Higher-Order Balances: Unifying Views of Data?

Heikki Hyötyniemi

Cybernetics Group
Helsinki University of Technology, Control Engineering Laboratory
P.O. Box 5500, FIN-02015 HUT, Finland
heikki.hyotyniemi@hut.fi
http://www.control.hut.fi/hyotyniemi

Abstract. It seems that many of the most challenging complex systems are *cybernetic.* This cybernetic nature means that there are special properties about them that should also be reflected in observations that are to be used for capturing their behaviors. It seems that the cybernetic intuitions offer fresh ways to attack the problems of semantics, and, further, the mysteries of cognition, for example, can be seen in a new perspective.

1 Introduction

In ancient Greece, the philosopher Heraclitus observed that

> *... The way up and the way down are one and the same.*

There are many ways to interpret this — already in his times, Heraclitus was called *skoteinos,* or *obscure.* One interpretation is that if one wants to understand phenomena in the real world, the observations have to be perceived applying the same principles. Using the modern engineering-like terminology, one could say that *the system and its model should have the same underlying structure.*

Another issue emphasized by Heraclitus was that

> *... Harmony consists of opposing tension, like that of the bow and the lyre.*

Complex systems are seemingly static but full of potential ready to burst out. The world is governed by harmonies even though there are hidden tensions underneath. And, as also noted by Heraclitus, the *hidden harmonies* are more powerful than the visible ones. The above ideas where there is competition and balancing between different strivings are today studied in the framework of *cybernetics.* To make the issues still more explicit, Heraclitus states that

> *... All things happen because of strife and necessity.*

Cybernetic studies have shown that local decisions (trying to stay alive!) truly can result in interesting system-level emergent behaviors (see [12]).

In this paper, the above visions are put in practice. The deep but hazy Heraclitus aporias are given concrete semantics, and efficient tools (that is, mathematics, and modern system theory) are used to functionalize the views. First, the

today's challenges in modeling of data from complex systems are discussed, and the cybernetic approach is presented. Linear cases are studied in detail to reach some intuition. The results are applied to analysis of how the mental machinery could be explained: For example, the idea of *functional chunks* is perhaps given new substance.

2 Cybernetic models for cybernetic systems

When aiming towards constructing *smart* systems capable of autonomously reacting to changes in their environments, some kind of *understanding* of the environment has to be implemented in those systems. Understanding cannot be implemented without *semantics.* It is evident that one is facing huge challenges here, but it turns out that the fresh intuitions from the field of *cybernetics* offer us new ways to look at the age-old problems. There are two main tasks here: First, it needs to be studied how such self-contained semantic representations could be defined in the first place; second, to transfer this semantics to the smart system, these representations somehow have to be coded in data.

2.1 World as complex data

After the era of rational (or, indeed, rather irrational) approaches to modeling reality, the empiristic views are now dominating: The world is given to us as observations only, and our task is to explain this data. There are new tools supporting these data-oriented approaches — computing capacity has grown exponentially, and also mathematical tools like multi-variate statistics has developed to support data analyses and different kinds of *data mining* techniques.

However, it is Plato's Cave Metaphor that still applies here: If one only studies the data in an empiristic manner, one cannot get beyond that, and the underlying system remains veiled. To connect observations to phenomena, an interpreter is necessary — a human. When trying to create truly *intelligent* systems that are capable of autonomously reacting to the changes in the environment, this approach is not satisfactory.

Are there any ways to circumvent this eternal dilemma of *noumena* versus *phenomena*? A modern approach (one of them) to attack this problem is to employ the *opposite* point of view: Forget about the Platonian Ideas. Thinking pragmatically, one can never know what is the actual structure of the system — the structure that is grounded on data is just as good as any. In *complexity theory* data takes the main role. The emergent structures that are based on the statistical relationships among data can be used for characterizing the system. One searches for the underlying patterns beyond the chaos of data, hoping that computing power can reveal something *interesting.* But, after all, again one is facing the vicious circle: How to tell the computer what kind of emergent structures are interesting?

The theory of complex systems is a close relative to artificial intelligence (AI) research. In both fields one is missing concrete definitions: What kind of behavior

represents *complexity* and what does not? For example, a *chair* can be a magnificent, deeply purposeful, aesthetic combination of simpler parts — is complexity manifested in it? In a complex system, there is a delicate balance between order and chaos: Complete order is not interesting, but neither is complete disorder.

If there are no grains of relevant information present in the data, mindless thrashing of that data only gives trash out. Only applying computing power does not help when one wants to implement intelligence in systems; computation cannot restore information if it has originally been ripped off. The essence of the system somehow has to be captured in the data. In concrete terms, when implementing some data processing application, one has to answer the question: Which variables to include, and which ones to ignore? This paper tries to give some intuition here.

2.2 Challenge of semantics

A good model should reflect the system being modeled. It is not only the structures of the system and the model that should be matched, but also the *functions;* one should somehow capture the *essence* of the system. To reach this, one needs to study the problems of *semantics.* Syntax or structure alone is dead if it is not supported by underlying semantics, or *meaning.*

The goal here is to define and implement *semantics outside brain.* In concrete terms, this means that data and its interpretation have to be integrated. There will be a self-contained package of data and program, or syntax and semantics, or information and interpretation. It is evident that only a narrow view of semantics can be captured in an engineering-like way — but such simplifications are necessary to reach a solid basis for further discussions.

The only system where there is self-contained semantics today is the human brain; perhaps this system can offer some intuition. As a starting point, it needs to be noted that a human does not just store data in the brain. Learning is a process of evaluating the data, assessing it, putting it in context. What are connections and consequences of that information? All human knowledge seems to be *functional,* information is stored in terms of cause/effect structures, or in action/reaction pairs. Another way to put this is to say that human data structures are *causal.* This suggests a practical approach to attacking the problem of semantics: Assume that *semantics is reflected in causality.*

The causal structures (in an expanded meaning) convey information of what will result from the current situation. In this sense, such structures implement a *path* in the space of states of the world, binding successive "world snapshots" (or mental snapshots) together. The key point is the connections to other entities and their mutual dependencies. It turns out that this view is an extension of the ideas of *naturalistic* and *contextual semantics* from static to *dynamic* context. In concrete terms, as naturalistic semantics binds entities to actual system inputs (sensory signals, or to signals from separate subsystems), and contextual semantics binds the entities to each other statically, the *causal semantics* binds them to each other dynamically. The spatial model has to be substituted with a spatio-temporal model.

To capture such dynamic phenomena, it seems that some kind of dynamic model is needed (for example, see [8]). The problem is that in high-dimensional cases dynamic models are still more difficult to handle than static ones are: There typically exist very many additional degrees of freedom when the parameters are to be determined. What is more, the traditional models are too rigid — the causal chain is not always temporal. The succession should typically be characterized in terms of transitions rather than imposing smooth dynamics on them. However, such event-based models models without explicit time variable have weaker mathematical structure; it seems that more powerful views are needed here.

