so far restricted that part of the data is filtered out. In situations like this, structure incremental effects might overshadow data incremental effects or vice versa.

For a long period the proposal of Newport[63,96] and the experiments of Elman[47] which suggested unrestricted advantages of a data incremental approach were generally accepted. However, recent studies[119,120,149] seem to question the superiority of data incremental learning.

Elman[46] used examples from natural language to train simple recurrent neural nets on a "one step, look ahead" prediction task. The aim was to learn to predict the order of words in sentences. In his later paper[47] Elman used a similar type of task and investigated two versions of incremental learning which he entitled: *incremental input* and *incremental memory.*

In the *incremental input* approach, Elman trained simple recurrent networks, while the complexity of the sentences in the training data was gradually increased. The training was conducted in five phases of increasing complexity, where each training phase used a different training set of 10,000 sentences. The sentences in the training set of each phase were divided into two classes according to two levels of complexity — "simple" and "complex":

- Phase 1: 10,000 simple sentences.
- Phase 2: 7,500 simple and 2,500 complex sentences.
- Phase 3: 5,000 simple and 5,000 complex sentences.
- Phase 4: 2,500 simple and 7,500 complex sentences.
- Phase 5: 10,000 complex sentences.

The striking result was: *The ANNs trained using incremental input learned the grammar represented by the 50,000 sentences better than those which were trained non-incrementally.*

In the *incremental memory* approach the full data set was used, but the time window of the simple recurrent network was initially restricted and then increased in five steps during training. This was achieved through elimination of the recurrent feedback connections, by resetting the corresponding context units after a number of inputs. This number, which corresponds to the time window, was then increased in stages. The network could only recognise that structure in the data whose complexity corresponded to the size of the time window. Variations of the incremental memory approach were investigated in Ref. 74.

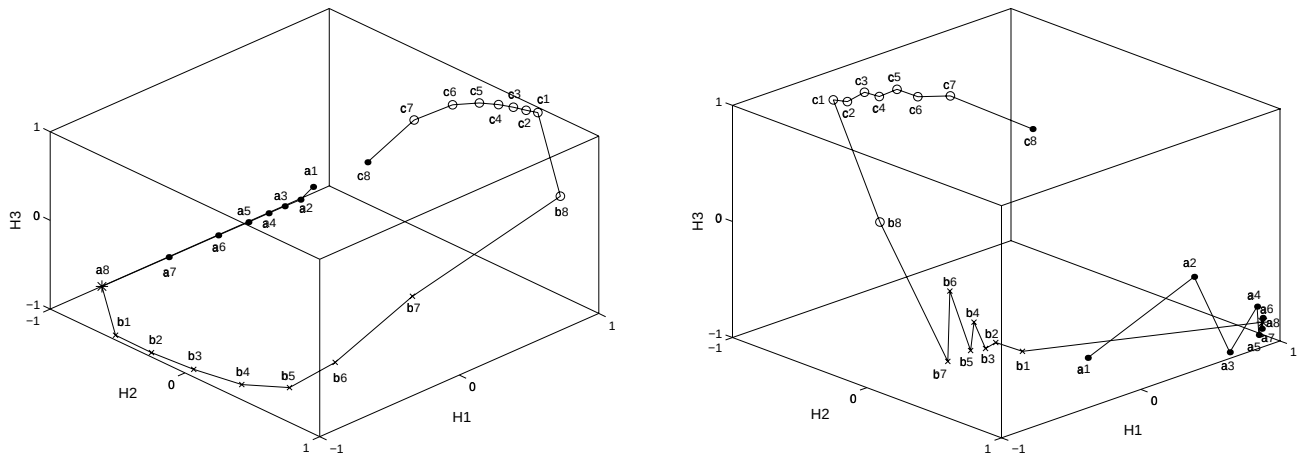
Elman[47] claimed that when the network was trained without using either of the two incremental learning techniques it was unable to learn the task. This observation seemed to be in accordance with the work of Newport[96] who proposed that "less is more"; that is, incremental learning is better or more powerful than non-incremental learning.

In contrast to Elman's work[47] later studies by Rohde and Plaut[119,120] report experiments using a similar language task that data incremental learning is not necessary. The networks learn the task when trained straight on the most complex data sets. It was claimed that recurrent networks statistically learn the task, and inherently extract simple regularities, before proceeding to the more complex structures. Therefore, an incremental learning scheme would work both automatically and implicitly and must not artificially be imposed on the network by additionally preprocessing the data or restricting the network's structure. Training which first uses simplified data would allow the network to learn an inappropriate data representation. The network would later experience difficulties in adjusting to more complex data.

