PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL PHYSICAL & ENGINEERING SCIENCES

Controlled analysis of neurocontrollers with informational lesioning

Alon Keinan, Isaac Meilijson and Eytan Ruppin

Phil. Trans. R. Soc. Lond. A 2003 **361**, doi: 10.1098/rsta.2003.1253, published 15 October 2003

10.1098/rsta.2003.1253

THE ROYAL SOCIETY

Controlled analysis of neurocontrollers with informational lesioning

By Alon Keinan[1], Isaac Meilijson[2] and Eytan Ruppin[1,3]

[1]*School of Computer Sciences, and* [2]*School of Mathematical Sciences, and* [3]*School of Medicine, Tel-Aviv University, Tel-Aviv, Israel*
(keinan@cns.tau.ac.il; isaco@post.tau.ac.il; ruppin@post.tau.ac.il)

Published online 18 August 2003

How does one aim to understand neural information processing? One of the difficult first challenges is to identify the roles of the network's elements. To this end a functional contribution analysis (FCA) method has been developed and applied for studying the neurocontrollers of evolutionary autonomous agents (EAAs). The FCA processes data composed of multiple lesion experiments and the corresponding performance levels that the agent obtains under these lesions. It calculates the *contribution values* (CVs) of the network's elements such that the ability to predict the agent's performance under new, unseen lesions is maximized. Previous analysis has found a strong dependence of the CVs and the prediction error on the specific type of lesioning method used, i.e. on the way in which the activity of lesioned neurons is disrupted. We present a new, *informational lesioning method* (ILM), which views a lesion as a noisy channel and applies a *controlled lesion* to the network by varying the *lesioning level* from large to arbitrarily small magnitudes. Studying the ILM within the FCA framework, our main results are threefold: first, that lower lesioning levels permit more accurate FCA predictions; second, that the usage of minute ILM lesioning levels can uncover the long-term effects of elements on the network's functioning; and third, that as the lesioning level decreases, the CVs tend to approach limit values, reflecting the importance of these elements in the intact, normal-functioning neurocontroller.

Keywords: neurocontroller analysis; lesioning; localization of function; performance prediction; FCA; long-term dynamics

1. Introduction

Recent years have witnessed a growing interest in the study of neurally driven evolved autonomous agents (EAAs). An EAA is a software program embedded in a simulated virtual environment, performing typical animat tasks such as gathering food, navigating, evading predators and seeking prey and mating partners. An EAA is

One contribution of 16 to a Theme 'Biologically inspired robotics'.

controlled by an artificial neural network 'brain'. This neurocontroller receives and processes sensory inputs from the surrounding environment and governs the agent's behaviour via the activation of motors controlling its actions. It is developed via genetic algorithms (Yao 1999) that apply some of the essential ingredients of inheritance and selection to a population of agents that undergo evolution. Much progress has been made in finding ways to evolve autonomous agents that successfully cope with diverse behavioural tasks (Floreano & Mondada 1996, 1998; Gomez & Miikkulainen 1997; Kodjabachian & Meyer 1998; Marocco & Floreano 2002; Scheier *et al.* 1998). However, considerably less effort has been spent on the important question of understanding how the neurocontrollers perform the tasks. One of the difficult challenges in analysing an agent's neurocontroller is identifying the roles of the network's elements, and assessing their contributions to the different tasks.

Several studies have used a variety of conventional neuroscience techniques to analyse neurocontrollers of EAAs. In Floreano & Mondada (1996), the activity of internal neurocontroller neurons as a function of a robot's location and orientation was charted by a simple form of receptive-field measurement. Neuronal functioning was generally highly distributed, but a specific interneuron that had an important role in path planning was also identified. Other researchers have studied the effects of clamping neuronal activity on the robot's behaviour (for example, inducing rotation, straight-line motion or more complex behaviours such as smooth tracking of moving targets (Harvey *et al.* 1994)). The 'command' neurons described in Aharonov-Barki *et al.* (2001) were discovered by studying the agent's behaviour following single lesions and by receptive field analysis. Finally, a more 'procedural' kind of ablation, in which different processes (and not just units or links) are systematically cancelled out, was employed by Stanley & Miikkulainen (2001). Overall, these studies have provided only glimpses of the neural processing that takes place in EAAs' neurocontrollers.

Aharonov *et al.* (2003) have presented a rigorous and quantitative method for localizing functional tasks in neurocontrollers. A definition of the elements' contributions to the performance has been presented, and the functional contribution analysis (FCA) framework was developed to measure them. The FCA framework assumes an existing dataset composed of numerous multiple lesions that have been inflicted on the network along with the agent's performance scores resulting from each such experiment. In each multiple lesion experiment several elements of the agent's neurocontroller are lesioned by concurrently disrupting their normal mode of operation. The FCA algorithm uses these data to calculate the elements' contributions, with the aim of yielding an accurate prediction of the performance of the agent when a new, untested, multiple lesion damage is inflicted upon it. The current paper adopts the FCA framework and studies a new, intriguing and challenging issue: *the role played by the method of lesioning*. As demonstrated in this paper, the lesioning method itself is a critical factor in determining the outcome of the FCA.

Previous FCA studies (Aharonov *et al.* 2003; Segev *et al.* 2002) have employed *stochastic lesioning*, in which lesioning of a neuron is performed by randomizing its firing pattern rather than by completely silencing its output, the latter lesioning alternative termed *biological lesioning*. In stochastic lesioning, at every network update a lesioned neuron fires with a probability equal to its overall mean firing rate, independent of its input field. This type of lesioning maintains the neuron's firing rate but disrupts the precise firing pattern it transmits to other neurons in its intact

state. It has been shown (Aharonov *et al.* 2003) that the FCA algorithm predicts the performance much more accurately when using stochastic lesioning in comparison with biological lesioning. Furthermore, the elements' contributions assigned by the FCA differ significantly when the two distinct lesioning methods are used. These results indicate that the lesioning method employed by the FCA is very important.

Considering these previous results, we hypothesize that FCA using stochastic lesioning outperforms that using biological lesioning because biological lesioning causes larger perturbations to the neurocontroller firing patterns: when using biological lesioning the effects of the lesions propagate strongly to other network elements. Thus, the effective part of the network that is damaged is much larger than that originally lesioned (Young *et al.* 2000). Stochastic lesioning also spreads the damage from lesioned neurons to other neurons, but to a much lesser extent. This discrepancy between the originally induced and actual lesions is a likely cause of error, reducing the FCA prediction accuracy. Furthermore, the larger multiple lesions occurring effectively with biological lesioning result in a stronger deviation of the network from its normal mode of activity and operation, our main region of interest. This is a manifestation of *the inherent paradox of the lesioning methodology*, where we want to learn about the network's *normal behaviour* by observing its activity when it is *perturbed*. Consequently, it is logical to conclude that lesioning perturbations should be *as small as possible* to reveal the network's normal working.

The challenge of this work is to develop a new lesioning method that will permit the FCA study of neurocontrollers without incurring the problem imposed by the large perturbations discussed above. To this end, we present a new *informational lesioning method (ILM)*. This method, in contrast with both stochastic lesioning and biological lesioning, keeps the output of a lesioned element *stochastically dependent* on its input field. As a consequence, the method allows the application of *controlled lesions* to the network, *keeping the lesion damage arbitrarily small*. The damage may be applied with varying lesioning probabilities, where the most extreme disruption is stochastic lesioning. Thus, when applying small ILM damage to a number of elements in a multiple lesion configuration, the neurocontroller continues to function in a relatively normal fashion. The smaller the amount of damage, the closer the lesioned neurocontroller to its normal functioning and the better the elements' contributions capture its intact state.

The specific goals that are addressed in this study are threefold: first, to study whether using ILM can lead to *a more accurate FCA*; second, to examine if applying subtle lesions can reveal the *long-term dynamics* of the neurocontroller (in both biological lesioning and stochastic lesioning the lesioning effect is too extensive and eliminates the possibility of measuring such long-term effects); third, in light of the motivation of introducing small perturbations to learn about normal behaviour, to test whether, as the ILM damage is made more minute, the elements' contributions approach a set of limit values, that is, whether the FCA with ILM lesioning reveals a set of contribution values (CVs) at the limit of normal functioning of the neurocontroller.

The rest of this paper is organized as follows: § 2 presents the EAA model used and reviews the basic FCA; § 3 presents the ILM; and § 4 describes and analyses the results of applying the ILM-based FCA to the localization of function in EAAs' neurocontrollers. Our results and their implications are discussed in § 5.