2.3 Cybernetic structures and functions

One needs a framework for efficiently studying the (extended) contextual semantics. It seems that the problems with dynamic models can efficiently be avoided when the framework is turned upside down: Rather than trying to capture movement, or changes in the system, let us concentrate on the *balances* — cases where there is *loss* of any dynamics. However, to exploit the above view and still have something non-static, the balance idea has to be interpreted in the correct way: Balances here are *dynamic equilibria.* Intuitively, such balances characterize the system by describing where the natural dynamics would finally take the system, if enough time was available.

A good framework for this kind of studies is that of *cybernetic systems.* As explained in more detail in [12], the essence of a cybernetic system is in *higher-order balances.* It turns out that the nature of such systems is that the underlying tensions are in equilibrium. Cybernetic systems offer a practical framework for studying semantics in such a narrow semse. From the point of view of *cybernetic semantics,* it can be claimed that cybernetic systems are such self-contained systems where the functions have been integrated with the structure. For examples of cybernetic systems, see [14].

Not all systems are cybernetic — but, luckily enough, the most challenging, and the most interesting systems *are.* These systems are those with "hidden tensions" (as Heraclitus put it). A cybernetic system interacts with its environment, being in dynamic balance with it, reacting immediately to environmental changes, searching for the new balance. After a disturbance, a cybernetic system will typically not be the same, there is adaptation to the environment — but the new system is again fully functional, better in the new circumstances, being again in balance.

Another central property of (higher-order) cybernetic systems is *optimality,* in terms of exhausting available resources in the best possible way (see [12]). This optimality must be shared by all sensible systems that have survived and prospered in the cybernetic evolution. What is more, cybernetic systems are typically fractal, consisting of lower-level cybernetic systems. Because there is optimality at each level, there are no superfluous structures or functions, there is minimality in representation. And because the representations are there only to implement functions, it can be said that in a cybernetic system syntax and

semantics are in one-to-one correspondence (in linear systems this holds only to a certain extent, up to the *principal subspace*). In a fully cybernetic system function dictates structure, and structure dictates function. This streamlined functionality can be seen only in the holistic perspective; whereas all structural elements contribute to the overall goal, on the local level the goals are not visible, and the parts are not indispensable[1]

This all means that cybernetically optimal systems are unique to an extent. This structure is determined by the properties of the environment visible in data (as explained in [12], the vector u characterizing the current environment determines the state x, and the statistical properties of u determines the system structure ϕ). The cybernetic system reflects its environment; or, indeed, it is a mirror image of the environment.

Now there is an escape from the Platonian cave: A sensible system must have the cybernetic structure, and within this restricted modeling framework, the free parameters fixing the behaviors can be identified. Note that this holds only to the system level as the single system components cannot be distinguished. What is more, indeed, all cybernetic systems must have the same structure. It is like in *computability theory:* The class of NP complete problems is such that solving one of those problems simultaneously solves the other problems (rather than being exponentially capacity demanding, polynomial amount of resources suffices). Similarly, one could speak of the class of "CS complete" systems: The intuition here is that solving the mysteries in one complex system simultaneously gives the tools to modeling the other complex systems as well (as has always been prophesized in complexity theory).

Still one more concrete criticism needs to be commented here — it is that of causality: No data is enough to certainly determine the dependency structures underlying observations. This Humean problem can be circumvented here, because one does not try to extract individual causality patterns. It is now implicitly assumed that there are causal connections between *all* entities, *all* variables are related to each other in a network, balancing each other's causal strivings. One-signal-at-a-time analyses truly are doomed — but they are not needed.

When having the data, two kinds of cybernetic models can be constructed applying the guidelines presented in [12], depending on the data properties: First, if there is no underlying structure beneath the static data, one can still have useful results; the cybernetic modeling machinery carries out (sparse coded) principal component analysis of data, so that optimal compression in terms of data variance is reached. But more interesting models can be reached if the data is well-conditioned, reflecting the dynamic balances within the system. How

[1] So, is the *chair* a cybernetic entity? No, it is not. If a chair is scratched, there is now self-healing capability, and gradually this "system" is ruined by the knocks coming from outside. A chair represents temporary, "dead" balance: It has been constructed once and for all applying external forces rather than trusting the internal ones. The structure and function of a chair are not in one-to-one correspondence. There can be scratches in a cybernetic system, too, but the *function is changed accordingly*

could structureless data carry such information, or, indeed, *knowledge,* about the domain field?

2.4 Data with balance

Static data is, by definition, in balance, and the cybernetic model of such data is balanced through associations among entities (as will be explained in more detail later). However, if the data should represent dynamics, one is facing more challenges.

Above, the problem of representing dynamics was already simplified applying cybernetic intuitions: There is no need to worry how to represent the actual trajectories, or functions of some free spatio-temporal variables as data; one only needs to represent the eventual dynamic equilibria. In practice, there is a dilemma: The balances cannot be easily determined. A cybernetic system itself typically never represents the balance state; balances could be simulated if there was a model, but to start with, there is no model — indeed, data is collected to determine that model! What is more, there are typically many alternative routes towards balance in complex systems, resulting in different equilibria — which one of them to select?

Again, it turns out that turning the problem upside down helps to proceed: Rather than looking at the hypothetical distant balance, one can *make the current state balanced.* Semantics is not necessarily in actual movements but in *causal tensions* causing those movements. One can also study what is the *opposing force* needed to neutralize the natural dynamic tendencies of the system. Rather than doing global analyses, local analyses are only needed: Looking at current state of the system and its "flow" in that state, the dynamic data can consist of temporally local observations. From the practical point of view, there is another advantage here — *relevance:* only those states that have actually been seen are represented in data, not some hypothetical future states that probably will never be reached.

The intuitive notion of *force* needs to be formalized. To do this in a consistent way, one needs to introduce the concept of *energy function.* This cost criterion defines a *potential field* in the state space; to move "higher" in that landscape one needs to apply force, and this force is accumulated as potential energy level. Indeed, the force can be defined in terms of energy that is needed to produce a certain movement in the potential field. In mathematical terms, this can be expressed in differential form as

$$F(x) = -\frac{dJ}{dx}(x), \tag{1}$$

that is, the (virtual) force is the virtual change in energy that would be needed to cause a hypothetical movement; the minus sign reveals that the force points against the gradient[2].

[2] It needs to be noted that coding of virtual movements is a deep philosophical problem, as manifested in the Zeno's "arrow paradox": *During each time point an arrow*

This energy function J is a central concept in discussions that follow. In different environments, it can have differing semantic interpretations: For example, it can be called a *fitness landscape* characterizing how far the system is from (local) minimum. Essentially, J determines the domain field; however, because the cybernetic systems are *constructivistic,* typically it is a function of not only the environment but also of the system itself, so that one has $J(x, u)$. It needs to be recognized that function J needs not be known explicitly, it is only important that it can be assumed to exist. Locally, when the system is in some specific state, the properties (gradient) of J can be directly observed in the system's behavior.

Assume that the system can be characterized in terms of a variable vector x. Then the *minimum-dimensional* data vector containing cybernetically relevant information of the system state (configuration, and the corresponding forces) could be constructed, for example, as[3]

$$u = \left(\frac{x}{\frac{dJ}{dx}(x)} \right). \tag{2}$$

If defined so, the balance can (at least in principle) be constructed as a linear function of the variables; because the cybernetic model is essentially linear, relevant information is now coded in the data in such a form that it can be exploited by the cybernetic modeling machinery. Because of the underlying PCA-based multivariate modeling structures, an appropriately constructed cybernetic model is robust against redundant variables, so that the dimension of u needs not be explicitly minimized; formula (2) represents the minumum set of variables that necessarily has to be included.

