Evolutionary Autonomous Agents: A Neuroscience Perspective

Eytan Ruppin
School of Computer Science and School of Medicine
Tel-Aviv University, Tel-Aviv, 69978, Israel
ruppin@math.tau.ac.il

December 29, 2001

Abstract

This paper examines the research paradigm of neurally-driven Evolutionary Autonomous Agents (EAAs), from a neuroscience perspective. Two fundamental questions are addressed: 1. Can EAA studies shed new light on the structure and function of biological nervous systems? 2. Can these studies lead to the development of new neuroscientific analysis tools? The value and significant potential of EAA modeling in both respects is demonstrated and discussed. While the study of EAAs as a neuroscience research methodology still faces difficult conceptual and technical challenges, it is a promising and timely endeavor.

1 Introduction

Recent years have witnessed a growing interest in the study of neurally-driven evolved autonomous agents (EAAs). These studies, part of the field of Evolutionary Computation and Artificial Life (see [1, 2, 3, 4] for general introductory textbooks), involve agents that live in an environment and autonomously perform typical animat tasks like gathering food, navigating, evading predators, and seeking prey and mating partners. Each agent is controlled by an Artificial Neural Network (ANN) "brain". This network receives and processes sensory inputs from the surrounding environment and governs the agent's behavior via the activation of the motors controlling its actions. The agents can be either software programs living in a simulated virtual environment, or hardware robotic devices. Their controlling networks are developed via Genetic Algorithms (GAs) that apply some of the essential ingredients of inheritance and selection to a population of agents that undergo evolution.

A typical EAA experiment consists of a population of agents that are evolved using a genetic algorithm over many generations to best survive in a given environment (see Figure 1). In general, agents may have different kinds of controllers and encode also sensors and motors in the their genome, but we focus in this review on agents with a genome that solely encodes their controlling neural network. At the beginning of each generation, a new population of agents is generated by selecting the fittest agents of the previous generation and letting them mate – i.e., form new agent genomes via genetic recombination followed by mutations that introduce additional variation in the population. The genomes formed in this process are "transcribed" to form new agents that are placed in the environment for a given amount of time, after which each agent receives a fitness score that designates how well it performed the evolutionary task. This ends a generation cycle, and a new generation is initiated. Typically, this evolutionary "search" process is repeated for many generations until the agents' fitness reaches a plateau and further evolutionary adaptation does not occur. *The result is a final population of best-fitted agents, whose emergent behavior and underlying neural dynamics can now be thoroughly studied in "ideal conditions":* One has full control on manipulating the environment and other experimental conditions. More important, one has complete knowledge of the agents behavior on one hand, and the controlling network's architecture and dynamics, on the other. This scenario is illustrated below in a concrete example of a typical EAA navigation and foraging experiment adapted from [5].

The agents in the model live in a grid arena of size 30x30 cells surrounded by walls (Figure 2).

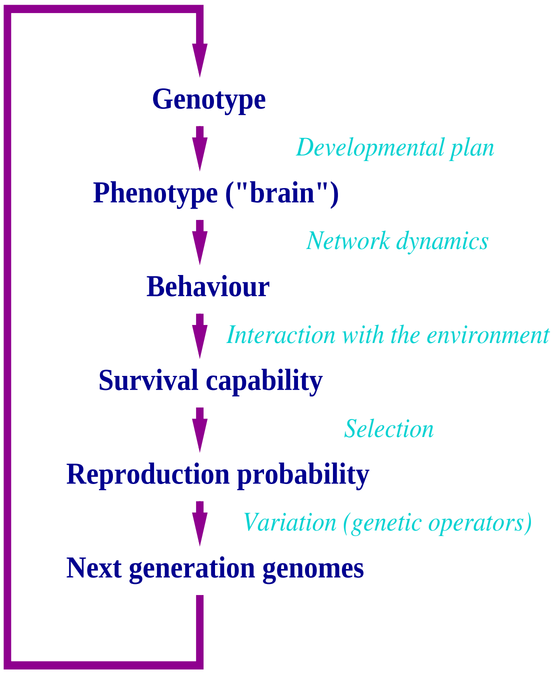

Figure 1: The Paradigm of Evolutionary Autonomous Agents

"Poison" is randomly scattered all over the arena (consuming this resource results in a negative reward). "Food", the consumption of which results in a positive reward, is randomly scattered in a restricted 10x11 "food zone" in the southwest corner of the arena. The agents' behavioral task is to eat as much of the food as they can while avoiding the poison. The complexity of the task stems from the partial sensory information the agents have about their environment. The agents are equipped with a set of sensors, motors, and a fully-recurrent ANN controller. The neurocontroller is coded in the genome and evolved; the sensors and motors are given and constant.

The initial population consists of 100 agents equipped with random neuro-controllers. Each agent is evaluated in its own environment, which is initialized with 250 poison items and 30 food items. The life cycle of an agent (an *epoch*) is 150 time steps, in each of which one motor action takes place. At the beginning of an epoch the agent is introduced to the environment at a random location and orientation. At the end of its life-cycle each agent is assigned a fitness score calculated as the total amount of food it has consumed minus the total amount of poison it has eaten. Simulations last for a large number of generations, ranging between 10000 and 30000.

Each agent is controlled by a fully recurrent binary neural network consisting of 15 to 50 neurons