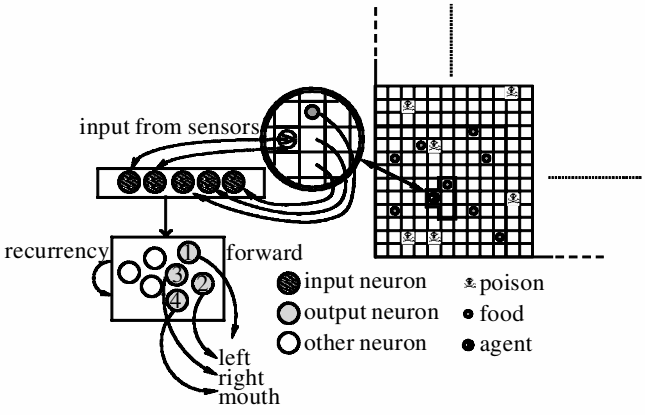

Figure 1. The EAA environment. An outline of the grid world and the agent's neurocontroller. The agent is marked by a small arrow on the grid, whose direction indicates its orientation. The curved lines indicate where in the arena each of the sensory inputs comes from. (Reproduced with permission after Aharonov-Barki *et al.* (2001).)

2. The evolutionary environment and the FCA

(*a*) *The EAA environment*

The EAA environment is described in detail in Aharonov-Barki *et al.* (2001). A brief overview is provided herein. The EAAs analysed here live in a discrete two-dimensional grid 'world' surrounded by walls. Poison items are scattered all around the world, while food items are scattered only in a 'food zone' in one corner. The agent's goal is to find and eat as many food items as possible during its life, while avoiding the poison items. The fitness of the agent is proportional to the number of food items minus the number of poison items it consumes. The agent is equipped with a set of sensors, motors and a fully recurrent neurocontroller.

Four sensors encode the presence of a resource (food or poison, without distinction between the two), a wall or a vacancy in the cell the agent occupies and in the three cells directly in front of it (see figure 1). A fifth sensor is a 'smell' sensor, which can differentiate between food and poison underneath the agent, but gives a random reading if the agent is in an empty cell. The four motor neurons dictate movement forward (neuron 1), a turn left (neuron 2) or right (neuron 3), and control the state of the mouth (open or closed, neuron 4). Importantly, eating only takes place if the agent is neither moving nor turning. Thus, eating is a costly process requiring a time-step with no other movement, in a lifetime of limited time-steps.

The neurocontroller is composed of binary McCulloch–Pitts neurons, with fully recurrent connections. Network updating is synchronous: in every 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 field of neuron i at time t is defined by

$$f_i(t) = \sum_{j=1}^{N} s_j(t-1)W(i,j), \tag{2.1}$$

where $s_j(t-1)$ is the state of neuron j at time $t-1$, $W(i,j)$ is the synaptic weight from neuron j to neuron i and N is the number of neurons, including the input

sensory neurons. The neuron fires ($s_i(t) = 1$) if its field exceeds a threshold, and it remains silent ($s_i(t) = 0$) otherwise. The synaptic weights are evolved using a genetic algorithm (Yao 1999).

Previous analysis (Aharonov-Barki *et al.* 2001) revealed that successful agents possess one or more *command neurons* that determine the agent's behavioural strategy. Artificially clamping these command neurons to either constant firing activity or to complete quiescence causes the agent to constantly maintain one of the two behavioural modes it exhibits, regardless of its sensory input. These two behavioural modes are *exploration* and *grazing*. Exploration, which takes place when the agent is outside of the food zone, consists of moving in straight lines, ignoring resources in the sensory field that are not directly under or in front of the agent and turning at walls. Grazing, which takes place when the agent is in the food zone, consists of turning towards resources to examine them, turning at walls and maintaining the agent's location on the grid in a relatively small region.

Throughout this paper, we mainly focus on the analysis of one of the successfully evolved agents, S10, with a neurocontroller consisting of 10 neurons. This agent achieves a fitness score which is above the performance levels obtained by several manually designed algorithms, as well as that obtained through reinforcement learning (Aharonov-Barki *et al.* 2001). This agent has one command neuron that determines its behavioural mode.

(*b*) *The FCA*

The FCA algorithm is described in detail in Aharonov *et al.* (2003). A brief overview is provided here. The FCA starts by measuring the performance (fitness) of the agent over many lesioning experiments. In each such experiment, a different lesioning configuration is inflicted upon the agent's neurocontroller. Each configuration specifies which of the agent's neurocontroller elements (which may be synapses, neurons or any higher-order modules) are lesioned. The FCA algorithm is designed to use these data in order to search for a *contribution vector* $\boldsymbol{c} = (c_1, \ldots, c_N)$, where c_i is the CV of element i to the task in question and N is the number of elements in the network. The goal of the FCA is to find a contribution vector that provides the best prediction of the agent's performance in terms of mean-squared error (MSE), under all possible multiple site lesions (including new, unseen ones).

More formally, a lesioning configuration is denoted by a vector $\boldsymbol{m}$, where $m_i = 0$ if the element is lesioned, and $m_i = 1$ if it is intact. The prediction of performance in this lesioned state is based on a linear model generalized by a nonlinear transformation. Given a contribution vector $\boldsymbol{c}$ and a non-decreasing function f, the predicted performance $\tilde{p}_{\boldsymbol{m}}$ when a lesion $\boldsymbol{m}$ is applied to the network is given by

$$\tilde{p}_{\boldsymbol{m}} = f(\boldsymbol{m} \cdot \boldsymbol{c}). \tag{2.2}$$

Denoting the actual performance of the agent under lesioning configuration $\boldsymbol{m}$ by $p_{\boldsymbol{m}}$, the mean-squared prediction error is

$$\mathrm{MSE} = \frac{1}{2^N} \sum_{\{\boldsymbol{m}\}} (\tilde{p}_{\boldsymbol{m}} - p_{\boldsymbol{m}})^2, \tag{2.3}$$

where the summation runs theoretically over all lesion configurations. A vector $\boldsymbol{c}$ which minimizes this error is a contribution vector for the task tested, and the cor-

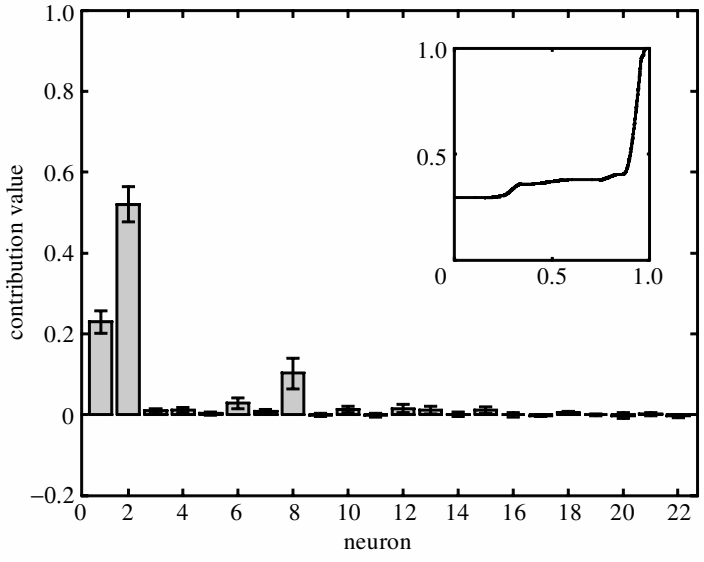

Figure 2. FCA of agent S22. The CVs c_i of S22's neurons computed by the FCA are plotted (mean and standard deviation across 10 FCA runs). The inset depicts a performance prediction function, where the x-axis is $\boldsymbol{m} \cdot \boldsymbol{c}$ and the y-axis is the predicted performance $f(\boldsymbol{m} \cdot \boldsymbol{c})$. The performance prediction functions in all 10 runs are very similar. Thus, only one is depicted for clarity of exposition.

responding f is its adjoint *performance prediction function*. The performance prediction function f is a non-decreasing piecewise polynomial. It is non-decreasing to reflect the notion that beneficial elements (those whose lesioning results in performance deterioration) should have positive CVs, and that negative values indicate elements that hinder performance. Since multiplying $\boldsymbol{c}$ and scaling f accordingly maintains the prediction, we arbitrarily normalize $\boldsymbol{c}$ such that $\sum_{i=1}^{N} |c_i| = 1$.