The above discussion proposed a generic approach towards modeling complex data based on local balances. The key issues where the energy function defining the "causal landscape", and its (negative) gradients determining the "causal forces". In practice, what do the energy functions typically look like? Such questions are studied closer in Sec. 3. The abstract discussions are made more concrete towards the end of the paper, offering interesting new views for AI research in Sec. 4.

---

is still, so that there can exist no movement at all. Similarly, each data sample is a snapshot — and, similarly, it cannot represent dynamics? What the Greeks did not understand was the power of mathematical concepts such as *differentials.* When time interval goes to zero, and the movement during that time interval correspondingly goes to zero, their ratio, or the velocity can still remain bounded. Just as velocity, also force is a mathematical abstraction based on changes in variables, and thus involving differentials. Differentials can efficiently be used to code rates of change, even though it seems that today's people also feel uncomfortable when facing them

[3] Alternative sets of variables can also be selected, as long as there are some kind of gradients involved: For example, when characterizing mechanical systems, the vector q of *generalized coordinates* suffices to uniquely determine the system configuration at any time instant; it turns out (see [6]) that when characterizing the dynamic system state, it is exactly the vectors q and $\dot{q}$ that are needed

2.5 Models of the *relevant*

Typically in engineering work, the global optimum of some criterion is searched for. This applies as well to analysis, or modeling of existing systems, as to synthesis, or design of new ones. In the case of complex systems one ends in problems, because there typically exist various local minima for the criterion — one does not know whether the search should be continued, and if so, in which direction. What is interesting is that *nature has the same problem.* Evolution in nature also has to be based on evaluation of some optimality criterion, and nature has no better optimization methods than we have — the clever ones (those that are not based simply on random search) being based on some kinds of gradients. Typically nature also ends in some local minimum when optimizing designs — and if the design is repeated, the system will not be the same again (as Heraclitus put it: "You cannot step in the same river twice").

Indeed, the absolutely global optimum does not necessarily reflect typical existing systems. A model, however good it is, only represents the current system being studied. Rather, one would need a "higher-level" model over the range of models. It seems, of course, that as a single model may be difficult to construct, a model of models is a still more difficult goal. However, this is not necessarily true: Just as in mathematics many problems of real analysis can easiest be solved by extending the problem to complex domain, escaping from the real axis to the complex space, some modeling problems can actually become simpler when one studies the whole space of systems rather than a single individual system at a time.

A more complete view of the variability range among the models in the environment is given by the range of local minima — and it is these local minima of the energy function that are being captured by the cybernetic model. A cybernetic model is a model over the possible solutions. For example, it seems that a promising approach towards understanding *gene expression* is to think that the set of active genes determines the "subspace", and other metabolic processes optimize within it, so that the system ends in the corresponding local minimum.

This issue deserves to be emphasized: A cybernetic model is a higher-level integrated model of many local minima rather than a model of a single global optimum. The cybernetically balanced global model is an optimized model of local balances. It is a *model of the possible,* or *potential* rather than *actual;* or, because the data comes from real observations, it is a *model of the relevant.* Here it is claimed that this approach captures the fundamentally random nature of Nature in a more consistent way than traditional approaches. Rather than pursuing absolute optimization (typically being an NP hard problem), one searches for the spectrum of "nearby" solutions, trying to characterize them in a compact way.

It needs to be noted that speaking of local minima is misleading, and does not characterize the nature of the solutions in the correct way: At least if the system is linear, the balance solutions are strictly optimal. It is just the environment that may change, either as a function of the location or the time. In this sense one could speak of "spatio-temporal optima". The optima are "parame-

terized" by the input vector u, the model mapping from the environment to the corresponding balance.

Of course, because of their huge importance, different kinds of "models of models" have been studied before in different environments. For example, in *Hopfield nets* (for example, see [9]) the properties are similarly analyzed in terms of energy functions, and, similarly, the contents of the associative structure are characterized by the stable equilibria. However, Hopfield nets are only used as an associative memory for storing distinct patterns — there is no model among the minima, and, because the variables are typically binary, there is no continuity. In *Genetic Algorithms* (see [1]), on the other hand, "fitness landscapes", or energy functions, are discussed, and one stores the set of candidate solutions, trying to preserve the good candidates in the population. No higher-level model of the solutions is created, individual models are just combined and modified to reach better candidates; indeed, GA is yet another methodology for implementing search for the global optimum.

3 Analyses and intuitions

The above discussions offer interesting views not only to analysis of existing systems, but also to analysis of systems *as they could be.*

3.1 Nature of cybernetic models

How can cybernetic data be characterized? No matter how the data vector is constructed, whether it contains information of the underlying forces or not, one interpretation is that the cybernetic model spans a (locally) linear subspace where the data samples are assumed to reside. The model consists of (locally) linear set of *features* characterizing the degrees of freedom in the data. Getting from data to model is a feature extraction problem.

Assume that a domain field is characterized in terms of linear features φ_i, where $1 \leq i \leq n$, the number of features n being lower than the data dimension m. The measurements u should be represented as a weighted sum of those features:

$$u(x) = \sum_{i=1}^{n} x_i \, \varphi_i = \varphi x, \tag{3}$$

where the $m \times n$ matrix φ contains the features, and x represents the system *state*. Assuming that an observation u is known, the goal is to determine the state; because of the dimensions of φ, this cannot generally be done exactly, only in some approximate sense. An estimate for the state $\hat{x}$ corresponding to u can be characterized in terms of a criterion

$$J(x) = \frac{1}{2} \left(u - \varphi x\right)^T \left(u - \varphi x\right). \tag{4}$$

The minimum for this is given by

$$\bar{x} = \left(\varphi^T \varphi\right)^{-1} \varphi^T u. \tag{5}$$

As explained in [13], solution this can be interpreted as a balance where the gradient of (4) vanishes. In this sense, the solution (5) is a cybernetic solution, and the original feature system is a cybernetic system — in the simplest sense (see [12]). Does the system also represent *higher-order balance?* In that case there should hold

$$\left(\varphi^T \varphi\right)^{-1} \varphi^T = \left(\mathrm{E}\{\bar{x}\bar{x}^T\}\right)^{-1} \mathrm{E}\{\bar{x}u^T\}. \tag{6}$$

Without going into details (see [13]), the uniqueness of features can only be reached up to the *principal subspace,* that is, only the subspace spanned by the features is found, not the features themselves (nonlinear cases, and sparse coding, are another issue).

3.2 Properties of *virtual data*

It is interesting to study whether one can solve the *inverse problem,* or determine the properties of such data that are produced by some given structure; one could speak of "virtual data".