Chalup and Blair[26,28,29] describe staged learning of a context-sensitive language with first order recurrent neural networks. Strings from the $\{a^n b^n c^n; n \geq 1\}$ language can be staged to form the basis for a data incremental learning approach:

$$... \rightarrow a^3b^3c^3 \rightarrow a^4b^4c^4 \rightarrow a^5b^5c^5 \rightarrow ...$$

The 3-dimensional graphs in Fig. 2 show the activation of the three hidden units (H1–H3) of two representative solution networks while processing the string $a^8b^8c^8$. The state trajectory for $a^8b^8c^8$ has 24 states, each of them corresponding to an input symbol ('a1–a8', 'b1–b8', 'c1–c8') and a predicted symbol ($a =$ '$\bullet$', $b =$ '$\times$', $c =$ '$\circ$' and undetermined = '$*$').



(a) (b)

Fig. 2. The trajectories of hidden unit activation of two solution networks with hidden units H1, H2 and H3 while processing the string $a^8b^8c^8$. One of them was trained incrementally (a) and shows an almost monotonic behavior. The other network was trained non-incrementally (b) and shows an oscillating behavior. The symbols '•', '×' and '○' correspond to the *predicted* symbols a, b and c, respectively. The first b of the string cannot be predicted which is indicated by the asterisk '∗'. In addition to the predicted symbols, each state has been assigned its corresponding *input* symbol 'a1–a8', 'b1–b8' and 'c1–c8'.

The states are connected by lines to show qualitatively how the trajectory proceeds through the symbol clusters. In the left graph the hidden unit dynamics of a network is displayed which was incrementally trained and in the graph on the right side it was non-incrementally trained. In the majority of our experiments the hidden unit trajectories of the incrementally trained networks were almost monotonic while the hidden units' activation trajectories of the non-incrementally trained networks were oscillating, as can be seen in Fig. 2. Further, the experiments with incremental learning produced more and earlier solutions than the ones with non-incremental learning. This result indicates that the advantage in efficiency of incremental learning can come with the cost of obtaining a qualitatively different solution from non-incremental learning. However, it must be taken into account that training recurrent neural networks is extremely sensitive to the initial conditions (cf. Batali[9]). Training results with other network types might be more stable with respect to switching between incremental and non-incremental learning.

The result of Chalup and Blair[28] can be used to mediate between Elman's claim[47] that incremental learning is necessary for certain tasks and Rohde and Plaut's protest[119,120] that explicit incremental learning is not necessary and sometimes a hindrance. First, it should be noted that Chalup and Blair were using a context-sensitive language for training, which is more difficult than the languages used in the studies of Elman or Rohde and Plaut. Chalup and Blair's results show that both incremental and non-incremental learning were successful, which confirms that incremental learning was not necessary, exactly as argued by Rohde and Plaut. However, Chalup and Blair also showed that incremental learning was more efficient than non-incremental learning which counts for Elman, because this difference in efficiency might under certain constraints, for example a restriction in training time, inhibit successful non-incremental learning while incremental learning can still be successful.

Rohde and Plaut's proposal[119,120] that incremental learning is happening implicitly and does not need to be imposed to the learning system from outside contrasts with the view of Kirby and Hurford[80] who associate the term "incremental learning" with

> " ... the idea that some learning-related resource starting at a low value, which then gradually increases while (but not necessarily because) the organism matures ... " and " ... the initial low (immature) value of

> resource actually facilitates, or even enables, the early stages of learning. Later stages of learning are in turn facilitated, or enabled, by higher-valued settings of the resource concerned ... "[80]

Kirby and Hurford review Elman's approach[47] and propose that the elements of timing and evolution are missing from his models. They claim that " ... any innate pre-programming for input-sensitive growth must have evolved ... "[80] and they propose a method to incorporate both "less is more" and "starting small"[47] ideas into an evolutionary framework to model the evolution of incremental learning.

7. Characterization of Incremental Learning

One possibility would be to generalize and claim that all learning is somehow incremental. Another popular use of the term "incremental learning" is to use it synonymously with on-line, pattern or adaptive learning. However, as we have discussed above in Sec. 5 each of these three terms stands for a different learning concept, and therefore it is better not to use all of them synonymously with incremental learning. Our characterization of machine incremental learning is primarily inspired by the parameterized increment of learning resources as was suggested and employed in the approaches taken by Elman[47] and Kirby and Hurford.[80] It also aims to take into account the different concepts of incremental learning that our review of biological and machine learning mechanisms revealed. The proposed characterization is based on two key observations:

- The analysis of sensitive period mechanisms has shown that timing is an essential feature and that the constructive incremental mechanisms in nature involve growing *and* pruning. Biological incremental learning in the developing brain exhibits a complex interaction of both processes (compare the development of neurons, axons and synapses in the first 100 days after birth, as displayed in Fig. 1).
- Biological learning involves interacting learning mechanisms on different time scales.

The following characterization of machine incremental learning incorporates the concepts behind these observations and applies them to three dimensions of machine incremental learning: structural changes, learning parameter adjustments and input data variations.