In practice, the goal of the FCA algorithm is to find a contribution vector $\boldsymbol{c}$ and a performance prediction function f which minimize equation (2.3) given a subset of the full 2^N configurations set. The optimal $\boldsymbol{c}$ and f are determined using a training set of lesioning configurations $\boldsymbol{m}$ and the accompanying performance levels $p_{\boldsymbol{m}}$. The FCA algorithm works as follows.

(i) **Initialize** $\boldsymbol{c}$ by selecting each element c_i randomly in the range $[-1, 1]$. Normalize $\boldsymbol{c}$ such that $\sum_{i=1}^{N} |c_i| = 1$. Compute f as in step (iii).

(ii) **Compute** $\boldsymbol{c}$ by gradient descent to minimize equation (2.3) while keeping f fixed. Normalize $\boldsymbol{c}$.

(iii) **Compute** f to minimize equation (2.3) while keeping $\boldsymbol{c}$ fixed, by performing an isotone regression on the pairs $\{\boldsymbol{m} \cdot \boldsymbol{c}, p_{\boldsymbol{m}}\}$, and smoothing the result with a cubic spline.

Steps (ii) and (iii) are repeated until convergence or for a fixed number of iterations.

To illustrate the FCA's workings, figure 2 (after Aharonov *et al.* (2003)) shows the results of applying the FCA for the analysis of S22, an agent with a neurocontroller consisting of 22 neurons. The FCA is applied to the neurocontroller by

performing random lesioning experiments (using stochastic lesioning) on the neurons, while measuring the performance levels of the agent. The result of applying the FCA to this training set is a contribution vector $\boldsymbol{c}$ and a performance prediction function f. The FCA succeeds in finding a pair $\{\boldsymbol{c}, f\}$ that brings the error in equation (2.3) to a very low value even when trained on a very limited subset of the 2^{22} configuration space.

3. The informational lesioning method

(*a*) *Overview*

As discussed in § 1, previous analysis (Aharonov *et al.* 2003) has suggested that the method of lesioning has a considerable effect on the FCA. Using biological lesioning results in a much larger prediction error than that obtained using stochastic lesioning. The lesioning method also has a significant effect on the CVs found. To study the effect of the lesion magnitude on localization of function in more depth, we have devised a method for applying controlled lesioning, allowing us to modulate the amount of perturbation from the neurocontroller's intact firing state within a given lesioning configuration.

The concept of informational lesioning is based on viewing each lesioned element in a given lesion configuration as a noisy transmission channel, where the channel inaccuracy (noise) determines the magnitude of the lesion. It is convenient to think first about the elements as neurons, but we later extend the analysis to synaptic elements. The input to the channel representing lesioning of a given neuron is the activity state (0 or 1) the neuron would select given its input *if it had been intact*. The output of the channel is the actual firing (0 or 1) of that neuron in its lesioned state. The lesioning is controlled by setting the channel fidelity, i.e. the similarity of the actual firing of the lesioned neuron to its firing in its intact state. We use *mutual information* (MI) as an information-theoretic measure of the channel-transmission fidelity.

In the rest of this section we introduce the notion of MI, formulate the ILM paradigm and present the experimental protocol for the analysis of EAA neurocontrollers using the ILM-based FCA.

(*b*) *A channel view of lesioning*

The information theory measures used here were first defined by Shannon (1948). Our notation and definitions follow Cover & Thomas (1991). We begin by introducing the concepts of entropy, conditional entropy, MI and channel. Let X and Y be discrete random variables with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and with probability mass functions:

$$\left.\begin{aligned} p(x) &= \Pr(X = x), \quad x \in \mathcal{X}, \\ q(y) &= \Pr(Y = y), \quad y \in \mathcal{Y}. \end{aligned}\right\} \tag{3.1}$$

Let $p(x, y)$ be their joint probability mass function. The *entropy* $H(X)$ of the discrete random variable X, which is a measure of uncertainty, is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{3.2}$$

The *conditional entropy* $H(Y \mid X)$ is defined as

$$H(Y \mid X) = \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x). \tag{3.3}$$

The MI $I(X;Y)$ of two discrete random variables, X and Y, is defined by

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)q(y)}, \tag{3.4}$$

which is symmetric with respect to X and Y. It can be shown that

$$I(X;Y) = H(X) - H(X \mid Y). \tag{3.5}$$

Thus, the MI $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y. More intuitively, the MI quantifies the amount of information Y gives on X, and vice versa. $I(X;Y) \geqslant 0$ and equality holds if and only if X and Y are independent. A *discrete channel* is a system consisting of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y \mid x)$ that expresses the probability of observing the output symbol y given that we send the symbol x.

In ILM, each lesioned neuron in a given lesion configuration is modelled as a discrete channel. The input to the channel is the original firing of that neuron (given its input field) in its intact state (X). The output of the channel determines the actual firing of the neuron under lesioning (Y). An ILM lesion configuration is simulated and implemented in the network as follows: at every network update, the intact firing state of all neurons is calculated as usual from their input fields. Then, the firing state of each lesioned neuron is passed through its lesioning channel to determine its actual firing. Based on the probabilities induced by the channel, the firing of a lesioned neuron may be *inverted* or left intact. The firing state of unlesioned neurons is always left intact.

As the neurocontrollers analysed in this paper are composed of binary McCulloch–Pitts neurons, the lesioning channel input and output are binary. There are two parameters to set in the channel, namely

$$p = \Pr(Y = 1 \mid X = 1) \quad \text{and} \quad q = \Pr(Y = 1 \mid X = 0).$$

p is the probability that the lesioned neuron fires when it indeed should fire and q is the probability of firing due to inversion (table 1). The marginal distribution of X is fixed, and may be estimated by observing the neuron across many lives of the agent. We will denote $\Pr(X = 1) = a$, which is the neuron's average firing rate.

Hence, a lesion channel can be characterized by the MI, $I(X;Y)$, between the intact (X) and lesioned (Y) outputs of a lesioned element. This MI value can serve to quantify the amount of lesioning employed. As $I(X;Y)$ varies in the range $[0, H(X)]$ (where $H(X) = -a \log a - (1-a)\log(1-a)$), we define the *lesioning level* λ to be the proportion of input entropy $H(X)$ lost by the channel. That is,

$$\lambda = 1 - \frac{I(X;Y)}{H(X)} = \frac{H(X \mid Y)}{H(X)}. \tag{3.6}$$

The same lesioning level λ results from different channels. The next subsection presents additional constraints under which the lesioning level determines a unique channel.

Table 1. *Illustration of the lesioning channel between the intact firing (input) and the lesioned firing (output) of an element*

(Each value in the channel denotes the probability of the corresponding output given the corresponding input.)

	lesioned	
intact	0	1
1	$1-p$	p
0	$1-q$	q

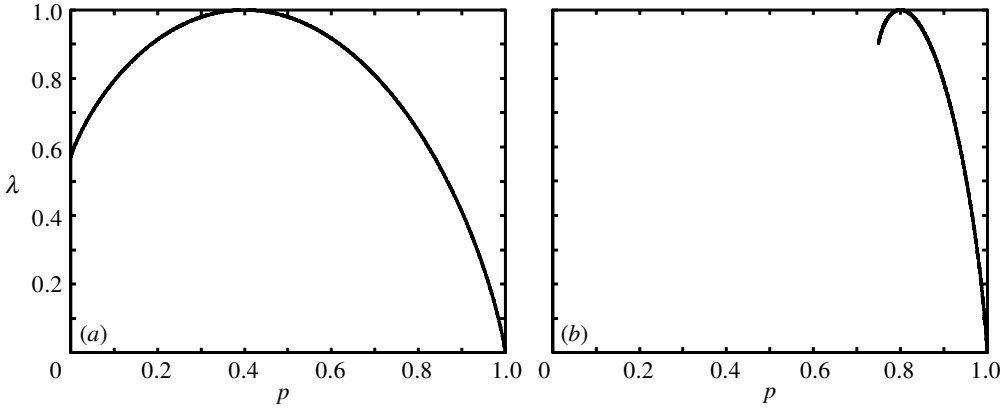

Figure 3. Constructing a lesioning channel. The lesioning level λ of the channel as a function of p, for average activity of (*a*) 0.4 and (*b*) 0.8. Only $p \geqslant 0.75$ is plotted for $a = 0.8$ since, in order to satisfy $q \leqslant 1$, the inequality $p \geqslant 2 - 1/a$ must hold. The region of interest is only $p \geqslant a$ (see main text).