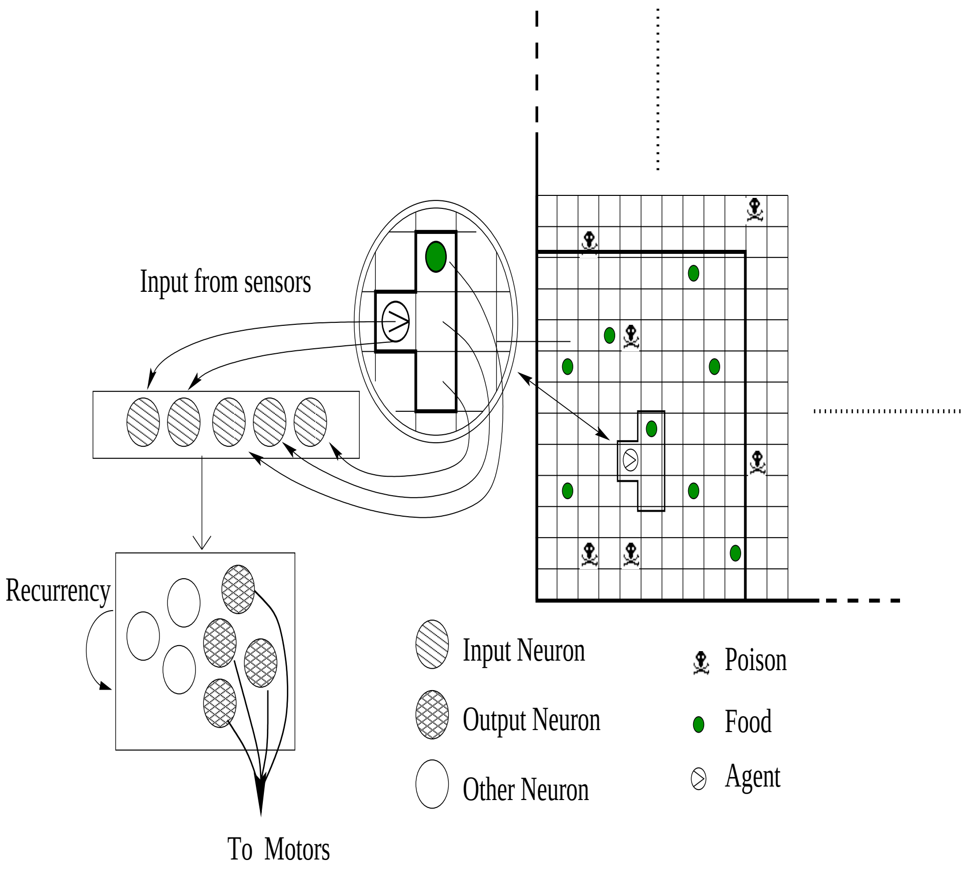

Figure 2: *An outline of the grid arena* (southwest corner) and the agent's controlling network. For illustration purposes the borders of the food zone are marked, but they are invisible to the agent. The agent is marked by a small arrow on the grid, whose direction indicates its orientation. The T-shape marking in front of the agent denote grid cells which it senses; the curved lines indicate where in the arena each of the sensory inputs comes from. Output neurons and interneurons are all fully connected to each other.

(the number is fixed within a given simulation run). Of these, 5 are dedicated sensory neurons, whose values are clamped to the sensory input, and have no input from other neurons. Four output motor neurons command the agent's motors. Network updating is synchronous. In each step a sensory reading occurs, network activity is then updated, and a motor action is taken according to the resulting activity in the designated output neurons.

The agents are equipped with a basic sensor consisting of five probes. Four probes sense the grid cell the agent is located in and the three grid cells immediately ahead of it (see Figure 2). These probes can sense the difference between an empty cell, a cell containing a resource (either poison or food – with no distinction between those two cases), and the arena boundary. The fifth probe can be thought of as a *smell probe*, which can discriminate between food and poison if these are present in the cell occupied by the agent. The motor system allows the agent to go forward, turn 90 degrees in each direction, and attempt to eat. Eating is a costly process as it requires a time step with no other movement, in a lifetime of limited time-steps.

Each agent carries a chromosome defining the structure of its N-neuron controlling network, consisting of $N(N-5)$ real numbers specifying the synaptic weights. At the end of a generation a phase of sexual reproduction takes place, composed of 50 reproduction events. In each of these events, two agents from the parents population are randomly selected with probability proportional to their fitness (the amount of food minus poison eaten). Then, their chromosomes are crossed over and mutated to obtain the agents of the next generation. A point-crossover with probability 0.35 was used, after which point mutations were randomly applied to 2% of the locations in the genome. These mutations changed the pertaining synaptic weights by a random value between -0.6 and +0.6. The resulting chromosomes define two agents of the next generation. Essentially, the genetic algorithm performs a search for the best synaptic weight values in the space of all possible network architectures that may be composed of the controlling neurons. Figure 3 shows a typical evolutionary run. The average initial population fitness is very low (around -0.05). As evolution proceeds better controllers emerge and both the best and average fitness in the population increase until a plateau is reached.

Current EAA studies have been able to successfully evolve artificial networks of several dozen neurons and hundreds of synapses, controlling agents performing non-trivial behavioral tasks (see [6, 7, 8, 9, 10] for recent reviews). These networks are *less biased* than conventional neural networks used in neuroscience modeling as their architecture is not pre-designed in many cases. They are

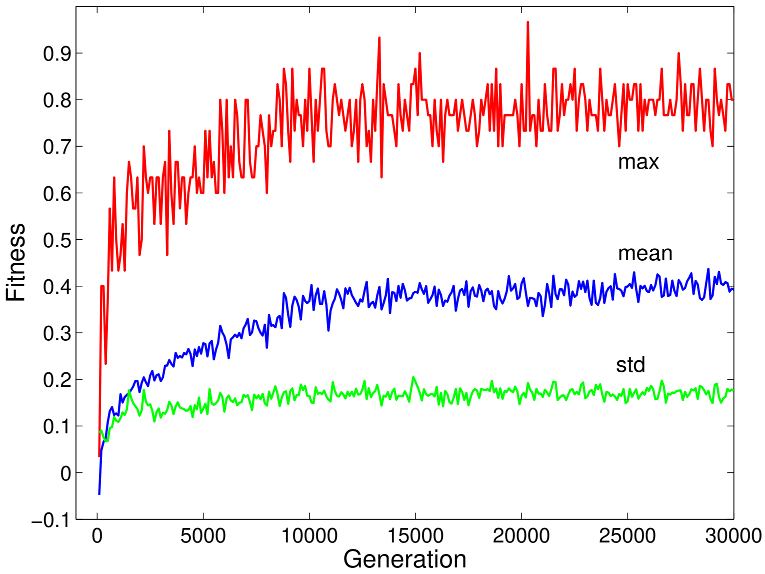

Figure 3: *A typical evolutionary run:* maximum, mean and standard deviation of fitness in a population plotted over 30000 generations of an evolutionary run. Values are plotted every 100 generations. The fitness is evaluated over a single epoch for each agent and the mean is the average of the fitness in the population (100 agents).

the *emergent result* of a simplified and idealized process that models the evolution of intelligent, neurally-driven life forms. This fundamental property naturally raises the possibility of using these agents as a vehicle for studying basic questions concerning neural processing. This potential is further substantiated by two additional observations: First, *Feasibility:* the small size of the evolved networks, the simplicity of their environments and behavioral tasks, coupled with the *full information* available regarding the model dynamics, form conditions that help make the analysis of the network's dynamics an amenable task. Second, *Relevancy:* since the networks are evolved in biologically motivated animat environments, their analysis may potentially reveal interesting insights into the workings of biological systems.

The use of EAA models as a neuroscience research tool is a very complex and challenging scientific endeavor. Even the seemingly simple task of understanding the neural processing of small, fully transparent EAA agents is a very difficult one. Furthermore, the relevance of findings in EAA studies should be considered with caution (as is perhaps true regarding the relevance of neural modeling studies in general). We should always be aware of the many simplifications involved in these models; while enabling one to address systems that would be otherwise too complex to investigate, it may well be that interesting and even vital components of a system are missed.

Yet, some of the results already obtained testify to the beneficial role of EAA as a fundamental and simple methodology of neuroscience investigation. The goal of this paper is to review these results, many of which have not yet been brought to the attention of neuroscientists due to the different scientific communities involved. The paper is further organized to answer two fundamental questions: 1. *Do EAA studies produce results that bring new* insights *to neuroscientific issues?* – Section 2 aims to answer this question by providing a selective overview of EAA studies that bear upon current neuroscientific research topics. 2. *Can EAA studies lead to the development of new* tools *for neuroscience research?* – Section 3 describes methods for systematic analysis of the evolved agents' controlling networks that may be used for understanding neural information processing in both animat and biological systems. The last section returns to discuss the critique and the future of EAA studies in neuroscience.

2 EAA studies: A Neuroscience Perspective

EAA studies typically evolve neurally-driven computer simulated animats or robots that solve a variety of cognitive and behavioral tasks. As such, they form an intuitively appealing approach for modeling and studying biological nervous systems. *However, do current studies really begin to realize this potential? And what can be learned from these studies?* Here we selectively review a few studies that explore specific questions that are of relevance to neuroscience. Our selection is by no means exhaustive. We begin with studies modeling simple animal systems, and proceed with models of evolution and learning. Finally, a description of evolutionary computation investigations of cortical organization leads us to briefly review and discuss various existing models of genotype-to-phenotype encodings.