Characterization: Let $s = (s_i)_{i=0\cdots l}$, $l = (l_j)_{j=0\cdots m}$ and $d = (d_k)_{k=0\cdots n}$ be families of real numbers. An *Incremental Learning System* $I = (s, l, d)$ is a learning system[b] which is parameterized by three families of *incremental learning parameters* which can be modified during training:

(i) *Structure parameters* $s = (s_i)_{i=0\cdots l}$ which are, for example, the number of neurons, density of connections, or other parameters which determine structure and functionality of a neural network.
(ii) *Learning parameters* $l = (l_j)_{j=0\cdots m}$ which are, for example, evolutionary or other learning parameters, such as the stepsize.
(iii) *Data-complexity parameters* $d = (d_k)_{k=0\cdots n}$ which can represent any complexity measure of the training data.

Accordingly, there are three main forms of *incremental learning.* Each of them modifies members of one of the parameter families defined above:

(i) *Structure Incremental Learning*: The structure or functional capacity of the neural network is changed during learning.
(ii) *Learning Parameter Incremental Learning*: A selection of learning parameters from $l = (l_j)$ is adapted during learning.
(iii) *Data Incremental Learning*: The data set or its complexity is increased in stages during learning controlled by parameter changes of $d = (d_k)$.

In many cases a mixture of these three forms of incremental learning may occur. An example for an incremental learning system is a neural network, together with a training algorithm and training data, where the complexity of the training data is parameterized by a sequence of parameters $d = (d_k)$. For example, the study of Elman[47] which was discussed in the previous section, has employed a learning system of this category.

[b]For simplicity it may be assumed that a *learning system* is a neural network together with a training algorithm and training data.

Table 2. The *Three Timephases Model*: The left column contains a list of biological mechanisms within the three time phases of evolution (I), neurodevelopment (II) and learning (III). The right column contains a proposed list of candidates for corresponding machine learning techniques.

Timephase	Neurobiology	Machine learning
(I)	Evolution of the brain and its modules	Evolutionary algorithms
(IIa)	Neural development (prenatal)	Growing algorithms genetically programmed, structure incremental learning
(IIb)	Sensitive phases (postnatal), early learning, child language acquisition	Pruning and growing algorithms, Hebbian learning and formation of cell assemblies, all types of incremental learning, self-organizing maps
(IIIa)	Learning relevant for short-term memory, adaptive learning processes which continue after the sensitive phases	ANN learning algorithms, data incremental learning, self-organizing maps
(IIIb)	Learning relevant for long-term memory, neuromodulation	Supervised learning, reinforcement learning, data incremental learning, metalearning and parameter incremental learning

8. Discussion and Summary

The review has shown that biological learning mechanisms are incremental in various ways. The main time phases, as seen from the perspective of an individual, are: evolution, neurodevelopment and (late ontogenetic) learning. They operate on different time scales and can themselves contain learning processes which operate on different time scales. For example, learning involves short-term processes but also long-term processes. In summary, there are two concepts which may be important to explain the general phenomenon of incremental learning:

- *Interaction of time scales*: The incremental learning processes of evolution, neurodevelopment and learning can interact on different time scales.
- *Algorithm of intertwined neurodevelopmental processes*: A review of the sensitive period mechanisms has shown that constructive incremental mechanisms in nature include growing and pruning of different neuron components at different times during development (compare timeline in Fig. 1).

The complex interaction of intertwined biological incremental learning processes on all levels of the three time phases of our model, and in particular during the sensitive periods, is still barely understood. Many of the amazing effects of biological incremental processes might stem from that interaction.

Incremental learning was further discussed as a traditional concept of machine learning, but was also motivated as machine learning concept that possibly "corresponds" to biological incremental learning. A summary is given in Table 2 with hypothetical correspondences between neurobiological and machine learning mechanisms. At this point, we want to emphasize that these correspondences are hypothetical; and it seems to be widely agreed that artificial neural networks and evolutionary algorithms have little or no biological plausibility.[108] However, it was also argued that finding appropriate links or correspondences between biological and machine learning theory is a matter of both interpretation and finding the right level of abstraction (cf. Ref. 4).

Our review on experimental work focused on data incremental learning. Here for a long period, the simulation results of Elman[47] which suggested unrestricted advantages of an incremental approach over a non-incremental approach, were fairly generally accepted. However, other studies[119,120,149] seemed to question the superiority of data incremental learning. To better understand input or data incremental learning and the above mentioned contradicting example studies, we compared them with our own experiments where incremental learning and non-incremental learning produced qualitatively different solutions. It is indicated that on certain difficult learning tasks where non-incremental learning has a low probability of success, incremental learning is more efficient. However, an explanation of why and when incremental learning works depends on several factors which are inherent to the specific learning task. Therefore it cannot be concluded that incremental learning is generally better or more powerful. Incremental learning can be more successful than

non-incremental learning provided the learning task allows a suitable incremental learning scheme and incremental learning of the lower stages does not inhibit access to learning the higher stages by guiding the learning system into fixation or paralysis.