Table 2. *Several lesioning channels calculated by the ILM for the indicated values of λ and a*

	$\lambda = 0.8$; $a = 0.3$		$\lambda = 0.5$; $a = 0.3$		$\lambda = 1$; $a = 0.8$		$\lambda = 0.2$; $a = 0.8$	
	0	1	0	1	0	1	0	1
1	0.348	0.652	0.159	0.841	0.2	0.8	0.014	0.986
0	0.851	0.149	0.932	0.068	0.2	0.8	0.945	0.055

(*c*) *Constructing a lesioning channel*

Let us construct a channel with the desired lesioning level λ, i.e. with a certain level of MI between the intact and lesioned outputs (3.6).

First note that the MI, like all information theory quantities, depends only on the distributions' probabilities, and not on the actual variable values. In the extreme case, a channel with $p = 0$ and $q = 1$ gives an optimal MI value, just like the case $p = 1$ and $q = 0$. While the latter channel denotes an intact neuron, a neuron lesioned using the former channel always fires when it should not and remains quiescent when

it should fire, reflecting the most extreme lesioning. In order to solve this ambiguity problem, the lesioning is bounded by stochastic lesioning as the most severe case ($p = q = a$) and the channel lesioning space is restricted by adding the constraint

$$p \geqslant a. \tag{3.7}$$

We further demand that the channel maintains the average activity of the intact neuron, i.e. $\Pr(Y = 1) = a$. This constraint is motivated by the need to inflict small and unbiased perturbations on the neurocontroller, following the previous findings that large perturbations result in fairly poor FCA prediction accuracy (Aharonov *et al.* 2003). Maintaining average activity provides the constraint†

$$a = (1 - a) \cdot q + a \cdot p. \tag{3.8}$$

Once a p value is assigned,

$$q = \frac{a \cdot (1 - p)}{1 - a}$$

is set according to this linear constraint. Thus, given a desired lesioning level λ, it is sufficient to find a p value yielding the desired λ, in the range $p \geqslant a$ (3.7), where λ is monotonously decreasing in p (figure 3): $p = a$ corresponds to stochastic lesioning, where $\lambda = 1$, since the lesioned output is independent of the intact output. For $p \geqslant a$ the lesioning level decreases in p, spanning all values of λ in the range $[0, 1]$, ensuring a unique solution for a desired λ, which is found numerically. Table 2 illustrates a number of such channels, calculated for the indicated λ and average activity a.

As a first test of the ILM, we measured the agent's performance when applying lesioning of level λ to *all its neurons*, for different λ values. Figure 4 shows that the higher the lesioning level the lower the agent's performance. The decrease is monotonous and very smooth. Thus, we verify that the introduced measure serves as a reliable index for lesioning levels.

(*d*) *The experimental protocol*

In the subsequent experiments presented in this paper, the FCA algorithm is used to analyse evolved neurocontrollers with ILM. That is, in each lesioning configuration (§ 2 *b*), the activity of those elements that are lesioned is disrupted by applying an ILM channel with a predetermined lesioning level. Given the desired λ value, the appropriate channel is calculated for each lesioned neuron of the neurocontroller (§ 3 *c*). Note that the actual channel parameters are different for different neurons since their average activity differs, but all channels maintain the same λ value. The performance $p_{\boldsymbol{m}}$ for each lesioning configuration $\boldsymbol{m}$ in the training and test sets is obtained by running the lesioned agent in its environment and measuring its fitness. Thus, the only difference between these *ILM-based FCA* experiments and the FCA employed in Aharonov *et al.* (2003) is the lesioning method.‡

† Note, however, that under lesioning the neurocontroller's firing patterns change. Hence, the designed channel may fail in achieving an activity level of a exactly. Moreover, the desired value of a, which is the activity of the intact element, may change due to the lesioning of other elements. Nevertheless, the smaller the perturbation to the neurocontroller, the closer the activity level is to a and the smaller the effect that a lesion has on the average activity of other elements.

‡ As in the FCA using stochastic lesioning, and also in the ILM-based FCA, when lesioning motor neurons we do not alter the activity transmitted to the motors themselves, in order to isolate the role of the motor neurons in the computation of the recurrent neurocontroller.

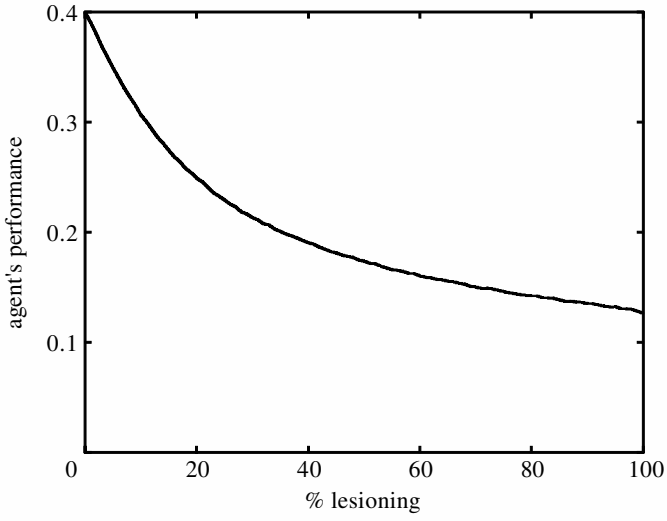

Figure 4. Agent's performance versus the percentage of lesioning applied. A uniform level of lesioning is applied to all the neurons of agent S10, introduced in § 2 *a*. The lesioning level in this experiment ranges between 0% ($\lambda = 0$), which implies no lesioning at all, and 100% ($\lambda = 1$), which corresponds to stochastic lesioning. The lesioning level is sampled every 1%.

A single FCA run consists of 10 trials, each executed until convergence after starting from a different random initial contribution vector choice, choosing the results of the trial which reaches the lowest training error. We note, however, that under all λ values, the sensitivity to initial conditions is very small. All FCA results presented below are mean and standard deviations of 10 FCA runs (each consisting of 10 trials). The test MSE results are normalized by the variance of the distribution of the agent's performance $p_{\boldsymbol{m}}$ in the test set. Thus, the normalized MSE equals $1 - R^2$, where R^2 is the explained fraction of the variance. That is, if one would predict for all configurations a performance which is equal to the mean performance value, then the normalized MSE would equal one ($R^2 = 0$). The smaller λ is, the smaller the perturbation and, therefore, the range of performance values obtained in different lesion configurations is reduced. Since both the mean and the variance of the distribution of the agent's performance values tend to covary with λ, this normalization of the MSE is crucial for comparing the FCA of different λ values. *It should be emphasized that this measurement of normalized error puts the task of predicting the performance under small lesion perturbations on an equal footing with the task of predicting the performance under large lesion perturbations, making it, at least in principle, as difficult.*

4. Neurocontroller analysis with ILM

(*a*) *FCA accuracy*

We first measure the performance scores of agent S10 described in § 2 *a* under the entire set of 2^{10} lesioning configurations. This calculation is repeated for 10 lesioning levels, $\lambda = 0.1, 0.2, \ldots, 1$. We then apply FCA to the dataset obtained at each lesioning level. In § 4 *c* we show that the ILM-based FCA can rely on a small training set and successfully generalize to a test set.

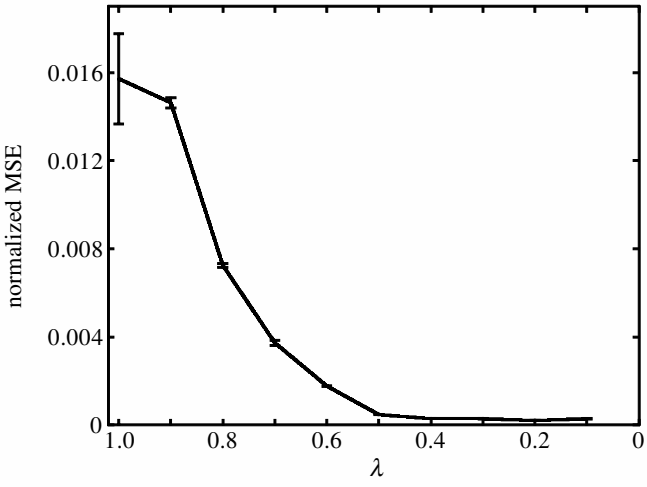

Figure 5. FCA accuracy with ILM. The mean and standard deviation of the MSE are plotted against the lesioning level λ. For each λ, the FCA is trained using the full set of 2^{10} lesioning configurations with that lesioning level. The λ values are plotted in decreasing order, reflecting our aim of starting at large lesioning levels and examining the results as λ is decreased gradually towards the zero limit. The vanishing standard deviation of small λ values is too small to be detectable.