The task of evolving neural network controllers of detailed models of animal systems is yet too difficult for the power of existing evolutionary computation systems, but a few interesting investigations in this direction have already been carried out. One study has evolved neural networks that reproduce the habituation of the touch sensitive behavior of the nematode C. elegans [11]. This behavior is comprised of backward movement in response to stimuli applied to the worm's head, and forward motion in response to stimulation of the tail. A comparison of data gathered from real worms and from lesion analysis of the evolved networks showed that lesioning the corresponding artificial and biological interneurons in both systems had similar global regulatory effects, resulting in disturbances in habituation. But a mismatch in the processing of the corresponding artificial and biological sensory neurons has lead to the formulation and testing of a more refined EAA model, which included the simulation of developmental events during network formation. The revised model succeeded in resolving the mismatch, manifesting emergent sensory neurons that play a major role in mediating touch sensation and others that play only a minor role, in a manner similar to that observed in real worms. This new model also suggests that such a partially asymmetric neural mechanism is responsible for motor response habituation, a prediction that remains to be tested.

EAA models were also used to study the evolution and development of central pattern generation for the swimming of the lamprey [12]. The evolved neural networks guide the agents' swimming by forming connections to the muscles located along both sides of the agent's body, and generating the coordinated oscillatory patterns required for propelling the body. These oscillatory patterns could be modulated by varying the external excitation applied to the network, resulting in varying directions and speed of swimming. The best evolved controllers cover a broader ranges of frequencies, phase lags and speeds than the original, hand-crafted model [13]. Using the EAA approach it was also possible to develop biologically plausible controllers with ranges of oscillation frequency that were closer to those actually observed in the lamprey than those produced by the hand-crafted model. In agreement with the experimental findings of [14], these oscillations may be produced without excitatory interneurons. Finally, the synaptic connections formed in some of the evolved agents were, at least to some extent, similar to those observed in the real lamprey [12].

These results were obtained using fitness (optimization) functions that explicitly express the desired outcome. However, interesting biological-like phenomena may emerge in EAA models even when they are not modeled explicitly. The model of [5] described in detail above demonstrates this potential, when the agents' emergent controller networks are analyzed using conventional neuroscience methods of lesioning and receptive fields measurements. This analysis reveals a command neuron whose firing state triggers a switch between navigation and foraging behaviors that occurs immediately after the agent ingests food (recall that the agents are placed in a random location on the grid and they first have to navigate and find the food zone and then remain and forage in that zone). The firing of this command neuron (or a few such neurons) essentially switches between two evolved different input-output subnetworks controlling navigation vs. foraging. These EAA findings closely resemble findings of command neurons in animals, including crayfish [15], Aplysia [16, 17, 18], Clione [19], crabs [20, 21] and lobsters [22]. In some cases the animal behavior is modulated by the command neurons on the basis of certain sensory stimuli [16], and in particular, as in the EAA simulation, by food arousal [17, 18]. This activity has been shown to control a variety of motor repertoires, mainly by inducing different activity patterns in the same network by modulating neuronal activity [22, 21], again, in a manner similar to that found in the EAA study. Obviously, biological reality is much more complex and despite the resemblance, there are many significant differences. For example, chemical neuro-modulation plays an important role in command neuron activity [23, 19], while absent from the current model (this is not, however, an inherent limitation as such neuro-modulation has been incorporated in EAA studies, e.g., [24, 25]). All together, even though biological reality is much richer than that of a simple EAA model, these studies demonstrate that networks emerging in EAA systems can manifest interesting biological-like characteristics and provide additional computational insights.

We see two main motivations for studying biological systems with EAA models:

- The first one stems from the observation that *biologically relevant neural network models should be studied in a comprehensive system containing not only the networks themselves but also the "bodies" in which they reside*, i.e., the agent's sensors and motors and the environment in which the agent acts. As shown in various EAA studies, this embodiment is of paramount importance for providing constraints that reduce the degeneracies involved in the neural-to-behavioral mappings [26]. Moreover, EAA agents can utilize the evolved motor behaviors to augment their sensory processing, for example, by moving and turning around such that important objects in the environment are viewed from a fixed angle, making their recognition much simpler [27, 28].

- The second motivation stems from the recognition that EAA studies are a natural computational framework for studying the interaction between learning and evolution, two prime adaptation mechanisms occurring on different time scales, but interleaved and interacting in complex ways (see [29] for a review). A primary focus of EAA studies of learning and evolution has been the Baldwin Effect which states that learning can influence the course of evolution even if learned traits are not inherited (e.g., [30, 31]). Below we review just one example of EAA research studying this multi-faceted interplay between learning and evolution.

In this study, a population of EAAs was subject to both evolutionary and learning processes [32]. Each agent's controller is composed of two subnetworks; one network guides its sensorimotor processing and the other is a "teacher" network. Both networks are encoded in the agents' genome and receive the same set of inputs from the environment. The teacher network processes these inputs to calculate desired responses that are then used to modify the synaptic connections of the sensorimotor network using a supervised learning algorithm. In a typical exploration task, the combination of learning and evolution in these agents enables them to obtain significantly higher performance than agents with a similar genetically encoded sensorimotor subnetwork but without learning, i.e., without the teacher subnetwork. The key to the success of the learning-able agents is their ability to develop a genetically inherited predisposition for learning. This predisposition stems from the selection of initial weights at birth that guides behavior to select the right set of inputs, thus "channeling" learning in a successful direction. *This power of evolution to select specific emergent learning predispositions points to the potential pitfalls of studying learning in isolation, as is done at times in conventional neural networks and connectionist models.*

EAA studies have been utilized to *study the evolution of learning itself. Such an investigation can span across several levels of neuroscientific research [33].* This is demonstrated in a study evolving (near-)optimal neuronal learning rules in a simple EAA model of reinforcement learning in bumblebees [34]. Following the neural modeling study of [35], this EAA investigation studied bee foraging in a simulated 3D arena of blue and yellow flowers. During its flight, the bee processes visual information about the flower arena it currently sees to determine its next flight heading. Upon landing, it receives a reward according to the nectar produced by the flower it has landed upon. This reward is then used by the bee as a reinforcement learning cue to modify the synaptic efficacies of its controlling network. Unlike the study of [35] where the synaptic learning rules governing these modifications are pre-specified, the learning rules themselves (and not just the synaptic weights) undergo evolution. These evolved synaptic plasticity rules include a new component of heterosynaptic Hebbian learning [36, 37] that was not specified in the previous hand-crafted solution [35], giving rise to varying exploration/exploitation levels. On a higher level of description, these synaptic micro-dynamics give rise to risk aversive behavior, providing a novel, biologically founded, parsimonious explanation for risk aversion. *Thus, even simple EAA models are capable of producing complex emergent behaviors that were not specified in the model in any explicit manner.*

Evolutionary computation optimization methods can be used to directly challenge a variety of important open questions in neuroscience. For example, it is well known that a localized excitatory stimulus applied directly to the cerebral cortex can produce a surrounding peri-stimulus inhibitory zone ("Mexican Hat Pattern" of activity). This has long been viewed as surprising by some because lateral intra-cortical connections within the 150-250 micron distance involved are predominantly excitatory (asymmetric synapses; pyramidal/stellate neuron to pyramidal/stellate neuron) [38, 39, 40, 41, 42]. In this context, a recent study used a genetic algorithm to inquire whether a neuronal circuit for cortical columns could be evolved that produces peri-stimulus inhibition under the constraint that the only horizontal, inter-columnar synaptic connections permitted are excitatory ones [43]. Starting with columnar neuronal circuits having arbitrary, randomly-generated excitatory and inhibitory synaptic strengths, it proved possible to evolve, within at most a few thousand generations, neuronal circuits that produce Mexican Hat patterns of activity (see Figure 4). The apparent lateral inhibitory effects that evolved were due to the turning off of baseline horizontal spread of excitatory activity between neurons in neighboring columns. This result is interesting, not just because the evolutionary process discovered a novel candidate neural circuit for cerebral cortex, but more importantly because it suggests a new approach to generating hypotheses about the nature of complex neural circuitry in general, via a simulated evolutionary process.