Not very much can be said about data in general terms. Now, however, following the discussions in the previous section, assume that $\bar{x}$ can be interpreted as a zero point of a gradient of some optimality criterion $J(x, u)$. To proceed, one can, for example, write the *Newton algorithm* for iteratively refining the estimate for the zero point:

$$\bar{x}(\kappa + 1) = \bar{x}(\kappa) - \left(\frac{dJ^2}{dxdx^T}(\bar{x}(\kappa), u)\right)^{-1} \frac{dJ}{dx}(\bar{x}(\kappa), u). \tag{7}$$

Selecting the optimality criterion is difficult — indeed, as noted above, the whole problem domain is characterized by this function. From the point of view of probabilistic modeling, a well-motivated energy function form is the (inverse of) *log-likelihood criterion.* For simplicity, assume that the distribution of the multivariate data u is Gaussian, so that log-likelihood is quadratic. The second derivative, or Hessian, of a quadratic criterion is constant:

$$\frac{dJ^2}{dxdx^T}(x, u) = \frac{dJ^2}{dxdx^T}. \tag{8}$$

Because of the probabilistic framework, one has to use expectations. In the case of log-likelihood functions, the Hessian can be expressed in terms of the *information matrix* as

$$\mathrm{E}\left\{\frac{dJ^2}{dxdx^T}\right\} = -\mathrm{E}\left\{\left(\frac{dJ}{dx}\right)\left(\frac{dJ}{dx}\right)^T\right\}. \tag{9}$$

Applying (9), and remembering that for a quadratic cost criterion the Newton iteration becomes a one-step process, regardless of the initial guess $\bar{x}(0)$, from (7) one has

$$\bar{x} = \left(\mathrm{E} \left\{ \left(\frac{dJ}{dx} \right) \left(\frac{dJ}{dx} \right)^T \right\} \right)^{-1} \frac{dJ}{dx}(x, u). \tag{10}$$

This expression has a structure that is formally very near to that in (5). This gives intuition of how the features can be interpreted: Assume that the landscape has been characterized by the "gradient prototypes" θ_j, where $\theta = \varphi^T$: They construct a "gradient landscape" in the space of u, so that along u_j the gradient in x space is θ_j. The actual gradient is a weighted sum of the gradient prototypes, or

$$\frac{dJ}{dx}(x, u) = \sum_{j=1}^{m} u_j \, \theta_j. \tag{11}$$

This can be written in the matrix form

$$\frac{dJ}{dx}(x, u) = \theta u = \varphi^T \, u. \tag{12}$$

When integrated, one has the cost criterion $J(x, u)$ (this is unique up to the constant of integration):

$$J(x, u) = x^T \varphi^T \, u. \tag{13}$$

There are two variable vectors, and one would like to simplify this expression. The most natural assumption is that one is only interested in cybernetically motivated, or balanced, situations, so that $x(u) = \bar{x}(u)$; to do this, (5) has to be taken into account:

$$J(u) = u^T \, \varphi \left(\varphi^T \varphi \right)^{-1} \varphi^T \, u. \tag{14}$$

This expression reveals that the landscape is quadratic in the original input space, the degrees of freedom being determined by the vectors φ_i. Variables u_i are used as indexes for selecting the appropriate combinations of the gradient prototypes; to make (5) and (10) hold, one must have

$$\mathrm{E} \left\{ \left(\frac{dJ}{dx} \right) \left(\frac{dJ}{dx} \right)^T \right\} = \varphi^T \varphi, \tag{15}$$

meaning that the input space has to be covered so that $\mathrm{E}\{uu^T\} = I$. From the log-likelihood function formulation (14) one can readily reconstruct the covariance matrix characterizing the Gaussian distribution (however, note that the matrix is uninvertible, and the data does not span the whole space of u).

In Fig. 1, simple examples of typical linear "balance landscapes" are presented: The balances, or possible fixed points of the system, are found on the bottoms of the valleys, along subspaces that are orthogonal to the subspace spanned by the n irreducible gradient vectors in the m dimensional input space.

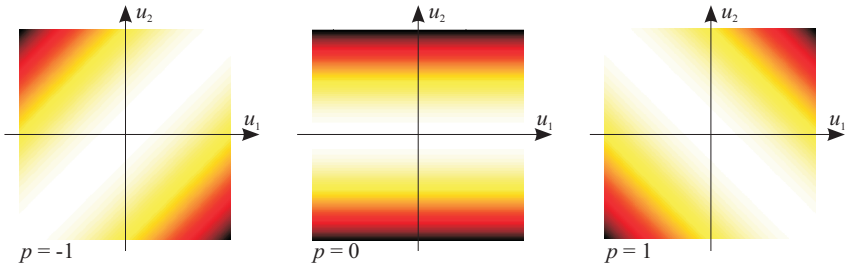

Fig. 1. "Landscapes" for systems with $n = 1$, $m = 2$, and $\varphi = (p\ 1)^T$

3.3 "Truth landscape"

In what follows, a more heuristic example is presented; it gives intuition, illustrating that one does not need to have a vector space with well-defined distance measures to model something interesting in the cybernetic framework. Not all locations in the space of u are now assumed to be meaningful; as an example, the *backward reasoning* framework from [13] is applied. Study the simple sequence of rules

$\mathtt{A} \rightarrow \mathtt{B}$
$\mathtt{B} \rightarrow \mathtt{C}$
$\mathtt{C} \rightarrow \mathtt{D}$.

As shown in [13], the corresponding *rule matrix* containing the information for implementing backward reasoning in the cybernetic framework (note that in this case nonlinearity is also needed; see [13]) has the characteristic structure

$$\varphi^T = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & -1 & 1 \end{pmatrix},$$

and, further,

$$\varphi^T\varphi = \begin{pmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{pmatrix}.$$

Now, assume that there is a continuous variable x_c spanning an "axis" between $\mathtt{A}$ to $\mathtt{D}$ in the space of logical entities, so that the vector x contains (equally spaced) discretized samples of it. Further, assume that there exists a continuous, differentiable function $J(x_c)$ along the axis. The derivative in $\mathtt{B}$, for example, can be expressed as

$$\begin{aligned} \frac{dJ}{dx_c}(B) &= \lim_{\Delta x \to 0} \frac{J(B + \Delta x) - J(B)}{\Delta x} \\ &\approx \frac{J(B+1) - J(B)}{1} \\ &= J(C) - J(B), \end{aligned}$$

assuming that the "distance" between sampling points in the approximated derivative is $\Delta x = 1$. Correspondingly, the second derivative in B can be approximated as

$$\begin{aligned}\frac{d^2J}{dx_c^2}(B) &= \lim_{\Delta x \to 0} \frac{\frac{J(B+\Delta x)-J(B)}{\Delta x} - \frac{J(B)-J(B-\Delta x)}{\Delta x}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{J(B+\Delta x) - 2J(B) + J(B-\Delta x)}{\Delta x^2} \\ &\approx \frac{J(B+1) - 2J(B) + J(B-1)}{1^2} \\ &= J(A) - 2J(B) + J(C).\end{aligned}$$

Comparing these expressions to those of φ and $\varphi^T\varphi$ above, one can see that the rule matrix φ essentially represents the derivatives as collected together; $\varphi^T u$ represents the gradient, u selecting the appropriate derivative expression from the matrix. Matrix $\varphi^T\varphi$ is the (negated) Hessian.

Again, it turns out that implementing the search for steady state of the logical problem in the proposed framework can be interpreted as Newton search for the "truth equilibrium". In more complex cases, when the space of the entities cannot be written as a one-dimensional continuum, if there are branches, etc., one has a *network* of logic interactions, and the gradient interpretation has to be abandoned (or somehow generalized). The conclusion here is that gradients seem to pop up automatically at least in some complex enough functional knowledge representations.