Figure 5 displays the mean and standard deviation of the normalized MSE in these experiments as a function of λ (the results for $\lambda = 1$ coincide with the results using stochastic lesioning presented in Aharonov *et al.* (2003)). It is evident that the MSE monotonously decreases as λ decreases; the lesser the lesioning, the better the ILM-based FCA is capable of describing the neurocontroller. Furthermore, the standard deviation of the MSE among several FCA runs is much smaller for low λ. This testifies that the convergence of the FCA is more stable for smaller perturbations. The accuracy of the description obtained with smaller lesioning perturbations may seem straightforward, since smaller lesions obviously result in a narrower range of performance values relative to larger perturbations. But note that since accuracy is measured by the *normalized* MSE, this possibility is unlikely (as explained in § 3 *d*). Rather, the larger accuracy observed with smaller ILM perturbations reflects the inherently 'well-behaved' dynamics of the system when it lies close to its normal operating mode.

(*b*) *The contributed values*

The previous subsection shows that the ILM-based FCA accuracy is improved as the lesioning level λ is decreased. This subsection examines the CVs obtained in these experiments across different lesioning levels, to see whether the CVs approach limiting values as λ is decreased and to gain further understanding of the network processing.

The CVs of the neurons of S10 are shown in figure 6*a*. The identity of the significant neurons is consistent for different lesioning levels; for all λ, neurons 1, 2, 3, 5 and 10 have significant CVs while neurons 4, 6, 7, 8 and 9 all have near-vanishing CVs. Figure 6*b–d* zooms in on some of these CVs (neurons 1, 2 and 5) to examine in more detail how they vary in a broad range of lesioning levels. The small standard

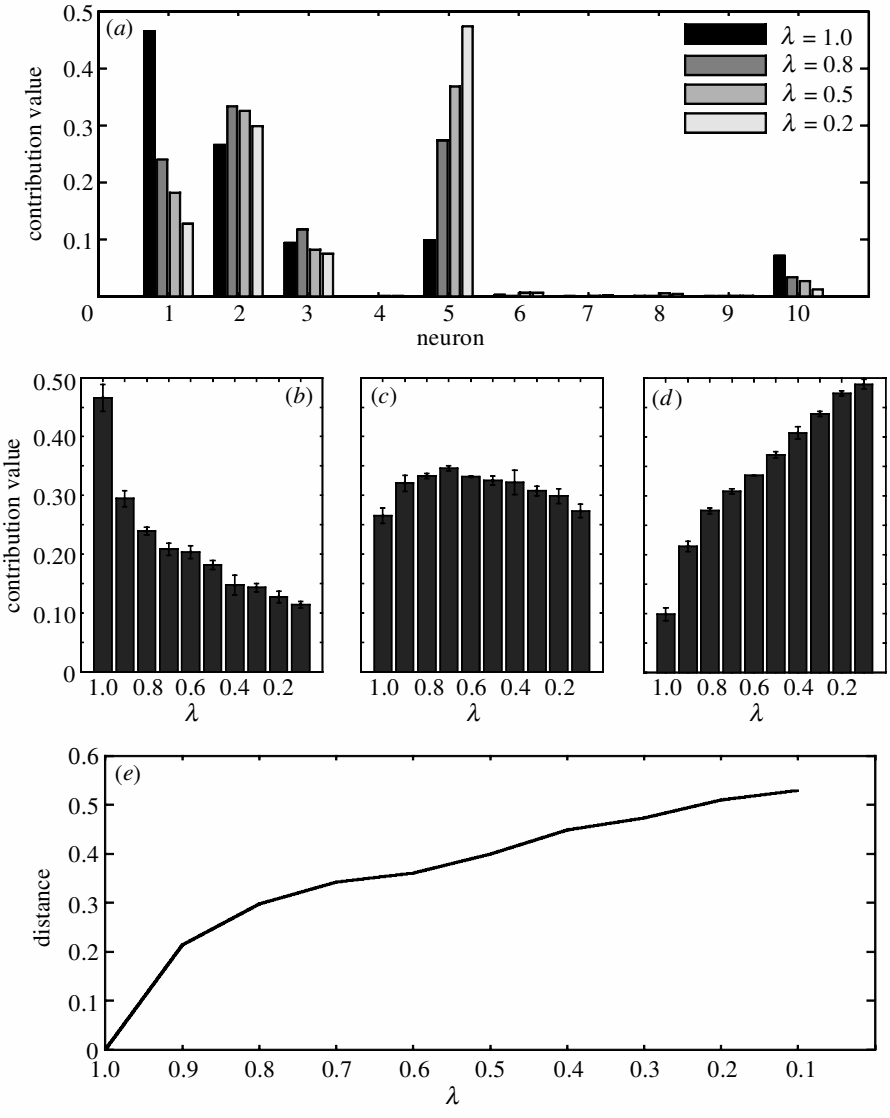

Figure 6. CVs of the neurons of agent S10. (*a*) CVs of the 10 neurons, plotted for several values of lesioning level λ. (*b*)–(*d*) Mean and standard deviation of the CVs of (*b*) neuron 1, (*c*) neuron 2 and (*d*) neuron 5 for all lesioning levels studied. (*e*) The distance of the 10-dimensional contribution vector obtained with lesioning level λ from the contribution vector of lesioning level $\lambda = \mathbf{1}$ (stochastic lesioning).

deviation testifies to the consistency of the algorithm: for each lesioning level, the FCA converges to similar CVs in all runs, starting from random CVs.

Comparing the CVs on a finer scale, they vary for different lesioning levels (figure 6*a*). In particular, the CVs of neurons 1, 5 and 10 vary monotonously in λ (decreasing for 5; increasing for 1 and 10). The change in the CV of a neuron as the lesioning level is decreased is explained as follows: the smaller the value of λ, the more rare the occurrences of inversions of the firing of the lesioned neuron from its

intact firing state. Thus, its CV might be expected to drop for lower λ. However, the CVs are normalized to a sum of absolute values of 1, which may lead one to expect that as long as the lesioning level is equal for all neurons (which is the case in these experiments), the CVs will remain similar in all lesioning levels. This is not the case, however, *since a rare inversion has a different effect on different neurons in the network*: if a neuron fires during a certain network update when it should not (or, vice versa, does not fire when it should), this event may affect other neurons, starting in the next network update. Depending on the evolved network's dynamics, the effects of this inversion may continue to propagate and affect the firing of many other neurons in subsequent network updates, in principle continuing even until the end of the agent's life. When these effects hinder the agent's performance, the longer they last, the larger that neuron's CV will become relative to the other neurons as λ decreases. Since the contribution vector is normalized, this relative increase is transcribed to an absolute increase in the CV. The shorter this effect, the lower the neuron's CV is, due to the normalization.

The two largest monotonous variations in the CVs are observed for neurons 1 and 5. Previous receptive field analysis of the agent revealed neuron 5 to be a command neuron modulating the agent's behaviour (Aharonov-Barki *et al*. 2001). When this neuron is active the agent manifests exploratory behaviour, whereas when it is quiescent the agent switches to a grazing mode (see § 2 *a*). The computational basis for the activity of the command neuron is its short-term memory, maintaining grazing mode for several time-steps after eating. An inversion in this neuron switches the agent to the wrong behavioural mode. *Such a switch, even lasting only one network cycle, affects the agent's behaviour for quite a long time, and as a consequence influences greatly the agent's performance*. That is, switching an exploring agent to the grazing mode may cause the agent to take a wrong turn and it will then have to explore for a much longer time before reaching the food zone. Furthermore, due to the memory dynamics maintaining grazing, the agent will keep the grazing behaviour for several cycles. In turn, switching a grazing agent to the exploration mode abolishes its memory of the grazing mode. Consequently, the agent will eat less in the next several cycles. Furthermore, it might leave the food zone during those cycles and will have to find its way back. Therefore, the effect of a rare inversion in a command-type neuron is long lasting. Thus, its CV monotonously increases as λ is decreased, as observed. The command neuron (5) has the highest CV among all neurons for small λ, while being only the third most significant neuron when stochastic lesioning is employed. For $\lambda = 0.1$ it governs almost half the sum of all CVs (figure 6*d*). The long-term importance of lesioned neurons is hidden when severe lesioning is used since every firing inversion is followed shortly after by other inversions, masking these long-term effects. *Small lesioning, in contrast, allows for the measurement of the propagation in time of the lesion effects by opening a 'temporal window of opportunity' and thus reveals the neuron's long-term effects on normal network dynamics*. Neuron 1, the motor neuron enabling the forward movement, shows an inverse dependency of its CV on λ. An inversion in this neuron's firing in one network cycle might cause the agent to perform a wrong action in the next cycle. *Nevertheless, this neuron does not exhibit memory and its effect on the rest of the network is mainly restricted to a single cycle after which it rapidly decays*. Therefore, this neuron's CV is monotonously decreasing as λ is decreased. It has the highest CV when stochastic lesioning is employed, while it is only in third place at low lesioning levels.