A kind of "reverse" approach has shown that "indirect" biologically-inspired genotype-to-phenotype encoding enables the successful evolution of a variety of basic neural network architectures that are assumed to participate in cortical neural processing [44]. In contrast to "direct" encodings where every gene explicitly specifies a synaptic connection, "indirect" encodings include a program for determining the network architecture or its connections' weight values in a compact manner. This approach serves a dual role: From a computational perspective, its aim is to find efficient genotype-to-phenotype encodings, a key to the evolution of smart agents. It provides for compact genetic encodings of complex networks, and for filtering out genetic changes [45]. From a neuroscience perspective, its aim is to test different hypotheses about how the architecture and operation of various biological neural networks may be specified by the genome. The work on developing indirect encodings has attracted ample efforts, focusing on a few different avenues (see Box 1) .

Box1: Indirect Genotype-to-Phenotype Encodings

- *Grammar Rewriting encodings* – which employ a set of rewriting rules that are encoded in the genome. For example, in [46] the genome contains numerous blocks. Each block has five elements and is interpreted as a rewriting rule that states that the first element in the block should be transcribed to a matrix composed of the next four elements in the block. Via such an iterative decoding process a matrix specifying the network architecture is formed (and the synaptic efficacies are then determined by other mechanisms, e.g., learning). Simple grammar rewriting encodings typically generate restricted tree-like network architectures, but utilizing graph grammars one may develop more general, recurrent networks [47, 48]. The latter have been used to develop encodings which lead to the emergence of modular subnetworks – repetitive building blocks, mimicking cortical columnar structures. Such grammar-like encodings generate fairly compact genomes and hence reduce the search space of possible solutions. A variant, [49] enables the evolution of compact target networks by including a complexity term in the fitness function.

- *Developmental, Ontogenetic encodings (Geometric grammars)* – where the genome expresses a program for cell division and axonal migration that determines the phenotypic neural architecture [50, 51, 52]. In these encodings the objects undergoing development are localized on a 2-dimensional (or 3D) space, allowing for context-dependent effects from neighboring neurons, and the developmental program has a more biological flavor. Yet, the genomes generated are less compact than those generated by encoding graph grammars (scaling linearly with network size) and do not strongly bias toward the evolution of modular networks. The temporal dimension of such ontogentic development has been studied by encoding "maturation times" that regulate the expression of different blocks in the genome [52]. Overall, it is shown that this mechanism can successfully select genomes which dictate early maturation of functional vs. dysfunctional blocks. "Layering" ontogenesis in time delays full functionality to later stages in the agent's life and hence may damage its fitness, but it may enhance genetic variation by sustaining mutations that occur in genetic blocks expressed late in maturation. A few developmental encodings have aimed at developing more biologically motivated "regulatory" encodings, specifying the development of a multi cellular organism [45, 53, 54]. The identity of the subset of genes active in each moment in a given cell is determined by a complex interaction with the "transcription factors" it receives from the environment and from other cells. These models however require extensive computational resources and are still unable to evolve agents solving complex tasks.

- *Compound encodings* – recently, newly developed genetic encodings that flexibly encode both the synaptic weights and their learning rules have been shown to be superior to direct synaptic encodings, but these results await further corroboration in a wider set of EAA models [55, 56]. The efficacy of such encodings may be enhanced by self-organizing processes that refine the developing networks by exploiting regularities in the environment, e.g., by incorporating activity-based pruning in the developmental programs [57]. Their efficacy may be further enhanced by EAA models that incorporate neuro-modulatory agents. Such a diffusible agent has already been used to dynamically modulate the neurons' transfer function in a concentration dependent manner, enabling efficient evolution of more compact networks than those obtained with direct encodings [24]. In a manner analogous to [34], neuromodulation may be harnessed to guide learning in agents. Lastly, self organizing compression encodings may be used to adaptively guide search to optimal subspaces [58].

Yet, finding efficient indirect encodings still remains an extremely important open problem. The superiority of existing indirect encodings over direct ones has not yet been shown in a convincing manner: First, due to the absence of examples of agents solving complex tasks with indirect encodings that were otherwise unsolvable with direct encodings or by hand-crafting the solutions [56]. Second, because some of these encodings do not scale up well with network size [59]. The idea that an encoding successfully capturing some of the essential computational principles of biological encodings could lead to a breakthrough in our abilities to evolve complex agents is certainly compelling, but it remains to be seen if this can be done. If however some success will be obtained, then simulating such developmental processes in EAAs will probably teach us a lot about the organization and functioning of biological neural networks.

3 Analysis Of Neural Information Processing in EAAs

This section reviews research that analyzes the evolved controller networks. The dual goal of this research is to uncover principles of neural processing in animat and biological nervous systems, and to develop new methods for their analysis.

A series of studies have developed a rigorous, quantitative analysis of the dynamics of central pattern generator (CPG) networks evolved for locomotion [60, 26]. The networks evolved are of very small size, composed of 3,4 or 5 neurons. A high-level description of the dynamics of these CPG networks was developed, based on the concept of a dynamical module: a set of neurons that have a common temporal behavior, making a transition from one quasi-stable state of firing to another together. The evolved networks can be decomposed to a varying number of multi-stable dynamical modules that are traversed via successive destabilizations. In some networks the dynamic modules do not remain fixed but change over a slow time scale. Dynamical modules give new insights to CPG operation, describing them in terms of a finite state machine, and enabling a rigorous analysis of their robustness to parameter variations. They provide one possible concrete realization of Getting's hypothesis that biological CPG's are constructed from basic dynamical "building blocks" [61, 62]. Interestingly, in some cases the dynamical modules could be assigned specific functional roles but in others this assignment was not possible. This observation touches upon the fundamental topic of "understanding" neural processing by localizing and assigning specific procedures and functions to component subnetworks, an issue we shall return to in the Discussion.

Examining the variety of CPG networks evolved, it becomes evident that there is a degeneracy in the mapping between structural and functional levels [26]. That is, many different network architectures give rise to the same functional level of rhythmic activities and, consequently, to similar walking performance. Such degeneracies may be ubiquitous in biological systems, and should be considered by neuroscientists constructing biologically realistic models of CPGs [26]: due to technical recording difficulties, biophysical properties of neurons are usually measured across many individuals, and models are constructed using values from several animals. In light of the degeneracies pointed above, this construction of "chimera-like" model networks may be artificial, and may lead to the "brittleness" often observed in realistic neuronal simulations [63], where small parameter variations may lead to large changes in model dynamics. The negative findings of this EAA study are of no lesser interest; some of the evolved 5-neuron CPGs were not functionally decomposable [26]. Similarly, in spite of the progress made in neuroscience in analyzing and modeling biological CPGs, their functional decomposition remains enigmatic [62]. This may not be surprising, since, after all, behavior rather than circuit architectures are selected for in evolution [64].