Even though the above example is highly hypothetical, there are some interesting connections to real world. For example, in ant colonies, it has been recognized that the flow of activity (the number of ants within a time period) is approximately proportional to the *pheromone gradients* along that path!

3.4 Additional application fields

When employing simple models, one of the advantages is that *there are more systems and environments than there are possible model structures.* This means that there often are analogues available — one just has to interpret concepts in another way.

Above, the discussions involved abstract *information balances.* The balances can also be concrete, like *mass balances,* etc. Indeed, the above approaches could directly be applied for modeling and analysis of industrial control systems. The system state is captured in the state variables (including the states of the controllers and controlled processes alike), and the force is, of course, the control action driving the system towards the intended balance: Stabilizing control is what is the key point in automation. This means that to model such systems in a cybernetic way, the control and the state need to be learned together. Remember that there are always feedback loops in complex automation systems; this is a painstaking problem from the point of view of traditional SISO (single input, single output) approaches to system identification. Now, on the other hand, the

closed loop problem is no more an issue: The system with feedbacks is a complete cybernetic system, and it can be modeled accordingly.

The differential-oriented interpretation of x makes it easy to see applications also in mechanics — indeed, one could speak of "cybernetic mechanics". Assuming that elements in x can be interpreted as velocities (or angular velocities), the energy criterion presented in [12] can have new relevance:

$$\mathcal{J}(x) = \frac{1}{2} x^T \mathrm{E}\{\bar{x}\bar{x}^T\} x - x^T \mathrm{E}\{\bar{x}u^T\} u. \tag{16}$$

If $\mathrm{E}\{\bar{x}\bar{x}^T\}$ is interpreted as an inertia matrix, so that masses and inertial moments are combined in the same matrix, the first term has the interpretation of *kinetic energy.* Comparing to the derivations in *Lagrangian mechanics* (for example, see [6]), there are differences — the latter term cannot be interpreted in the standard way as potential enargy. If u is the vector of external forces and torques affecting the system, the latter term defines *viscous friction.* Perhaps the cybernetic considerations can be applied for optimizing constructions of mechanical systems?

4 Models of mental machinery

From the point of view of applying systemic thinking to modeling of cognitive processes, one of the problems has been that modeling consists of engineering techniques, and cognitive science consists mostly of high-level hypotheses; these two worlds have been incompatible. Applying the modern cybernetic visions the philosophical and pragmatic realms coincide. In this latter part of the paper, it is illustrated what this claim means in practice. These discussions go far beyond the original goal of only studying the properties of informative data as seen from the strictly technical point of view; here the principles of the data processing machinery are studied no matter if the machinery is man-made or natural.

4.1 Critique of pure data

It was explained above that cybernetic models reflect their environment. Neural machinery is cybernetic, and so is the cognitive machinery (as will be explained later). There is optimality (in terms of exhausting available activation reasonably), and this promises that the original system and the resulting model are qualitatively equal. Now, if a "synthetic brain" is constructed applying the same principles, the emerging data structures must again be mirror images of the environment, and, further, they must be mirror images of the mental representations (see Fig. 2). The problem is that such functional structures cannot be directly mapped from a domain to another, and the cybernetic contents cannot be explicitly coded.

Heinz von Foerster coined the term *second-order cybernetics,* meaning that the observer of a system must be taken into account as another cybernetic system [5]. Indeed, if the ontology process (or generation of information in the system) is

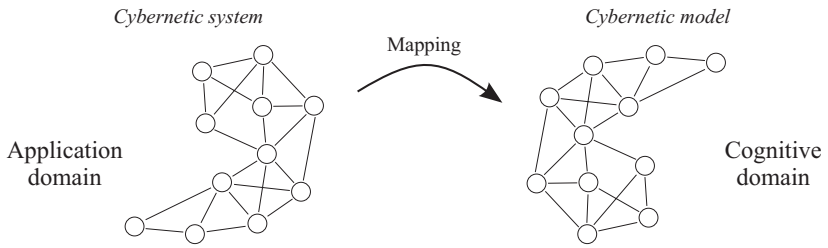

Fig. 2. Cognitive model triying to reflect reality

cybernetic, also the epistemic process (or gaining understanding of that information) should be cybernetic. But contrary to Foerster's vision, this second order nature does not complicate things — it *simplifies* them. It seems that Foerster thought that cybernetic systems always have to be interpreted by humans; as discussed above, the case is completely opposite. As semantics can be linked in the systems directly, the human can be liberated from the loop altogether: Also the human cognitive process can be simulated outside the brain.

It seems that Immanuel Kant was the first cognition theorist — and in some respects it seems that his intuitions have not been surpassed. He also seems to be the founder of *constructivistic* approaches to cognition[4]. As observed by Kant, there are not too many hardwired structures in the brain; only the principles of how new structures are constructed are fixed. Observations alone are not enough, something has to be assumed about the world and about the processing machinery. About the world Kant says that there are two basic dimensions of the world, space and time, that are essential for a human to be capable of constructing a perception. The observations have to be bound to such quantities to make them mentally modellable. Indeed, this is in one-to-one correspondence with the above studies: Spatial and temporal dimensions are manifested as relational and causal relationships among entities. About the functionalities in the brain, Kant proposes 12 distinct *categories* that are the minimum conceptual apparatus for making sense of the world; it seems that these assumptions can be relaxed.

4.2 Associative semantics

The above visions of different kinds of semantic data need to be given more content. It was claimed that in a simple case semantics can be based on static spatial context, and in more complex cases on dynamic spatio-temporal "extended context". First, the simpler case is studied, giving a practical example. In such a domain where the relationships are based on connotations and associations, no matter what is the nature of these connections, the entities can be called *casual.*

[4] Immanuel Kant can be said to be the founder of "social cybernetics", too: His *categorical imperative* defines the principle of how an agent in a cybernetic system should behave so that the overall system behavior were cybernetically optimized (indeed, in successful religions, like in Christianity, similar ideas of social feedback are also emphasized)

As an example of a such system, a document modeling environment was implemented [11]. In this application, contextual similarities were searched among textual documents, and they were stored in a cybernetically optimized model. Vectors u contained the document "fingerprints", that is, the histograms of words within the documents. After convergence, the columns of φ represented the "generalized keywords" assumedly being relevant for characterizing the documents and distinguishing between them. In technical terms, a sparse coded representation based on principal components was constructed for the data; because of this, the converged representations were statistically more or less independent of each other. The elements in $\bar{x}$ revealed the relevances of those generalized keywords when explaining a specific fingerprint u. The resulting model looked like a higher-level table of contents into the body of documents. A practical search engine was implemented where the above views were employed. In this application, learning of structures was based on explicit sparsity pursuit technique rather than on the nominal Hebbian/anti-Hebbian learning.

When seen from the semantic point of view, the network among generalized keywords in the above case purely represented non-causal associative structures: In this case, all entities have the same semantic "dimension". Such semantic networks could theoretically best be discussed in the framework of *fuzzy subsets* or *relevance networks* (see [14]).