After reviewing incremental learning in biological and machine learning systems we proposed a characterization of incremental learning which is a generalization of the concept employed by Elman[47] and Kirby and Hurford.[80] Future research could extend our review and the existing experimental results and complement them with a theoretical analysis.

Acknowledgments

The author is grateful to all colleagues and reviewers who made useful comments, and in particular to Alan Blair and Frederic Maire. This project was supported by the University of Newcastle RMC ECR grant *Machines Learn via Biologically Motivated Incremental Algorithms*.

References

1. Y. S. Abu-Mostafa 1989, "The Vapnik–Chervonenkis dimension: Information versus complexity in learning," *Neural Computation* **1**, 312–317.
2. A. A. Agmon, L. T. Yang, E. G. Jones and D. K. Dowd 1994, "Topologic precision in the thalamic projection to the neonatal mouse," *Journal of Neuroscience* **13**, 5365–5382.
3. M. A. Arbib 1995, *The Handbook of Brain Theory and Neural Networks* (Cambridge, MA: MIT Press).
4. W. Atmar 1994, "Notes on the simulation of evolution," *IEEE Transactions on Neural Networks* **5**, 130–147.
5. G. Auda and M. Kamel 1999, "Modular neural networks: a survey," *International Journal of Neural Systems* **9**(2), 129–151.
6. T. Bäck, D. B. Fogel, Z. Michalewicz and S. Pidgeon, ed., *Handbook of Evolutionary Computation* (IOP Publishing Ltd. and Oxford University Press).
7. A. D. Baddeley 1992, *Memory Theory and Memory Therapy* (In Wilson and Moffa,[144] 2nd edition), 1–31.
8. C. H. Bailey and E. R. Kandel 1995, *Molecular and Structural Mechanisms Underlying Long-term Memory*, Chap. 2, (In Gazzaniga,[59] 1st edition), 19–36.
9. J. Batali 1994, *Innate Biases and Critical Periods: Combining Evolution and Learning in the Acquisition of Syntax* (In Brooks and Maes[21]), 160–171.
10. E. Bates, B. L. Finlay and B. Clancy 2002, *Early Language Development and Its Neural Correlates* (Vol. 6 of Rapin and Segalowitz[115], 2nd edition).
11. S. A. Bayer and J. Altman 1991, *Neocortical Development* (Raven Press, New York).
12. D. Birdsong, ed. 1999, *Second Language Acquisition and the Critical Period Hypothesis* (Lawrence Erlbaum Associates, Publishers).
13. M. H. Bornstein 1989, "Sensitive periods in development: Structural characteristics and causal interpretations," *Plasticity of Development* **105**, 179–197.
14. J.-P. Bourgeois 1997, "Synaptogenesis, heterochrony and epigensis in the mammalian neocortex," *Acta Paediatrica Supplement* **422**, 27–33.
15. J.-P. Bourgeois, P. S. Goldman-Rakic and P. Rakic 2000, *Formation, Elimination, and Stabilization of Synapses in the Primate Cerebral Cortex*, Chap. 4, (In Gazzaniga,[60] 2nd edition), 45–53.
16. J.-P. Bourgeois and P. Rakic 1993, "Changes of synaptic density in the primary visual cortex of the macaque monkey from fetal to adult stage," *Journal of Neuroscience* **13**, 2801–2820.
17. H. Braun 1997, *Neuronale Netze: Optimierung durch Lernen und Evolution* (Springer–Verlag).
18. H. Braun and T. Ragg 1995, *ENZO Evolution of Neural Networks* (User manual and implementation guide, version 1.0), ftp://i11ftp.ira.uka.de in/pub/neuro/ENZO (The University of Karlsruhe, Institute for Logic, Complexity and Deduction Systems).
19. H. Braun and P. Zagorski 1994, *ENZO-M – A Hybrid Approach for Optimizing Neural Networks by Evolution and Learning* (In Davidor *et al.*[36]), 440–451.
20. K. Brodmann 1909, *Vergleichende Lokalisationslehre der Grosshirnrinde* (Leipzig: Barth).
21. R. Brooks and P. Maes, ed. 1994, *Proceedings of the Fourth Artificial Life Workshop, Cambridge, MA* (MIT Press).
22. M. Brown, R. Keynes and A. Lumsden 2001, *The Developing Brain* (Oxford University Press).
23. J. A. Campos-Ortega 1996, *Ontogenie des Nervensystems und der Sinnesorgane* (In Dudel *et al.*[44]).
24. S. Carey and R. Gelman, ed. 1987, *The Epigenesis of Mind: Essays on Biology and Cognition* (Lawrence Erlbaum Associates).
25. S. M. Catalano, R. T. Robertson and H. P. Killackey 1996, "Individual axon morphology and thalamocortical topography in developing rat somatosensory cortex," *Journal of Comparative Neurology* **366**, 336–353.
26. S. Chalup and A. D. Blair 1999, *Hill Climbing in Recurrent Neural Networks for Learning the $a^n b^n c^n$ Language* Vol. 2 of Gedeon *et al.*,[61] 508–513.
27. S. K. Chalup 2001, "Issues of neurodevelopment in biological and artificial neural networks,"

Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES'2001), 40–45.
28. S. K. Chalup and A. D. Blair 2002, "Software for analysing recurrent neural nets that learn to predict non-regular languages," *International Colloquium on Grammatical Inference (ICGI'2002)* (LNAI 2484, Springer-Verlag), 296–298.
29. S. K. Chalup and A. D. Blair 2003, "Incremental training of first order recurrent neural networks to predict a context-sensitive language," in preparation.
30. B. Clancy, R. B. Darlington and B. L. Finlay 2000, "The course of human events: Predicting the timing of primate neural development," *Developmental Science* **3**(1), 57–66.
31. B. Clancy, R. B. Darlington and B. L. Finlay 2001, "Translating developmental time across mammalian species," *Neurocience* **105**(1), 7–17.
32. B. Clancy and B. L. Finlay 2001, *Neural Correlates of Early Language Learning* (In Tomasello and Bates[138]), 307–330.
33. E. Clark, ed. 1993, *The Proceedings of the 24th Annual Child Language Research Forum* (Center for the Study of Language and Information, Stanford, CA).
34. R. B. Darlington, S. A. Dunlop and B. L. Finlay 1999, "Neural development in metatherian and eutherian mammals: Variation and constraint," *Journal of Comparative Neurology* **411**, 359–368.
35. C. R. Darwin 1859, *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle of Life* (Muray, London).
36. Y. Davidor, H.-P. Schwefel and R. Männer, ed. 1994, *Parallel Problem Solving from Nature – PPSN III*, *International Conference on Evolutionary Computation*, *The Third Conference on Parallel Problem Solving from Nature*, *Proceedings* (LNCS 866, Springer-Verlag).
37. P. Dayan and L. F. Abbott 2001, *Theoretical Neuroscience. Computational and Mathematical Modeling of Neural Systems* (MIT Press).
38. K. De Jong, D. B. Fogel and H.-P. Schwefel 1997, *A History of Evolutionary Computation* (In Bäck *et al.*[6]).
39. J. A. DeCarlos and D. D. M. O'Leary 1992, "Growth and targeting of subplate axons and establishment of major cortical pathways," *Journal of Neuroscience* **12**, 1194–1211.
40. T. Downs, M. Frean and M. Gallagher, ed. 1998, *Proceedings of the Ninth Australian Conference on Neural Networks* (Brisbane, University of Queensland).
41. K. Doya 1999, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?" *Neural Networks* **12**, 961–974.
42. K. Doya 2000, "Metalearning, neuromodulation and emotion," *Proceedings of the 13th Toyota Conference on Affective Minds.*
43. K. Doya 2002, "Metalearning and neuromodulation," *Neural Networks* **15**, 495–506.
44. J. Dudel, R. Menzel and R. F. Schmidt, ed. 1996, *Neurowissenschaft: Vom Molekül zur Kognition* (Springer-Verlag).
45. G. M. Edelman and V. B. Mountcastle, ed. 1978, *The Mindful Brain* (MIT Press).
46. J. L. Elman 1990, "Finding structure in time," *Cognitive Science* **14**, 179–211.
47. J. L. Elman 1993, "Learning and development in neural networks: The importance of starting small," *Cognition* **48**, 71–99.
48. S. E. Fahlman and C. Lebiere 1990, "The cascade-correlation learning architecture," (In D. S. Touretzky, ed. *Advances in Neural Information Processing Systems 2*, 534–532).
49. S. E. Fahlmann 1988, "An empirical study of learning speed in back-propagation networks," (Technical Report CMU-CS-88-162, Carnegie Mellon University).
50. B. L. Finlay and R. B. Darlington 1995, "Linked regularities in the development and evolution of mammalian brains," *Science* **268**(5217), 1578–1584.
51. B. L. Finlay, R. B. Darlington and N. Nicastro 2001, "Developmental structure in brain evolution," *Behavioral and Brain Sciences* **24**(263).
52. D. B. Fogel 2000, *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence* (IEEE Press, New York, 2nd edition).
53. L. J. Fogel 1962, "Autonomous automata," *Industrial Research* **4**, 14–19.
54. B. Fritzke 1994, "Fast learning with incremental RBF networks," *Neural Processing Letters* **1**(1), 2–5.
55. B. Fritzke 1994, "Growing cell structures – A self-organizing network for supervised and unsupervised learning," *Neural Networks* **7**(9), 1441–1460.
56. J. M. Fuster 1995, *Memory in the Cerebral Cortex: an Empirical Approach to Neural Networks in the Human and Nonhuman Primate* (MIT Press).
57. S. Gallant 1990, "Perceptron-based learning algorithms," *IEEE Transactions on Neural Networks* **1**(2), 179–191.
58. C. R. Gallistel 2000, *The Replacement of General-Purpose Learning Models with Adaptively Specialized Learning Modules*, Chap. 81, (In Gazzaniga,[60] 2nd edition), 1179–1191.
59. M. S. Gazzaniga, ed. 1995, *The Cognitive Neurosciences* (MIT Press, 1st edition).
60. M. S. Gazzaniga, ed. 2000, *The New Cognitive Neurosciences* (MIT Press, 2nd edition).
61. T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck and K. Fukushima, eds. 1999,