The dynamic modules analysis was carried out in very small networks with a regular rhythmic pattern of activity and its application to significantly larger EAA networks with less regular dynamics remains a daunting task. How can we analyze these latter networks? Several studies have tried a variety of conventional neuroscience techniques to this end. Here we briefly review a few of these studies, each demonstrating the use of a different technique. The activity of the internal (hidden layer) neurons in the network as a function of a robot's location and orientation was charted via a simple form of receptive field measurement in [65]. The function of these intermediate neurons was generally highly distributed and dependent on previous states, but a certain interneuron playing a major role in path planning was also identified. Others have systematically clamped neuronal activity and studied its effects on the robot's behavior [66] (such as inducing rotation, straight line motion or more complex behaviors like smooth tracking of moving targets, etc.). Single-lesion analysis was used to discover the "command" neurons described in the previous section [5]. A more "procedural" kind of ablations to the network, where different processes (and not just units or links) are systematically canceled out was used recently in [67]. Overall, these studies have provided only glimpses of the processing in these networks. Moreover, an integrative EAA study systematically analyzing a neural network by employing all these methods together and comparing between them is still lacking.

In Neuroscience, assessing the importance of single neurons or cortical areas to specific tasks is traditionally done either by assessing the deficit in performance after lesioning a specific area, or by recording the activity in the area during behavior. These classical methods suffer from two fundamental flaws [68]: First, they do not take into account the probable case that there are complex interactions among elements in the system. E.g., if two neurons have a high degree of redundancy, lesioning of either one alone will not reveal its influence. Second, they are mostly qualitative measures, lacking the ability to precisely quantify the contribution of a unit to the performance of the organism and to predict the effect of new, multiple-site lesions. *The relative simplicity and the availability of full information about the network's structure and dynamics make EAA models an ideal test-bed for studying neural processing.* In this framework, a rigorous, operative definition for the neurons' (or, cortical regions) contributions to the organisms performance in various tasks and a novel Functional Contribution Algorithm (FCA) to measure them have been recently presented [68]. This operative definition permits an accurate prediction of the performance of EAA agents after multi-lesion damage, and yields, at least in principle, a precise quantification of the distribution of processing in the network, a fundamental open question of neuroscience [69, 70].

To understand the concept of "contributions", conceive of an agent (either natural or artificial) with a controlling network of N interconnected neurons (or more generally, units) that performs a set of P different functional tasks in the environment in which it is embedded. *Who does What?* - addressing this question, it is natural to think in terms of a *contribution matrix*, where C_{ik} is the contribution of element i to task k, as shown in Figure 5. The data analyzed for computing the contribution matrix is gathered by inflicting a series of multiple lesions onto the agent's network (obviously, there are quite a few different ways of lesioning networks, e.g., by knocking out neurons, or by severing all incoming or all outgoing synapses from a neuron. The latter option is assumed in the FCA). After each lesion, the resulting (corresponding) performance of the agent in different tasks is measured. Given this data, the FCA finds the contribution values C_{ik} which provide the best performance prediction for new, multiple-site lesions. Following the spirit of [71]), the *localization* L_k of task k can now be defined as a deviation from equipotentiality along column k (e.g., L_1 in Figure 5), and similarly, S_i, the *specialization* of neuron i is the deviation from equipotentiality along the row i of the matrix (e.g., S_2 in Figure 5).

The FCA algorithm was applied to the analysis of EAA neurocontrollers evolved in [5], which are recurrent neural networks with complex interactions among the elements, providing a precise prediction of the effects of new multiple lesions [68]. This remarkable performance has been obtained primarily due to the utilization of a general monotone but non-linear performance prediction function. However, we now find that more complex EAA neurocontrollers cannot be accurately described using an FCA based on contributions of *single* units only. Precise multi-lesion prediction in these networks requires the consideration of contributions from additional, functionally important *conjunctions* of the basic units. These findings strongly suggest that the classic, conventional, thinking in neuroscience aiming to decompose the processing of various tasks to a set of individual distinct regions is a gross oversimplification. It should also be noted that *the current ongoing development of more efficient derivatives of the FCA algorithm would not have been possible without the existence of the ample body of data provided by the EAA investigations.*

Multi-lesion analysis algorithms like the FCA are important in neuroscience for the analysis of reversible inactivation experiments, combining reversible neural cooling deactivation with behavioral testing of animals [72]. They can also be used for the analysis of transcranial magnetic stimulation studies which aim to induce multiple transient lesions and study their cognitive effects (see [73] for a review). Another possible use is the analysis of functional imaging data by assessing the contributions of each element to the other, i.e., extending previous network's effective connectivity studies employing linear models (e.g., [74]). Applying algorithms such as the FCA should prove useful in obtaining insights into the organization of natural nervous systems, and settling the long-lasting debate about local versus distributed computation.

4 Discussion

In his illuminating treatise on the fictional autonomous "vehicle" agents, Valentino Braitenberg states two important observations: ".. in most cases our analysis [of a certain "type 6" artificial brains] would fail altogether: the wiring that produces their behavior may be so complicated and involved that we will never be able to isolate a simple scheme. And yet it works.." [75]. Indeed, EAA models work and are a relevant neuroscience research tool. Yet, even the simple models currently studied may be fairly complex, and their analysis forms a difficult challenge. *But if this is so with regard to these much simplified systems, what about the task of understanding real, natural, nervous systems?* The "central dogma" of addressing the latter challenge in neuro and cognitive sciences research has been the knowledge-based engineering approach. This approach aims to analyze and conceptualize neural information processing in terms of operations that manipulate representations, as exemplified in Marr's work on Vision [76]. The necessity and computational value of such representations has, however, been seriously questioned by robotics and adaptive behavior researchers in recent years (e.g., [77]). A very interesting review of numerous EAA studies that have evolved visual processing agents was recently presented [78]. Like the Braitenberg Vehicle robots that do not need representations, all the agents in the studies surveyed in this review did not employ representations in the conventional sense, at least as far as the researchers' analysis techniques could tell. Rather, understanding the controller's dynamics requires treating the agent and its environment as a coupled, embedded, dynamical system [78, 79]. Furthermore, there is considerable evidence supporting the possibility that the same is true for animals [78].