Figure 6*b–d* also indicates that the CVs of the neurons tend to approach limit values as λ decreases. This observation supports the possibility that such limit values exist, reflecting the CVs of the neurons in the neurocontroller's normal, intact state. Further verification and substantiation of this observation requires extensive numerical studies at very low λ values (and therefore extremely computationally exhaustive), which are beyond the scope of this study.

Another factor one should consider in assessing the dependence of the CVs on the lesioning level is the neuron's activity level. In all lesioning levels, the probability that the neuron undergoes inversion in a certain cycle is higher for neurons with higher variance (i.e. with average activity closer to 0.5), due to the constraint of constant activity (3.8). The lower λ is, the lower the probability of an inversion, but importantly this decrease in the inversion rate as λ decreases is faster for neurons with higher variance. Thus, neurons with higher variance might exhibit lower CVs (in relation to the others) as λ decreases, independent of their real significance. Yet, according to our experiments, this effect vanishes and the ILM-based FCA is very robust in capturing the long-term effects of a neuron, practically independent of its average activity. In particular, the command neuron analysed above has the highest variance among all neurons in the network. Nevertheless, its CV increases as λ decreases, rather than decreasing, as may be suggested by its average activity. Figure 6*e* plots the distance between the ILM contribution vector and that obtained using stochastic lesioning, as a function of λ. As expected, the distance increases as λ decreases. Interestingly, at least in the example of S10, when a small lesioning level is employed, the ILM contribution vector obtains an almost random position in space relative to the stochastic lesioning contribution vector.

In summary, the neuronal CVs are stable across different runs of the FCA. The lesioning level has a strong effect on the CVs: for small lesioning levels, the contribution vector reflects the longer-term contribution of the elements to the network's processing. For quantifying the shorter-term contributions *solely*, one should use a higher lesioning level. To gain further insights into task localization in the neurocontroller we may compare short-term and long-term CVs by using several lesioning levels. Finally, the CVs tend to approach limit values as λ is decreased.

To test these results further, we have performed ILM-based FCA on two more agents: S5, with a sensorimotor environment identical to S10 but with a neurocontroller consisting of only five neurons, and G10, which is evolved in the same environment as S10 with one additional constraint; it has to close its mouth after eating, to 'digest' the food. Otherwise, the sensors and motors as well as the fitness measure of the agent are identical to those of S10. Figure 7 compares the accuracy of the description obtained by applying stochastic lesioning ($\lambda = 1$) and a lower ILM lesioning level ($\lambda = 0.5$) on these three agents. It is apparent that the FCA based on an ILM lesioning level of $\lambda = 0.5$ outperforms the one based on stochastic lesioning. A comparison of the CVs obtained in both lesioning levels reveals similar phenomena to those presented above in all these agents (data not shown). That is, the CVs vary with the lesioning level, capturing the lesions long-term effects and approaching limit values at small lesioning levels.

(*c*) *Generalization and prediction from small training sets*

The results of the analyses in the previous sections were based on training the FCA with the full set of 2^N configurations, where N is the number of neurons in

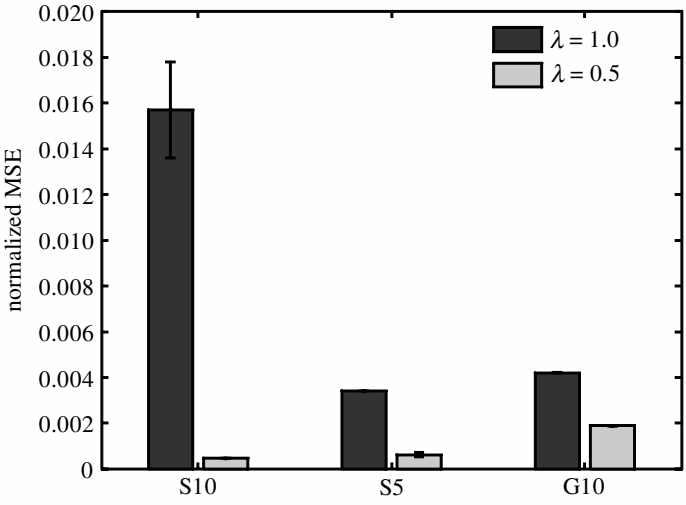

Figure 7. Normalized MSE obtained with two lesioning levels. Mean and standard deviation of the normalized MSE are plotted for three agents: S10, S5 and G10. Training and test sets are the full 2^{10} configurations sets for S10 and G10, and the full 2^5 configurations set for S5.

the network. Previous work (Aharonov *et al*. 2003; Segev *et al*. 2002) showed that a much smaller training set suffices in order to achieve good performance prediction for the test set. This work has been based on stochastic lesioning. We test here the FCA generalization in the more general framework of ILM.

For each lesioning level ($\lambda = 0.1, 0.2, \ldots, 1.0$) we apply the FCA to S10 with a training set of 200 configurations chosen at random. The test MSE is calculated on the test set consisting of all 2^{10} configurations. The results, summarized in figure 8*a*, testify that the FCA generalizes well for any lesioning level. The test MSE is monotonously decreasing as the lesioning level is decreased, demonstrating that the FCA predicts more accurately the effects of smaller perturbations. The lowest test MSE is achieved for the case of $\lambda = 0.1$ and is equal to 5.6×10^{-4}. This corresponds to explaining 99.944% of the variance of the test set. Figure 8*b*,*c* shows that the CVs assigned by the FCA using the 200 configurations sets are very similar to those obtained by using the full sets of 2^{10} configurations†, presented in § 4 *b*. This similarity further testifies to the generalization capabilities of the ILM-based FCA.

(*d*) *Analysis of the synaptic network*

Since the FCA finds the CVs of system elements in general, it is possible to analyse the neurocontroller on the level of its synapses. This is important both for further understanding the underlying architecture of the network and for studying the scalability of the FCA at various lesioning levels (working on the synaptic level considerably increases the number of system elements).

† For simplicity, the figures illustrate these results for two ILM lesioning levels. The same holds for all other lesioning levels.