Proceedings, 6th International Conference on Neural Information Processing (*ICONIP'99*).
62. S. Geman, E. Bienenstock and R. Doursal 1992, "Neural networks and the bias/variance dilemma," *Neural Computation* **4**, 1–58.
63. B. N. Goldowsky 1993, *Modeling the Effects of Processing Limitations on the Acquisition of Morphology: The Less is More Hypothesis* (In Clark[33]), 124–138.
64. J. L. Gould and P. Marler 1987, "Learning by instinct," *Scientific American* **256**, 74–85.
65. S. Haykin 1999, *Neural Networks. A Comprehensive Foundation* (Prentice Hall, 2nd edition).
66. R. Hayward 2000, *Analytic and Inductive Learning in an Efficient Connectionist Rule-based Reasoning System* (Ph.D. thesis, School of Computing Science, Queensland University of Technology, Brisbane, Australia).
67. D. O. Hebb 1949, *The Organization of Behavior: A Neuropsychological Theory* (John Wiley, New York).
68. J. Hertz, A. Krogh and R. G. Palmer 1991, *Introduction to the Theory of Neural Computation* (Addison-Wesley Publishing Company).
69. J. H. Holland 1962, "Outline for a logical theory of adaptive systems," *Journal of the ACM* **9**, 297–314.
70. J. H. Holland 1992, *Adaptation in Natural and Artificial Systems*: *An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (MIT Press).
71. D. H. Hubel and T. N. Wiesel "Binocular interaction in the striate cortex of kittens reared with artificial squint 1965, *Journal of Neurophysiology* **21**, 1041–1059.
72. P. R. Huttenlocher and C. deCourten 1987, "The development of synapses in striate cortex of man," *Human Neurobiology* **6**, 1–9.
73. J. S. Johnson and E. L. Newport 1989, "Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language," *Cognitive Psychology* **21**, 60–99.
74. J. Joyce 1996, "When/why/of what is less more? (Master's thesis, Centre of Cognitive Science, University of Edinburgh).
75. J. H. Kaas 2000, "Why is brain size so important: Design problems and solutions as neocortex gets bigger or smaller," *Brain and Mind* **1**, 7–23.
76. R. Kalil 1989, "Synapse formation in the developing brain," *Scientific American* **261**, 76–85.
77. E. R. Kandel, J. H. Schwartz and T. M. Jessell 2000, *Principles of Neural Science* (McGraw-Hill, 4th edition).
78. L. C. Katz, M. Weliky and J. C. Crowley 2000, *Activity and the Development of the Visual Cortex*: *New Perspectives*, Chap. 13, (In Gazzaniga,[60] 2nd edition), 199–212.
79. P. S. Katz 1999, *Beyond Neurotransmission*: *Neuromodulation and its Importance for Information Processing* (Oxford University Press).
80. S. Kirby and J. R. Hurford 1997, "The evolution of incremental learning: language, development and critical periods," *Edinburgh Occasional Papers in Linguistics*.
81. E. I. Knudsen and P. F. Knudsen 1990, "Sensitive and critical periods for visual calibration of sound localization by barn owls," *Journal of Neuroscience* **10**, 222–232.
82. T. Kohonen 1984, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin).
83. A. S. LaMantia and P. Rakic 1990, "Axon overproduction and elimination in the corpus callosum of the developing rhesus monkey," *Journal of Neuroscience* **291**, 520–537.
84. A. S. LaMantia and P. Rakic 1994, "Axon overproduction and elimination in the anterior commissure of the developing rhesus monkey," *Journal of Computational Neuroscience* **340**, 328–336.
85. E. H. Lenneberg 1967, *Biological Foundations of Language* (Wiley).
86. M. S. Lidow and P. Rakic 1992, "Scheduling of monoamineric neurotransmitter receptor expression in the primate neocortex during postnatal development," *Cerebral Cortex* **2**, 401–416.
87. Y. Liu and X. Yao, "A cooperative ensemble learning system," *Proc. of the 1998 IEEE International Joint Conference on Neural Networks* (*IJCNN'98*) (Anchorage, USA).
88. P. Marler 1987, *The Instinct to Learn* (In Carey and Gelman[24]), 37–66.
89. T. M. Martinetz 1993, "Competitive Hebbian learning rule forms perfectly topology preserving maps," *ICANN'93*: *International Conference on Artificial Neural Networks*, 427–434.
90. S. K. McConnell, A. Ghosh and C. J. Shatz 1994, "Subplate pioneers and the formation of descending connections from cerebral cortex," *Journal of Neuroscience* **14**, 1892–1907.
91. B. W. Mel 1999, *Why Have Dendrites? A Computational Perspektive* (In Stuart *et al.*[133]).
92. Z. Michalewicz 1996, *Genetic Algorithms + Data Structures = Evolution Programs* (Springer-Verlag).
93. G. F. Miller, P. M. Todd and S. U. Hedge 1989, *Designing Neural Networks Using Genetic Algorithms* (In Schaffer[124]), 379–389.
94. V. B. Mountcastle 1978, *An Organising Principle for Cerebral Function*: *The Unit Module and the Distributed System* (In Edelman and Mountcastle[45]).
95. J.-P. Nadal 1989, "Study of an algorithm for a feedforward network," *International Journal of Neural Systems* **1**(1), 55–59.