Indeed, it is just the end of the beginning. Though we hope that the examples presented are convincing, one should not underestimate the difficulties – the most important one being that it still remains to be seen if the EAA paradigm can generate really complex agents on the scale of animate systems, and if so, if we will be able to analyze them. Current EAA modeling is limited in quite a few important ways: First, the vast majority of current EAA models employ simple binary neurons and the investigation of EAAs driven by spiking networks and temporal synaptic dynamics seems to be a necessary step if we want to make closer contact with animate neural processing. Second, even the brains of the most simple biological organisms employ hundreds and thousands of neurons with orders of magnitude larger amounts of synapses. It is quite obvious that we shall not be able to evolve neural controllers of this magnitude using simple direct encodings. The development of new genotype-to-phenotype encodings is critical to enable the evolution of smarter agents. Third, current EAA models employ very simplistic forms of embodiment, i.e., very rudimentary sensors and motors in very elementary environments. More elaborate and realistic sensors and motors must be developed if we wish to really study sensorimotor processing. Moreover, these sensors and motors should probably be co-evolved with their controllers. And finally, as the scale and the complexity of the evolved networks grows, the challenge of finding new ways for analyzing the evolved networks will become more complicated. In its current nascent stage, the EAA paradigm is still fairly limited, but a gradual, incremental approach for its further development is feasible and within our reach. One should also bear in mind that until now there have been many fewer EAA studies of neuroscience questions compared with the volume of more traditional computational neuroscience investigations.

This brings us to what is perhaps the major critique of EAA in Neuroscience – the notion that evolution may take many paths and directions, and hence the findings observed in EAA models may not teach us anything significant about biological nervous systems. As outlined throughout this paper, there are various reasons to believe that this is not the case, and that biologically relevant principles of neural information processing may best be studied in these models. But the importance of EAA research goes beyond that: In my mind, its primary value for neuroscience is first and foremost its ability to serve as a very simple, but *emergent, accessible and concrete, even if artificial test-bed*, for thinking about neural processing principles and for developing new methods for deciphering them. EAA is a promising way of making neuroscience modeling simple as it should be, but not more. “Simple as it should be” may turn out eventually to be fairly or very complex, but even if so, I believe EAA studies is one of our best bets in this quest.

In summary, many avenues for further development of EAA studies await further exploration in the near future. But the combination of the results reviewed here, the clear challenges awaiting in the near future and the continuing fast growth of computing resources that open new ways for more realistic EAA modeling, make us confident that the study of EAAs as a neuroscience research methodology is a promising and timely endeavor.

Acknowledgments: Supported by the FIRST grant of the Israeli Academy of Sciences. We are grateful to Ranit Aharonov, Tuvia Beker, Naomi Gal-Ruppin, David Horn, Isaac Meilijson, Yael Niv, Yoram Oron, James A. Reggia and Zach Solan for careful reading of this manuscript and many helpful comments.

References

[1] M. Mitchell. *An Introduction to Genetic Algorithms.* MIT Press, Cambridge, Massachusetts, 1996.

[2] C. Langton. *Artificial Life: An Introduction.* MIT Press, Boston, MA, 1995.

[3] D. B. Fogel. *Evolutionary Computation - Toward a New Philosophy of Machine Intelligence.* IEEE Press, Piscataway, NJ, 1995.

[4] C. Adami. *Introduction To Artificial Life.* Springer-Verlag, New York, NY, 1998.

[5] R. Aharonov-Barki, T. Beker, and E. Ruppin. Emergence of memory-driven command neurons in evolved artificial agents. *Neural Computation*, (13):691–716, 2001.

[6] J-A. Meyer and A. Guillot. From SAB90 to SAB94: Four years of animat research. In D. Cliff, P. Husbands, J-A. Meyer, and S.K. Wilson, editors, *Proceedings of the Third International Conference on Simulation of Adaptive Behavior.* MIT Press, Cambridge, MA, 1994.

[7] J. Kodjabachian and J.A. Meyer. Evolution and development of neural controllers for locomotion, gradient-following and obstacle-avoidance in artificial insects. *IEEE Transactions on Neural Networks*, 9(5):796–812, 1998.

[8] X. Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, 1999.

[9] A. Guillot and J-A. Meyer. From SAB94 to SAB2000: What's new, animat? In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior.* MIT Press, Cambridge, MA, 2000.

[10] A. Guillot and J.A. Meyer. The animat contribution to cognitive systems research. *Journal of Cognitive Systems Research*, (2):157–165, 2001.

[11] A. Cangelosi and D. Parisi. A neural network model of Caenorhabditis Elegans: The circuit of touch sensitivity. *Neural Processing Letters*, (6):91–98, 1997.

[12] A.J. Ijspeert and J. Hallam amd D. Willshaw. Evolving swimming controllers for a simulated lamprey with inspiration from neurobiology. *Adaptive Bahvior*, 7:151–172, 1999.

[13] O. Ekeberg. A combined neuronal and mechanical model of fish swimming. *Biological Cybernetics*, 69:363–374, 1993.

[14] R. Jung, T. Kimmel, and A.H. Cohen. Dynamical behavior of a neural network model of locomotor control in the lamprey. *J. of Neurophysiology*, 75:1074–1086, 1996.

[15] Donald H. Edwards, William J. Heitler, and Franklin B. Krasne. Fifty years of command neuron: the neurobiology of escape behavior in the crayfish. *Trends in Neuroscience*, 22(4):153–161, 1999.

[16] Y. Xin, K.R. Weiss, and I. Kupfermann. A pair of identified interneurons in Aplysia that are involved in multiple behaviors are necessary and sufficient for the arterial-shortening component of a local withdrawal reflex. *Journal of Neuroscience*, 16(14):4518–4528, 1996.

[17] T. Nagahama, K. Weiss, and I. Kupfermann. Body postural muscles active during food arousal in Aplysia are modulated by diverse neurons that receive monosynaptic excitation from the neuron CPR. *Journal of Neurophysiology*, 72(1):314–25, 1994.

[18] T. Teyke, K. Weiss, and I. Kupfermann. An identified neuron (CPR) evokes neuronal responses reflecting food arousal in Aplysia. *Science*, (247):85–87, 1990.

[19] Y.V. Panchin, Y.I. Arshavsky, T.G. Deliagina, G.N. Orlovsky, L.B. Popova, and A.I. Selverston. Control of locomotion in the marine mollusc Clione limacina. XI. Effects of serotonin. *Experimental Brain Research*, 109(2):361–365, 1996.

[20] B.J. Norris, M.J. Coleman, and M.P. Nusbaum. Recruitment of a projection neuron determines gastric mill motor pattern selection in the stomatogastric nervous system of the crab, Cancer borealis. *Journal of Neurophysiology*, 72(4):1451–1463, 1994.

[21] R. A. DiCaprio. An interneuron mediating motor programme switching in the ventilatory system of the crab. *Journal of Experimental Biology*, 154:517–535, 1990.

[22] D. Combes, P. Meyrand, and J. Simmers. Motor pattern specification by dual descending pathways to a lobster rythm-generating network. *Journal of Neuroscience*, 19(9):3610–3619, 1999.

[23] M. Thoby Brisson and J. Simmers. Neuromodulatory inputs maintain expression of a lobster motor pattern generating network in a modulation-dependent state: evidence from long-term decentralization in vitro. *Nournal of Neuroscience*, 18(6):2212–2225, 1998.

[24] P. Husbands, T. Smith, N. Jacobi, and M. Oshea. Better living through chemestry: Evolving GasNets for robot control. *Connection Science*, 10:185–210, 1998.

[25] A. Ishiguro, K. Otsu, A. Fujii, Y. Uchikawa, T. Aoki, and P. Eggenberger. Evolving an adaptive controller for a legged-robot with dynamically-rearranging neural networks. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, MA, 2000.

[26] R.D. Beer, H.J Chiel, and J.C. Gallagher. Evolution and analysis of model CPGs for walking II. general principles and individual variability. *Journal of Computational Neuroscience*, (7):119–147, 1999.