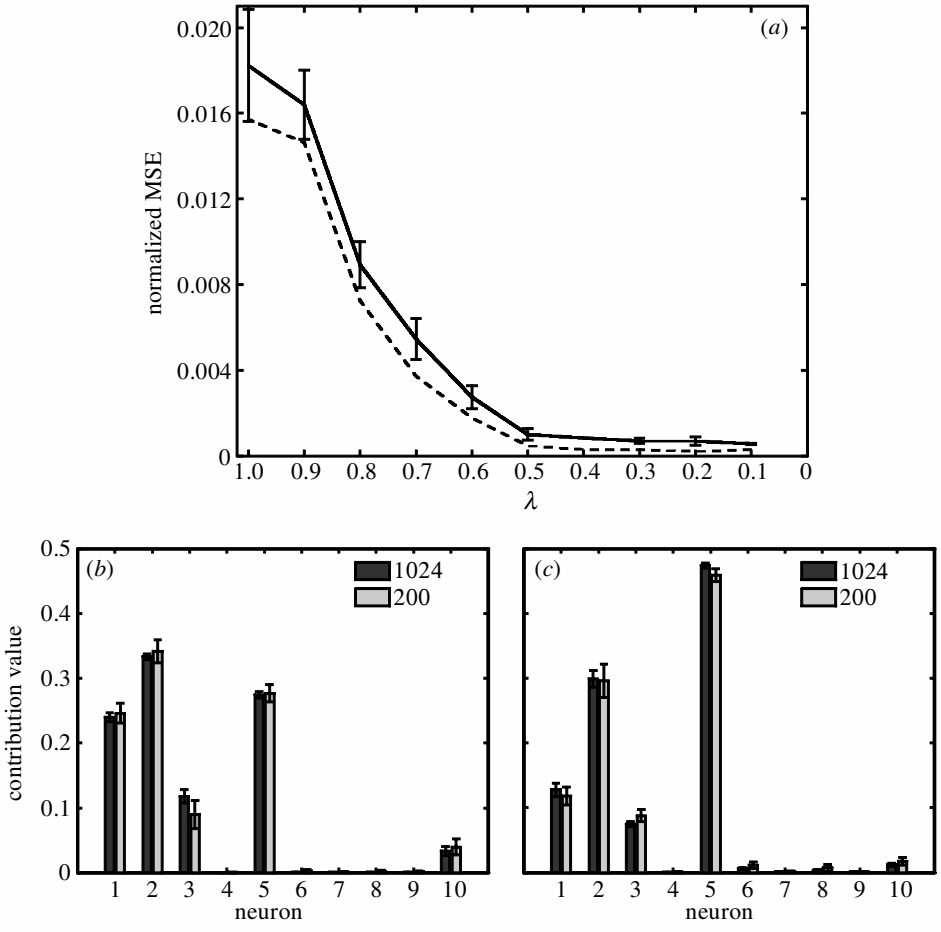

Figure 8. Generalization from a small training set. For each lesioning level λ, the FCA is trained using 200 random lesioning configurations. The test sets are the full 2^{10} configurations sets. (*a*) The mean and standard deviation of the test MSE are plotted against λ (solid line). For comparison, figure 5 is reproduced here (dashed line). (*b*), (*c*) The mean and standard deviation of the CVs obtained using 200 random lesioning configurations are plotted alongside the CVs obtained using the full training sets of 2^{10} configurations, for a lesioning level of (*b*) $\lambda = 0.8$ and (*c*) $\lambda = 0.2$.

We have applied the ILM-based FCA to S10's internal recurrent synaptic network, for a total of $10 \times 10 = 100$ synapses.† Each synaptic lesion configuration $\boldsymbol{m}$ indicates for each synapse whether it is lesioned or left intact. The ILM assigns each lesioned synapse a noisy channel yielding the desired lesioning level. The input to the channel is the original firing of the presynaptic neuron in its intact state. The output of the channel serves as the presynaptic activity under lesioning. The training set consists of 5000 random synaptic lesion configurations, extremely small relative to the full space of 2^{100} possible lesion configurations. The test set consists

† The sensory input synapses are not included in this analysis because our prime goal is to characterize the recurrent processing ongoing in the network.

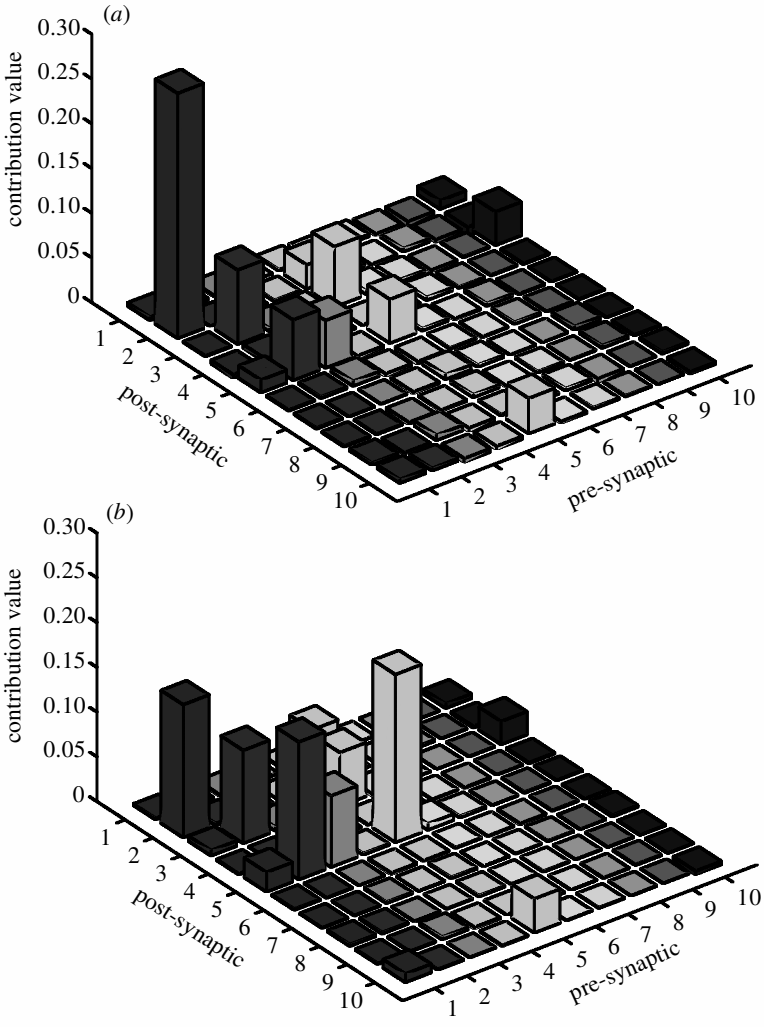

Figure 9. Comparison of the synaptic CVs obtained with different lesioning levels. The mean CVs for the internal synapses are plotted for lesioning levels of (*a*) $\lambda = 1$ and of (*b*) $\lambda = 0.5$.

of another 9000 random configurations. This is repeated for both lesioning levels of 1 and of 0.5. The FCA using a lesioning level of 1 results in quite a high mean test MSE of 0.078, corresponding to explaining 92.2% of the variance. The FCA using a lower ILM lesioning level of 0.5 improves the prediction, yielding a mean test MSE of only 0.007, corresponding to explaining 99.3% of the variance. *This testifies to the potential scalability of the FCA using small ILM lesioning levels.*

Figure 9 compares the synaptic CVs yielded by the FCA for the two lesioning levels. The results are similar in the sense that the same synapses have a significant CV in both lesioning levels (as in the neuronal analysis in § 4 *b*). Nevertheless, the significant CVs themselves differ greatly between the two lesioning levels. The most significant CV increase when decreasing the lesioning level is exhibited by the recurrent synapse from neuron 5 to itself. This increase indicates that the contribution of this synapse is

mainly due to its long-term contribution to the network processing. If one suspected from the behaviour of the agent that a memory mechanism had evolved in this neurocontroller, *one could conclude based solely on the ILM analysis, that this synapse plays a major role in this memory mechanism.* Indeed, neuron 5 is the command neuron and its recurrent synapse comprises the short-term memory which maintains the grazing mode for several time-steps after eating and maintains the exploration mode otherwise (Aharonov-Barki *et al.* 2001). Its main role in the network's long-term processing is also supported by the large CVs of both its incoming (row-wise) and outgoing (column-wise) synapses (figure 9*b*).

5. Discussion

Studying ILM within the FCA framework, we have shown that the lower the lesioning level, the more accurate the FCA description of the agent's performance levels. The improvement is by at least one order of magnitude, which permits the usage of the ILM for the accurate analysis of more complex networks than those that could have been previously analysed with stochastic lesioning. Analysing the contribution vector as a function of the lesioning level λ reveals new insights concerning the short-term versus long-term effects of lesioning, showing that the usage of minute levels of ILM lesions can uncover the long-term effects of elements on the network's functioning. We have also observed that as the lesioning level decreases, the CVs tend to approach limit values. These results testify to the efficiency of the ILM-based FCA as an enhanced method for a systematic and rigorous analysis of function localization in EAA neurocontrollers.

As the lesion level is decreased, some elements' CVs monotonously decrease, while the CVs of other elements monotonously increase. Specifically in S10, the neuron that attains the highest values in low lesion levels is neuron 5, the command neuron (capturing almost half of the total neuronal contributions). As shown previously in Aharonov-Barki *et al.* (2001), the dynamics of this neuron's processing are strongly influenced by a synapse to itself which effectively induces memory-dependent bistable dynamics; the neuron switches from one firing state to the other and back again, each state commanding a distinct mode of behaviour of the agent. As a result, a rare inversion of the neuron's firing state (low ILM lesioning levels) will have a profound long-lasting effect on the agent's behaviour, as it will switch the agent to a different, discordant behaviour mode, in which it will then remain for a long time. Note that the long-term effects play a less drastic role in the CV estimation when frequent inversions of the state of the command neuron are employed (higher levels of ILM lesioning) because the long-term effects of an inversion are masked by the effects of subsequent inversions.† It is clear that long-term contributions are an essential part of intact network processing, and one must trace them to understand the latter. This gives rise to the intuitive notion that CVs obtained with low lesioning levels better reflect the 'real' contribution of neurons to the intact network processing than the CVs obtained with high lesioning levels.