It can be assumed that underneath the document data there is a cybernetic system of *memes* that is reflected in the terminology and contextual correlations among documents. The grounding of semantics in this experiment was left floating, there were no any actual inputs; words determined document contents — but, on the other hand, it can be said that the contents of the documents gave new flavour to the words, and one could have implemented a truly cybernetic environment with words and documents determining each other, associations balancing each other. To reach homogeneity among documents and words as information entities, the input should have been augmented with some document identification entries (appropriately scaled).

The words and documents in the above case constitute an integrated interacting associative medium where concepts and "metaconcepts" (or the documents) determine each other's interpretation and meaning. In [11] it is assumed that this kind of associative network would be enough to constitute a basis for *expertise.* However, expert knowledge is not a static model of the domain field; it is a model with "flux structures", where causalities and functional connections play a central role. Reasoning is not simple associative regression.

How to make the above document model truly cybernetic so that such causal structures were also represented? The easiest way would be to add auxiliary information among the input data vector u: For example, the "memetic flow" could be modeled so that only links from the newer documents to older ones were taken into account (in the other direction it is assumed that there is no correlation, the connections are set explicitly to zero). The flow of ideas is, of course, also reflected in citations, or links to other documents. But, as explained in the first part of this paper, the essence of a cybernetic system can be expressed

also in terms of tensions. Next, an example of more "expert-like" representation is presented following this intuition.

4.3 Functional chunks

The abstract discussions in Sec. 2 can be given some semantic content best by studying an example. In this simulation, *chess configurations* were stored in a cybernetic model applying the presented guidelines.

Preliminary experiments were carried out already in [10]. The framework was intentionally made cognitivistic there: Piece configurations on the chess board were "shown" to the computer, and its task was to "recall" the configuration. The underlying assumption about the recall process was that the short-term memory (STM) contains references to chess-specific long-term memory (LTM) elements, or *chunks* (for chunking, see [3]). These chunks are used to reconstruct the piece configuration. It has been recognized that there are capacity limitations to STM: Only some 4–8 chunks can be engaged at a time. Thus, the board cannot be exactly reconstructed; the essence of chess expertise is that the set of chunks is optimized so that best possible reconstruction is reached. In the computer implementation, the chunks were the features φ_i, and they were optimized applying the sparse coded PCA structures. The sparsity level of the representation was determined by the assumed STM capacity. And, indeed, the simulations with the model resulted in very similar results that have been observed with human test subjects.

However, those simulations were plagued by the problems of insufficiently rich data — just as is explained in Sec. 2. Real chess expertise is not about remembering, or storing and recalling the past; it is more like *constructing the future* that expertise is about! To implement *functional chunks* rather than static ones, the piece configurations need to be linked with the *flow of the game.*

In chess, there are different possible paths from the beginning to the end, as determined by the rules of the game. Following the above discussions, it can be said that one starts from the top of the energy function mountain, proceeding along the slopes, or gradients. As long as there are possibilities, or hidden tensions in the piece configuration, there are slopes, or forces, so that one can get to lower energy levels. However, the tensions never completely vanish, there are preferred directions (even when there are only two kings left on the board). If some moves are not reasonable, or, better, if they are not relevant, there is no force in that direction, whereas if a move is stringent, the force is strong.

This force-based intuition can be applied to modeling chess not in a static but in a truly expert-like way: The local balances (or balancing forces) in different piece configurations have to be modeled. A chess expert could give good data for modeling purposes, explicitly being capable of telling where there are major gradients. However, now there is no expert available, only data from actual games. The gradients can be locally approximated — applying mathematical intuitions — by somehow substracting successive states, or configuration representations, from each other. The problem is that in a single game just one path realization

takes place, no matter how stringent it is; seen from outside the difference between successive configurations gives a narrow view to the "chesslikeness" of the move. When there is plenty of data from various more or less similar games, this problem is not so acute: If a specific move seems to be consistently encountered, that direction can gradually be made more emphasized in the gradient estimate. Whereas the rules of chess determine the landscape of possible gradients, the actual games give information of the relevant ones.

Of course, statistical approaches to modeling exclusive moves can be questioned: The gradients corresponding to alternative moves are not additive. When modeling correlations, the OR structures change to AND structures — rather than selecting one alternative, the model gives the spectrum of alternatives, appropriately weighted. Intuitively, this is appealing: When a visible configuration is matched against the model, a topographic map of causal forces is associatively constructed, revealing the "hot spots" on the board.

When implementing the simulation, "two-way gradients" were constructed, meaning that in each location, information of *from where* a piece disapperars and *in where* it again appears, was included in the data. The chess game is a highly nonlinear modeling problem, and the state (current configuration) has to be stored together with the gradient estimates. Because of efficient multivariate techniques, data dimension is not an acute problem, and sparse wasteful codings were used to make patterns maximally orthogonal, and to facilitate easier analyses: When representing piece configuration, there was available an entry for each of the piece types for each of the locations, meaning altogether $64 \cdot 12 = 768$ variables. The variables were binary, "1" meaning that the corresponding piece was in that location. Similarly, the two gradient estimates were coded as 768 dimensional vectors, so that the overall data dimension was $m = 3 \cdot 768$. No further data preprocessing was applied.

The model size in the experiments was $n = 100$, and the sparsity level (STM capacity) was $N = 5$: Each of the data samples was modeled as a linear combination of 5 most significant sparse components. Because of the extreme sparsity of the representations, an explicit sparsity pursuit algorithm (see [11]) was applied here, because localized (hierarchic) feature extraction is more efficient and better controllable than the strictly distributed cybernetic implementation. The model was linear, that is, no explicit *cut* function was applied (see [12]); nonlinearity was supplied by the sparse nature of the representations. As compared to [10], now only white moves were modeled. There were some 2500 data samples used in the training.

In Fig. 3, one (successful) case of configuration reconstruction is illustrated. In reconstruction, only the visible configuration was assumed to be available, and the entries corresponding to the gradient estimates in the data vectors were not matched — these vector elements were filled in associatively. The continuous-valued entries in the estimated configuration have been thresholded so that values above 0.5 are interpreted as the corresponding piece being in that location, whereas values below that are ignored (when interpreting the gradients, threshold was 0.1). The model seems to propose two alternative moves: Either one can

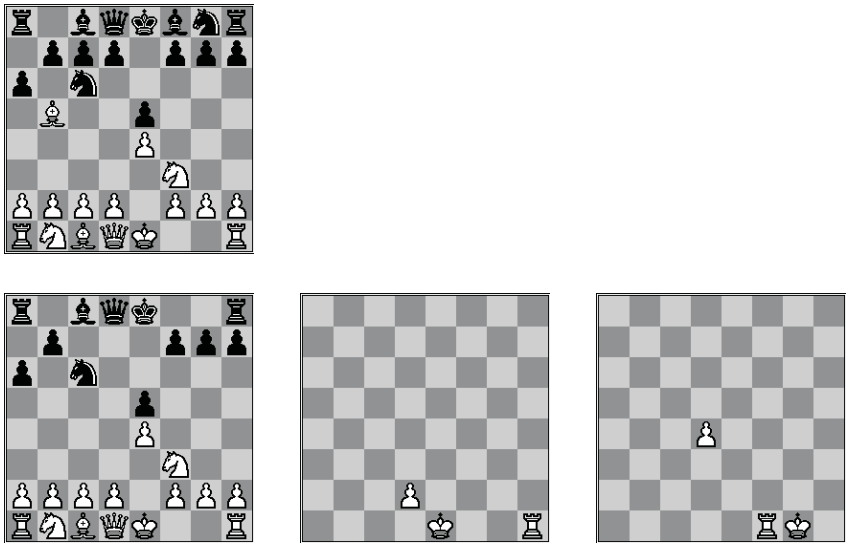

Fig. 3. Illustration of the functional chunks in chess. The topmost image represents the board that has been shown to the "test person", and the bottom row shows the reconstruction results: The recalled board on the left, the "from" pieces in the middle, and the "to" pieces on the right. This time, the two suggestions both are quite reasonable

move (simultaneously) the king and the rook, performing the castling operation, or one can move the pawn from d2 to d4. Truly, the "hot spots" are immediately detected without extra calculations. This kind of focusing of attention reminds the process that seems to be taking place when the real experts are involved.