96. E. Newport 1990, "Maturational constraints on language learning," *Cognitive Science* **14**, 11–21.
97. H. J. Novacek 1992, "Mammalian phylogeny – shaking the tree," *Nature* **356**, 121–125.
98. O. M. Omidvar and C. L. Wilson, eds. 1997, *Progress In Neural Networks* (Ablex Publishing Corporation, Norwood, New Jersey).
99. R. W. Oppenheim 1991, "Cell death during development of the nervous system," *Annu. Rev. Neurosci.* **14**, 453–501.
100. W. Penfield and L. Roberts 1959, *Speech and Brain Mechanisms* (New York, Atheneum).
101. B. Pettmann and C. E. Henderson 1998, "Neuronal cell death," *Neuron* **20**, 633–647.
102. S. Pinker 1994, *The Language Instinct: How The Mind Creates Language* (New York, William Morrow).
103. J. C. Platt 1991, "A resource-allocating network for function interpolation," *Neural Computation* **3**(2), 213–225.
104. V. W. Porto 1997, *Neural-Evolutionary Systems*, Chap. D1, (In Bäck *et al.*[6]), D1.1:1–D1.3:2.
105. L. Prechelt 1997, "Investigation of the CasCor family of learning algorithms," *Neural Networks* **10**, 885–896.
106. D. J. Price and D. J. Willshaw 2000, *Mechanisms of Cortical Development* (Oxford University Press).
107. S. R. Quartz and T. J. Sejnowski 1997, "The neural basis of cognitive development: A constructivist manifesto," *Behavioral and Brain Sciences* **20**(4), 537–596.
108. P. T. Quinlan 1998, "Structural change and development in real and artificial neural networks," *Neural Networks* **11**, 577–599.
109. P. Rakic 1985, "Limits of neurogenesis in primates," *Science* **227**, 154–156.
110. P. Rakic 1995, *Corticogenesis in Human and Nonhuman Primates*, Chap. 8, (In Gazzaniga,[59] 1st edition), 127–145.
111. P. Rakic 2000, *Setting the Stage for Cognition: Genesis of the Primate Cerebral Cortex*, Chap. 1, (In Gazzaniga,[60] 2nd edition), 7–21.
112. P. Rakic, J.-P. Bourgeois, M. E. Eckenhoff, N. Zecevic and P. S. Goldman-Rakic 1986, "Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex," *Science* **232**, 232–235.
113. J. Randløv 2000, "Shaping in reinforcement learning by changing the physics of the problem," *Proceedings of the Seventeenth International Conference on Machine Learning* (*ICML'2000*), 767–774.
114. J. Randløv, A. G. Barto and M. T. Rosenstein 2000, "Combining reinforcement learning with a local control algorithm," *Proceedings of the Seventeenth International Conference on Machine Learning* (*ICML'2000*), 775–782.
115. I. Rapin and S. Segalowitz, eds. 2002, *Handbook of Neuropsychology, Child Neurology*, Vol. 6, (Elsevier, 2nd edition).
116. R. Reed 1993, "Pruning algorithms – a survey," *IEEE Transactions on Neural Networks* **4**, 740–747.
117. M. Riedmiller 1994, "Advanced supervised learning in multi-layer perceptrons – from backpropagation to adaptive learning algorithms," *Computer Standards and Interfaces* **5**, (Special Issue on Neural Networks).
118. M. Riedmiller and H. Braun 1993, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm.," (In H. Ruspini, ed., *Proceedings of the IEEE International Conference on Neural Networks* (*ICNN*), *San Francisco*, *CA*, *March 28–April 1*, *1993)*, 586–591.
119. D. L. T. Rohde and D. C. Plaut 1997, "Simple recurrent networks and natural language: How important is starting small?" *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, 656–661.
120. D. L. T. Rohde and D. C. Plaut 1999, "Language acquisition in the absence of explicit negative evidence: How important is starting small?" *Cognition* **72**, 67–109.
121. A. J. F. Rooij, L. C. Jain and R. P. Johnson 1996, *Neural Network Training Using Genetic Algorithms* (World Scientific Publishing Co. Pte. Ltd.).
122. G. Roth and M. F. Wullimann 1996, *Evolution der Nervensysteme und der Sinnesorgane*, Chap. 1, (In Dudel *et al.*[44]).
123. W. S. Sarle 1999, comp.ai.neural-nets faq: Learning. posted in: comp.ai.neural-nets.
124. J. D. Schaffer, ed. 1989, *Proceedings of the Third International Conference on Genetic Algorithms and Their Applications* (Morgan Kaufmann, San Mateo, CA).
125. J. D. Schaffer, D. Whitley and L. J. Eshelman 1992, "Combinations of genetic algorithms and neural networks: A survey of the state of the art," *International Workshop on Combinations of Genetic Algorithms and Neural Networks* (*1992*), *Baltimore*, *Maryland* (Los Alamos, California, IEEE Press), 1–37.
126. H.-P. Schwefel 1965, "Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik. Master's thesis, TU Berlin, Hermann Föttinger Institut für Hydrodynamik.
127. H.-P. Schwefel 1977, *Numerische Optimierung von Computermodellen mittels der Evolutionsstrategie* (Birkhäuser).
128. H.-P. Schwefel 1995, *Evolution and Optimum Seeking* (John Wiley).
129. O. G. Selfridge, R. S. Suton and A. G. Barto 1985, "Training and tracking in robotics," *Proceedings of the Ninth International Joint Conference in Artificial Intelligence*, 670–672.