[27] C. Scheier, R. Pfeifer, and Y. Kunyioshi. Embedded neural networks: Exploiting constraints. *Neural Networks*, (7-8):1551–1569, 1998.

[28] R.D. Beer. Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, (4):91–99, 2000.

[29] S. Nolfi and D. Floreano. Learning and evolution. *Autonomous Robots*, (7):89–113, 1999.

[30] G.E. Hinton and S. Nowlan. How learning can guide evolution. *Complex Systems*, (1):495–502, 1987.

[31] G.F. Miller and P. Todd. Exploring adaptive agency: I. theory and methods for simulating the evolution of learning. In D.S. Touretzky, J.L. Elman, T.J. Sejnowski, and G.E. Hinton, editors, *Proc. of the 1990 Connenctionist models summer school*. Morgan Kaufmann, 1990.

[32] S. Nolfi and D. Parisi. Learning to adapt to changing environments in evolving neural networks. *Adaptive Behavior*, (5):75–98, 1997.

[33] P.S. Churchland and T.J. Sejnowski. *The computational brain*. MIT Press, Boston, MA, 1989.

[34] Y. Niv, D. Joel, I. Meilijson, and E. Ruppin. Evolution of reinforcement learning in foraging bees in neural terms. In *Tenth Annual Computational Neuroscience Meeting (CNS2001)*. 2001.

[35] P.R. Montague, P. Dayan, C. Person, and T.J. Sejnowski. Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, (377):725–728, 1995.

[36] S. Schacher, F. Wu, and Z.-Y. Sun. Pathway-specific synaptic plasticity: activity-dependent enhancement and suppression of long-term heterosynaptic facilitation at converging inputs on a single target. *The Journal of Neuroscience*, 17(2):597–606, 1997.

[37] K.E. Vogt and R.E. Nicoll. Glutamate and Gama-Amino Butyric Acid mediate a heterosynaptic depression at mossy fiber synapses in the Hippocampus. *Proceedings of the National Academy of Science, USA*, 96:1118–1122, 1999.

[38] R. Fisken, L. Garey, and T. Powell. Patterns of degeneration after intrinsic lesions of the visual cortex of the monkey. *Brain Res*, 53:208–213, 1973.

[39] R. Hess, K. Negishi, and O. Creutzfeldt. The horizontal spread of intracortical inhibition in visual cortex. *Exp Brain Res*, 22:415–419, 1975.

[40] R. Douglas and K. Martin. Neocortex. In *The Synaptic Organization of the Brain*, pages 389–438. Oxford University Press, 1997.

[41] J. Reggia, L. Autrechy, G. Sutton, and M. Weinrich. A competitive distribution theory of neocortical dynamics. *Neural Computation*, 4:287–317, 1992.

[42] V. Mountcastle. *Perceptual Neuroscience: Cerebral Cortex*. Harvard Univ Press, 1998.

[43] D. Ayers and J.A. Reggia. Evolving columnar circuitry for lateral cortical inhibition. In *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks*, pages 278–283. July 2001.

[44] E.T. Rolls and S.M. Stringer. On the design of neural networks in the brain by genetic algorithms. *Progress in Neurobiology*, (61):557–579, 2000.

[45] F. Dellaert and R.D. Beer. Toward an evolvable model of development for autonomous agent synthesis. In R. Brooks and P. Maes, editors, *Proceedings of the Fourth Conference on Artificial Life*. MIT Press, Cambridge, MA, 1994.

[46] H. Kitano. Designing neural networks using genetic algorithms with graph generation system. *Complex System*, 4:461–476, 1990.

[47] F. Gruau. Automatic definition of modular neural networks. *Adaptive behavior*, 3:151–183, 1994.

[48] J. Kodjabachian and J-A. Meyer. Evolution and development of modular control architectures for 1-D locompotion in six-legged animats. *Connection Science*, 10:211–254, 1998.

[49] B-T. Zhang and H. Muhlenbein. Evolving optimal neural networks using genetic algorithms with Occam's Razor. *Complex Systems*, 7:199–220, 1993.

[50] R.K. Belew. Interposing an ontogenetic model between genetic algorithms and neural networks. In J. Cowan, editor, *Advances in Neural Information Processing (NIPS5)*. Morgan Kaufmann, San Mateo, CA, 1993.

[51] A. Cangelosi, D. Parisi, and S. Nolfi. Cell division and migration in a 'genotype' for neural networks. *Network*, (5):497–515, 1994.

[52] S. Nolfi and D. Parisi. Evolving artificial neural networks that develop in time. *Advances in Artificial Life: Lecture Notes in Artificial Intelligence*, (929):353–367, 1995.

[53] A. Cangelosi and J.L. Elman. Gene regulation and biological development in neural networks: An exploratory model. *Technical report, CRL-UCSD, University of California at San Diego, www.citeseer.nj.nec.com/context/15377/132530*, 1995.

[54] P. Eggenberger. Cell interactions as a control tool of developmental processes for evolutionary robotics. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, MA, 1996.

[55] D. Floreano and J. Urzelai. Evolutionary robots with online self-organization and behavioral fitness. *Neural Networks*, (13):431–443, 2000.

[56] D. Floreano and J. Urzelai. Neural morphogenesis, synaptic plasticity and evolution. *Theory in Biosciences*, 2001.

[57] A.G. Rust, R. Adams, and S. George H. Bolouri. Activity-based pruning in developmental artificial neural networks. In P. Husbands and I. Harvey, editors, *Proceedings of the 4th European Conference on Artificial Life (ECAL 97)*. MIT Press, Cambridge, MA, 1997.

[58] S. Bushy and E. Ruppin. A self-organizing compressed encoding of evolutionary autonomous agents. *Preprint, www.math.tau.ac.il/ ruppin*, School of Computer Sciences, Tel Aviv University 2002.

[59] D. Cliff and G.F. Miller. Co-evolution of pursuit and evasion II: Simulation Methods and Results. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, MA, 1996.

[60] H.J. Chiel, R.D. Beer, and J.C. Gallagher. Evolution and analysis of model CPGs for walking I. Dynamical modules. *Journal of Computational Neuroscience*, (7):99–118, 1999.

[61] P. Getting. Emerging principles governing the operations of neural networks. *Ann. Rev. Neurosci.*, (12):185–204, 1989.

[62] E. Marder and R.L. Calabrese. Principles of rythmic motor pattern generation. *Physiol. Rev.*, (76):687–717, 1996.

[63] E. Marder and L.F. Abbott. Theory in motion. *Curr. Opinion Neurobiol.*, (5):832–840, 1995.

[64] J.P.C. Dumont and R.M. Robertson. Neuronal circuits: An evolutionary perspective. *Science*, (233):849–853, 1986.

[65] D. Floreano and F. Mondada. Evolution of homing navigation in a real mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 26(3):396–407, 1996.

[66] I. Harvey, P. Husbands, and D. Cliff. Seeing the light: Artificial evolution, real vision. In D. Cliff, P. Husbands, J.A. Meyer, and S. Wilson, editors, *From Animals to Animats 3, Proc. of 3rd Intl. Conf. on Simulation of Adaptive Behavior, SAB94*. MIT Press/Bradford Books, 1994.

[67] K.O. Stanley and R. Miikkulainen. Evolving neural networks through augmenting topologies. *Technical Report AI01-290*, Department of Computer Sciences, University of Texas at Austin 2001.