However, from the point of view of assisting in real game, the reconstructions are typically much too poor, specially as the game proceeds beyond the standard openings (for example, the very essential bishop in Fig. 3 was ignored altogether). But perhaps chess as a domain with very strict rules is not a very good example of real-life expertise. The configurations cannot now be efficiently modeled applying associative techniques, evaluations of configurations being too much dependent of one single piece.

From the point of view of practical implementations, and also from the point of view of cognitivist plausibility, additional gradient inputs are difficult — for example, one needs to determine appropriate weighting among the inputs. Explicitly increasing data dimension does not sound very elegant, specially when the property of cybernetic systems should be optimality at every level. One should pursue light-weight solutions; causal structures can be implemented also directly if the platform supports it — and the neural platform *does.* Are there ways to reach automatic balancing for originally unbalanced observations?

4.4 Essence of "deep structures"?

Just as data, also *language* reflects the underlying real world in a different form. However, modeling of language seems to be among the biggest challenges of all in cognitive science (for example, in [2] where the unification of mental faculties is propagated, it is still admitted that perhaps linguistic faculties need completely different mechanisms). Still, natural language is the most natural way for humans to process and transfer structured information — providing a plausible explanation of language is the real testbench for any cognitive theory.

In [4], it is explained how there is a difference between the surface form of the language and the *deep structures.* The actual language contains complex syntax to represent the structural hierarchies of the deep structures in a "unidirectional", linguistically representable form. What is the essence of these underlying deep structures? The above discussions give a new answer to this question.

Again, one needs to distinguish between two different types of semantic entities. First, there is the spatially connected associative medium where the concepts are grounded based on naturalistic and contextual semantics; these concepts are in balance, as presented in Sec. 4.2. The second set of entities define some kind of transitions, changes, or causal relationships; they are declarative "on the fly" structures, explicitly "programmed". When a new concept is defined it is connected to its qualifiers. The concept of a "concept" needs to be understood so that indeed all new connections of prior concepts are new concepts here. In a way, a concept (abstract as well as concrete) is defined through the process of *ostension,* defining examples and other concepts that characterize it. The new structures can equally well refer to established concepts in the associative medium, or to the possibly very volatile new structures. The linked sequence of structures defines a causal path. Whereas the associative medium is cybernetically balanced, the newly created sequence of novel concepts is *not yet.*

In a way, a *story* defines a path in the causal semantic force field towards lower energy, as measured in terms of uninstantiated possibilities. As compared to the chess game, the number of alternative directions to go is immense. The ideal story balances all opposing forces, so that finally the holes in the story get filled. Until that, there is tension and dynamic flow within the structures. As compared to the extremely static linguistic syntax trees, the dynamic functional interpretation does have some added intuitive appeal (see Fig. 4).

The new concepts are also first all declarative (to facilitate further references to them, the new concepts also need to be labeled somehow). The concepts have arbitrary attributes, or links to other concepts. The explicitly determined attributes are initially the only inputs into that concept (additionally, there is a "-1" link to itself to assure stability). However, as soon as the new concept is established, the Hebbian/anti-Hebbian processes are activated: If the concept is active, and there is simultaneous activation in *lower-level* structures, the connection is sthrengthened according to the Hebbian law; simultaneous activation with same level entities results in inverse adaptation in the anti-Hebbian way. This means that finally (if the concept is activated enough often) it becomes

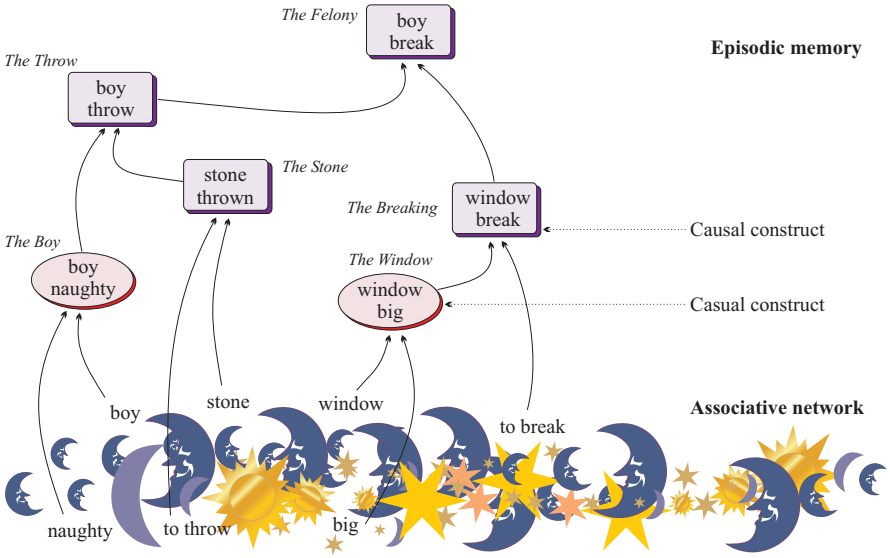

Fig. 4. "A naughty boy throws a stone. A big window breaks."

"swallowed" by the associative medium, interacting in the new cybernetic balance there. There also emerge direct links to lower-level "resources" — and if there once existed a declarative succession of same-level concepts, they start sharing those resources in a parallel manner, the temporally ordered master-slave structure disappearing. Mutual competition means that finally there will be only one associative concept standing for the whole sequence; this means that the representation becomes compressed and optimized if there are concepts with no differences whatsoever (in terms of their connections). The final dissolution into the associative medium takes a long time; there is a continuum from declarative to associative representations in the universe of concepts.

Originally the connections between sequential concepts is positive, and after adaptation the connections among concepts are negative — some kind of landmark in the process towards becoming "common knowledge" is perhaps when the mutual links cross the zero level (note that the connections to lower-level constructs become all the time stronger and stronger, so that the concept is still firmly grounded). If there are disturbances in the balanced system ("too active" concepts, etc.), they are soon compensated by inhibitory connections that get stronger.

This far in this context, it seems that all interesting phenomena and processes have been deemed as being *cybernetic.* So, is the deep structure of language cybernetic? Surprise — the answer here is *no.* Deep structures what comes to representing sequentially ordered linguistic utterances must be non-cybernetic. After becoming associative and cybernetically balanced, the contents of a structure cannot any more be decoded and explicated.