In addition to improving the accuracy of the FCA and revealing the long-term dynamics, this study has been motivated by a more important consideration; the

† One should bear in mind that this description is obviously a simplification because the normalization employed determines the contribution of a neuron *relative* to the others' CVs.

need to address the paradox of the lesioning methodology, which essentially involves learning about the network's normal mode of operation by examining the agent's behaviour in a series of perturbed states *which are different from its normal state*. The straightforward hypothesis underlying this study has been that this potential caveat may be overcome by a lesioning method that employs very minute lesions, and thus studies the network in a series of perturbed states which are very close to the network's normal mode of operation.† Indeed, our results show that this is feasible: if one gradually decreases the lesioning level in accordance with the controlled regime dictated by the ILM method, the CVs of the system's elements tend each to converge to its unique limit value. These values are likely to reflect the contributions of the system's elements in its normal, intact mode of operation, *when studied via ILM*.

Previous FCA studies have shown that when biological versus stochastic lesioning strategies are used, the CVs obtained are not universally identical and may strongly depend on the lesioning method (Aharonov *et al*. 2003). These lesioning methods have employed large perturbations and thus had not aimed to reflect the 'true' CVs of the network which characterize its intact mode. Bearing this in mind brings home the following question: will any small-perturbation lesioning method yield CVs similar to those found by ILM? Disappointingly, perhaps, we hypothesize that the answer to this question is negative; if one uses an 'ILM-like' variant in which different elements are lesioned with different λ levels, one is likely to obtain different CVs from those obtained with the original, uniform level ILM, even when one gets very close to the network's intact mode. Yet one should cautiously note that this still remains an open question, whose investigation shall require very extensive and time-consuming numerical studies involving very long runs that are necessary if one wishes to measure the CVs at vanishing lesioning levels in a discrete, stochastic framework.

Future studies of lesioning strategies should include an investigation of more general ILM methods that do not necessarily apply the same lesioning level at all lesioned elements. A further extension may be the use of even more general lesioning methods, abandoning the binary dichotomy of a lesioning configuration where some elements are lesioned but others are left intact, i.e. in such a lesioning method, *every* element is lesioned to some level in a given lesion configuration. These more general lesioning methods may be better suited for active learning adaptive algorithms (Cohn *et al*. 1995; Engelbrecht & Cloete 1999) that select the best lesion configurations with which the FCA should be trained, since they employ a continuous lesion configuration space. They are also likely to allow for more tractable studies of the CVs at the limit of vanishing lesioning levels.

ILM-based FCA has been developed and studied in this research primarily as a method for localizing function in EAA neurocontrollers. Yet, it should be noted that the FCA is currently being applied to study real biological nervous systems: to the analysis of reversible lesioning studies in the cat (i.e. biological lesioning), and to the analysis of 'virtual lesions' induced in humans via transcranial magnetic stimulation (i.e. stochastic lesioning). The basic elements analysed in these ongoing studies are fairly large brain regions containing millions of neurons. ILM-based lesioning is not applicable to such large-scale networks, but in the future it may become a useful

† From a technical perspective this makes further sense, since the FCA is essentially a projection pursuit model and such models are known to provide a good approximation to any well-behaved continuous function given small perturbations from a state where its value is known.

tool for the analysis of smaller neuronal circuits, where the elements of analysis of the system are its neurons and synapses. As previous studies have shown (Servan-Schreiber *et al*. 1990; Usher *et al*. 1999), the action of aminergic neuromodulation on neurons may be modelled as a variation in the gain of the neuronal response function. This gives rise to the possibility of pharmacologically manipulating the fidelity of neural response and carrying out an ILM-based FCA in biological neural networks on the neuronal level.

In summary, this paper presents an in-depth study of the classical lesioning paradigm, by developing a new, controlled lesioning method and studying it in an EAA model. We showed that the ILM-based FCA is an accurate lesion-based performance prediction algorithm, capable of unearthing neurocontroller long-term dynamics and approaching limit values when very small lesion levels are used.

We acknowledge the valuable contributions and suggestions made by Ranit Aharonov, Hezi Avraham, Yossi Mossel, Lior Segev, Ron Mertens and Haim Sompolinsky, and the technical help provided by Oran Singer. This research has been supported by the Adams Super Center for Brain Studies in Tel Aviv University and by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities.

References

Aharonov, R., Segev, L., Meilijson, I. & Ruppin, E. 2003 Localization of function via lesion analysis. *Neural Comput.* **15**, 885–913.

Aharonov-Barki, R., Beker, T. & Ruppin, E. 2001 Emergence of memory-driven command neurons in evolved artificial agents. *Neural Comput.* **13**, 691–716.

Cohn, D. A., Ghahramani, Z. & Jordan, M. I. 1995 Active learning with statistical models. In *Advances in neural information processing systems* (ed. G. Tesauro, D. Touretzky & T. Leen), vol. 7, pp. 705–712. Cambridge, MA: MIT Press.

Cover, T. M. & Thomas, J. A. 1991 *Elements of information theory*. Wiley.

Engelbrecht, A. & Cloete, I. 1999 Incremental learning using sensitivity analysis. *INNS/IEEE Int. Joint Conf. on Neural Networks*, paper 380. Washington, DC: IEEE Press.

Floreano, D. & Mondada, F. 1996 Evolution of homing navigation in a real mobile robot. *IEEE Trans. Syst. Man Cybern.* B **26**, 396–407.

Floreano, D. & Mondada, F. 1998 Evolutionary neurocontrollers for autonomous mobile robots. *Neural Netw.* **11**, 1461–1478.

Gomez, F. & Miikkulainen, R. 1997 Incremental evolution of complex general behaviour. *Adapt. Behav.* **5**, 317–342.

Harvey, I., Husbands, P. & Cliff, D. 1994 Seeing the light: artificial evolution, real vision. In *From Animals to Animats 3. Proc. 3rd Int. Conf. on Simulation of Adaptive Behavior* (ed. D. Cliff, P. Husbands, J.-A. Meyer & S. W. Wilson), pp. 392–401. Cambridge, MA: MIT Press.

Kodjabachian, J. & Meyer, J. A. 1998 Evolution and development of neural controllers for locomotion, gradient-following and obstacle-avoidance in artificial insects. *IEEE Trans. Neural Networks* **9**, 796–812.

Marocco, D. & Floreano, D. 2002 Active vision and feature selection in evolutionary behavioral systems. In *From Animals to Animats 7. Proc 7th Int. Conf. on Simulation of Adaptive Behavior, Edinburgh, UK* (ed. B. Hallam, D. Floreano, J. Hallam, G. Hayes & J.-A. Meyer), pp. 247–255. Cambridge, MA: MIT Press.

Scheier, C., Pfeifer, R. & Kunyioshi, Y. 1998 Embedded neural networks: exploiting constraints. *Neural Netw.* **11**, 1551–1569.

Segev, L., Aharonov, R., Meilijson, I. & Ruppin, E. 2002 Localization of function in neurocontrollers. In *From Animals to Animats 7. Proc 7th Int. Conf. on Simulation of Adaptive Behavior, Edinburgh, UK* (ed. B. Hallam, D. Floreano, J. Hallam, G. Hayes & J.-A. Meyer), pp. 161–170. Cambridge, MA: MIT Press.

Servan-Schreiber, D., Printz, H. & Cohen, J. D. 1990 A network model of catecholamine effects: gain, signal-to-noise ratio and behavior. *Science* **249**, 892–895.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423; 623–656.

Stanley, K. O. & Miikkulainen, R. 2001 Evolving neural networks through augmenting topologies. Technical Report TR-AI-01–290, Department of Computer Science, The University of Texas at Austin, TX, USA.

Usher, M., Cohen, J., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. 1999 The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554.

Yao, X. 1999 Evolving artificial neural networks. *Proc. IEEE* **87**, 1423–1447.

Young, M. P., Hilgetag, C. C. & Scannell, J. W. 2000 On imputing function to structure from the behavioural effects of brain lesions. *Phil. Trans. R. Soc. Lond.* B **355**, 147–161.