[68] R. Aharonov, I. Meilijson, and E. Ruppin. Who does what to whom: A quantitative approach. In *IEEE Conference on Neural Information Processing Systems (NIPS 2000)*. 2000.

[69] J. Wu, L. B. Cohen, and C. X. Falk. Neuronal activity during different behaviors in aplysia: A distributed organization? *Science*, 263:820–822, 1994.

[70] S. Thorpe. Localized versus distributed representations. In M. A. Arbib, editor, *Handbook of Brain Theory and Neural Networks*. MIT Press, Massachusetts, 1995.

[71] K. S. Lashley. *Brain Mechanisms in Intelligence*. University of Chicago Press, Chicago, 1929.

[72] S. G. Lomber. The advantages and limitations of permanent or reversible deactivation techniques in the assesment of neural function. *J. of Neuroscience Methods*, 86:109–117, 1999.

[73] V. Walsh and A. Cowey. Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience*, (1):73–79, 2000.

[74] K.J. Friston, C.D. Frith, and R.S.J. Frackowiak. Time-dependent changes in effective connectivity measured with PET. *Human Brain Imaging*, (1):69–79, 1993.

[75] V. Braitenberg. *Vehicles, Experiments in Synthetic Psychology*. MIT Press, Cambridge MA, 1984.

[76] D. Marr. *Vision*. W. H. Freeman, New York, 1982.

[77] R.A. Brooks. Intelligence without representations. *Artificial Intelligence*, (47):139–159, 1991.

[78] D. Cliff and S. Noble. Knowledge-based vision and simple visual machines. *Philosophical Transactions of the Royal Society of London: Series B*, (352):1165–1175, 1997.

[79] D. Cliff. Neuroethology, computational. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press / Bradford Books, 1995.

(a) (b)

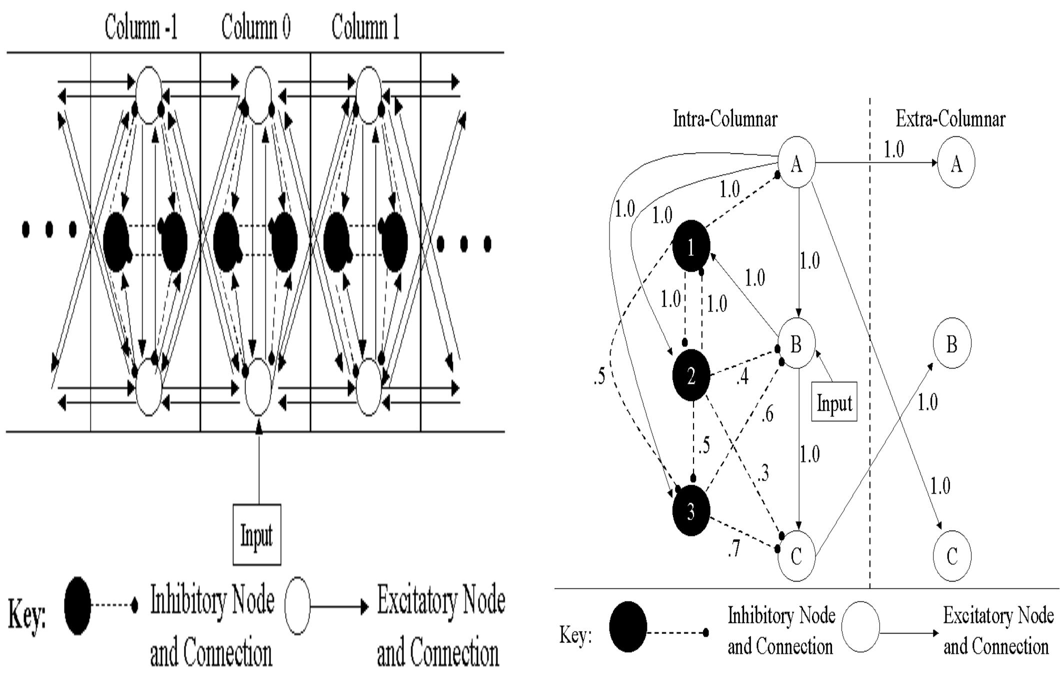

Figure 4: a) An example initial cortical circuit prior to evolution with two excitatory and two inhibitory nodes per column. Each node represents a small population of similar neurons. Initial synaptic strengths were random, and different evolutionary runs started with different numbers of excitatory/inhibitory nodes. b) Diagram of an evolved cortical circuit that produces a Mexican Hat pattern of activity to a point stimulus in the absence of horizontal, inter-columnar inhibitory connections. Only the magnitudes of inhibitory weights are shown (dotted connections); they are multiplied by a negative gain constant when used. Connections with weights less than 20% of the maximum were omitted. Connections for all columns are the same. Lateral inhibition arose because excitation of some excitatory neurons (nodes B and C) in a stimulated column caused other excitatory neurons (node A) in the same column (those sending lateral excitatory connections to other columns) to turn off, decreasing lateral excitation of adjacent columns and causing their mean activation levels to fall. Inhibitory neurons evolved to not only inhibit intra-columnar excitatory neurons, but to also inhibit each other in a highly specific, selective pattern.

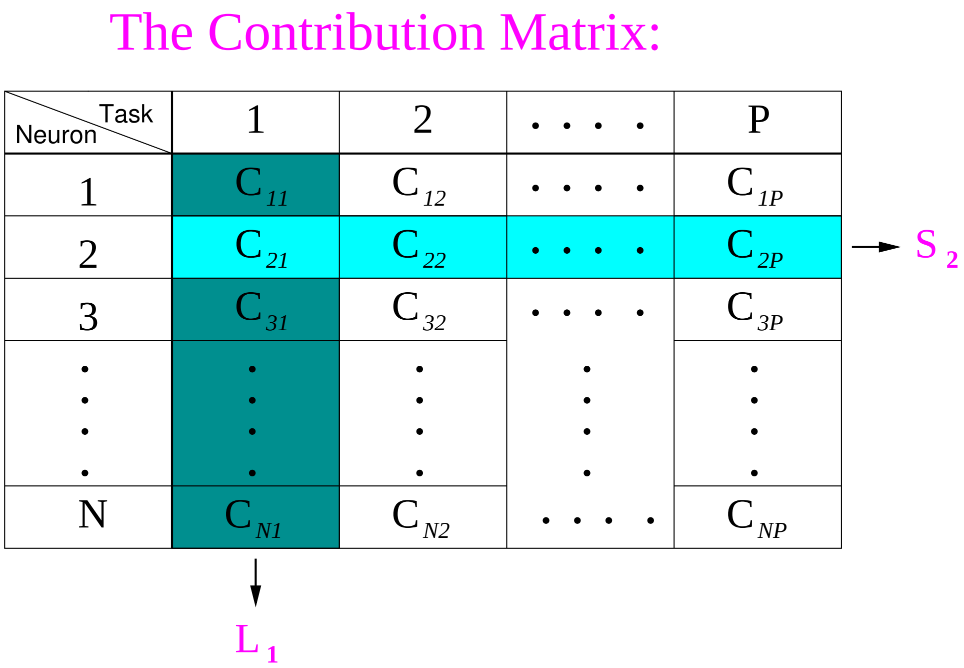

Figure 5: Contributions of Neurons (Units) to Tasks