130. A. J. C. Sharkey 1996, "On combining artificial neural nets," *Connection Science* **8**(3 & 4), 299–313
131. F. Smieja 1993, "Neural network constructive algorithms: Trading generalization for learning efficiency?" *Circuits, Systems and Signal Processing* **12**, 331–374.
132. L. R. Squire 1986, "Mechanisms of memory," *Science* **232**, 1612–1619.
133. G. Stuart, N. Spruston and M. Hausser, eds. 1999, *Dendrites* (Oxford University Press).
134. J. E. Sulston and H. R. Horvitz 1977, "Post embryonic cell lineages of the nematode Caenorhabditis elegans," *Developmental Biology* **56**, 110–156.
135. R. S. Sutton and A. G. Barto 1998, *Reinforcement Learning: An Introduction* (MIT Press).
136. G. Tesauro 1992, "Practical issues in temporal difference learning," *Machine Learning* **8**, 257–278.
137. G. Tesauro 1994, "TD-gammon, a self-teaching backgammon program, achieves master-level play," *Neural Computation* **6**, 215–219.
138. M. Tomasello and E. Bates, eds. 2001, *Language Development. The Essential Readings* (Blackwell Publishers).
139. J. Tooby and L. Cosmides 2000, *Toward Mapping the Evolved Functional Organization of Mind and Brain*, Chap. 80, (In Gazzaniga,[60] 2nd edition), 1167–1178.
140. R. Tuttle, Y. Nakagawa, J. E. Johnson and D. D. M. O'Leary 1999, "Defects in thalamocortical axon pathfinding correlate with altered cell domains in Mash-1-deficient mice," *Development* **126**, 1903–1916.
141. V. N. Vapnik 1995, *The Nature of Statistical Learning Theory* (Springer-Verlag).
142. G. Weiss 1997, *Towards the Sythesis of Neural and Evolutionary Learning*, Chap. 6, (In Omidvar and Wilson[98]), 145–176.
143. T. N. Wiesel and D. H. Hubel 1965, "Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens," *Journal of Neurophysiology* **28**, 1029–1040.
144. B. A. Wilson and N. Moffat, eds. 1992, *Clinical Management of Memory Problems* (Singular Publishing Group, Inc: San Diego, CA, USA, 2nd edition).
145. N. J. Woolf 1998, "A structural basis for memory storage in mammals," *Progress in Neurobiology* **55**, 59–77.
146. X. Yao 1999, "Evolving artificial neural networks," *Proceedings of the IEEE* **87**(9), 1423–1447.
147. X. Yao and Y. Liu 1997, "A new evolutionary system for evolving artificial neural networks," *IEEE Transactions on Neural Networks* **8**(3), 694–713.
148. X. Yao and Y. Liu 1998, "Making use of population information in evolutionary artificial neural networks," *IEEE Transactions on Systems, Man and Cybenetics* **28**(2).
149. Z. Zeng, R. M. Goodman and P. Smyth 1994, "Discrete recurrent neural networks for grammatical inference," *IEEE Transactions on Neural Networks* **5**(2), 320–330.
150. B.-T. Zhang 1994, "An incremental learning algorithm that optimizes network size and sample size in one trial," *Proceedings of the International Conference on Neural Networks* (*ICNN-94*), 215–220.