The deep structures are not only characteristic to language: It is claimed here that all ordered information is (originally) presented in this form, and it gives a solid basis for implementing complex functionalities. For example, a declarative feedforward reasoning system can be implemented in this framework (see [13]). As another example, study *learning by examples.* Assume that a new concept has been defined, but its input links have not been sufficiently instantiated (so that the concept "fires" inappropriately); the correct substance of the concept is given implicitly in terms of positive and negative examples of that concept. If the new concept is not too much more complex than the prior ones on the lower level, so that the new class is linearly separable in terms of other concepts, Hebbian learning adapts the weights among active neurons until correct classification is inevitably reached. To implement learning by examples, there is also no need for any structural assumptions in addition to the hierarchically ordered deep structure.

When the chess game is modeled in the deep structure form, the whole game is stored as a sequence of successive configurations. Because of the links among configurations, activating one of them simultaneously activates its successors — thus, the representation is again functional. It is not only so that configurations are linked; there is always the whole frame of reference that is stored. For example, related feelings — enthusiasm, fear, despair — can also be activated later in certain situations. There is also a huge number of memory units needed to store the chess player's memories; however, because there is only the fixed population of units available, less relevant ones will be reused (see later). And because it is also the associative matching that is taking place, the representations are compressed; a single unit can start representing various functionally similar configurations.

Also non-hierarchic information may be presented in a sequential form because of the properties of the available data processing machinery and data transfer channels. For example, the process of looking at an image consists of a sequence of eye fixations.

4.5 Mind and brain: Towards unifying principles?

The computer paradigm has been rooted in our understanding, and this makes us see all information procesing systems through the same viewpoint. However, it is evident that in the brain this analogue collapses altogether: There is no need to distinguish between hardware and software (or "wetware"), there is no need for a "central processing unit", or separate memory registers. Applying the above views, it turns out that the associative medium, the deep structures, and also the accompanying mental processes can be explained in terms of the *simple (nonlinear) perceptron model implementing Hebbian/anti-Hebbian learning, thus constituting a cybernetic system.* A single neuron is the atom of semantics, connecting other concepts into a balanced whole.

There is homogeneity what comes to mapping between the higher-level cognitivist concepts and neural constructs: The same basic construct, the neuron,

is only needed. First, it needs to be noted that no structural or functional difference between declarative casual and causal structures exists; this distinction had only a semantic role in the above discussions.

No centralized memory units are needed for any specific task, and there is no transfer of data between "registers" — logic functions need no logistic functions: A newly allocated neuron with its few links to prior concepts can be seen as a short-term memory element. An STM element changes into a long-term memory unit if it is *relevant* enough; and an LTM unit is "swallowed" by the associative medium if it is bound to other concepts densely enough. Using the traditional terminology, this process of becoming "common sense" can be called *shift from novice to expert.* There is no clear distinction between the declarative and associative representations. Indeed, all different kinds of hypothesized memory types can be emulated in the proposed framework with no essential extensions; the connections can be used also for explaining many AI concepts like *frames* and *schemas.* In technical terms, a converged associative structure is characterized by symmetric connections among entities, whereas declarative structures are characterized originally by unsymmetric connections (see next section).

Looking at the above discussions, it seems that some kind of an "operating system" is needed for constantly supplying free neurons to be allocated when new concepts are being defined. The operation of the neuro-cognitive system looks so goal-directed that some kind of an explicit organizer seems to be necessary. However, this is only an illusion: The operating system is distributed rather than centralized, being an emergent property of the neuronal competition for activation. There is no centralized "winner" selection among neurons; typically, many individual neurons can be simultaneously allocated for a single task, and only later the representation can be optimized because of mutual competition[5]. This competition for resources can be modeled applying cybernetic considerations. Now, there is a population of neurons rather than a fixed network to start with, just as studied in [12]. As seen from outside, the many competing entities make the system behavior mathematically better conditioned: rather than having to study individual, more or less random signals, only statistical average behaviors are needed. This distributedness of neuron allocation also makes it possible that simultaneous inputs can be given (subconscious) attention in a parallel manner.

It would seem that another task of the cognitive machinery where some centralized control seems necessary is that of "garbage collection", or freeing those neurons that are no more needed. Again, no organizer is needed here: Neurons that have too little overall activity (because of lack of active enough or *otherwise* relevant links), are free to be reused. Or, indeed, the neurons are not just passively exploited — they start actively creating new connections, or dendrites, towards the sources of activity. When a dendrite meets another neuron, a synapse is created in the junction (this has been shown to happen also in reality). There is no "10% activity rule" in the brain; the neurons are all

[5] This competition-like behavior among neurons was already observed by the Nobel laureate Gerald Edelman

the time fully loaded, searching for resources. There are no "free", completely unconnected neurons; they may just not be well "aligned" with other ones, thus receiving contradictory activations summing up to zero. There are also no timing or synchronization problems in the system: Whenever a neuron is in balance with its environment, it can start adapting according to other balanced activity levels that it observes in its environment.

When there is no input from the senses to the neurons, the neuronal activity pursuit makes them more sensitive, so that activity can be triggered also by random noise. The REMaining task of the hypothetical operating machinery is to connect neurons to other ones in a more or less random manner — and, again, this can be explained in completely local manner[6].

When studying semantically very loaded phenomena, one bangs into anthropocentric connotations everywhere. For example, above the term *competition* is dangerous: It sounds as if consciousness and free will were moved from the operating system level (mind) to the agent level (neurons). However, supplied with the Hebbian/anti-Hebbian learning principle, the neuron can do all tasks that are needed; what is more, as a cybernetic agent, it *cannot avoid* adapting accoding to the observations, no "motivations", etc., are needed. The underlying "Elan Vital" (why there is some adaptation instead of no adaptation) is neither a mystery: The emergent functionalities do give the system the evolutionary advantage. In the other end, there is the question of *consciousness;* there is not yet enough intuition available to say anything concrete about these higher level control issues. According to some definitions, consciousness is "consciousness of being conscious": If defined in terms of self-consciousness, self-self-consciousness, etc., one is again facing a deeply connected cybernetic feedback system.

Even though homogeneity of structures was emphasized above, it needs to be noted that the associative medium is *functionally* by no means homogeneous. There are different associative subsystems for different internally tightly coupled subdomains. Again, this organized-looking nature of the structures is just an emergent phenomenon: If different neurons are active together, as they will be if they represent related concepts, they become connected when applying the assumed learning principles. Actually, our way of seeing different domains as decoupled entities is a fallacy: There is a continuum from strongly connected neurons to less strong connections, and no strict boundaries between subsystems exist. As an example, study the decomposition of visual patterns: The low-level visual features constitute a tightly connected subsystem where alternative features (line segments, etc.) compete for activity; combining these into more complex feature structures is the next level — as seen by a human eye and reasoning style. However, there is no clear-cut boundary between layers, neither there are any well-defined "interfaces": There can be individual connections between the lower and highel level neurons.

[6] Of course, there also exist some global mechanisms, like sleep rhythms, etc., and different kinds of chemical levels (hormones and enzymes), controlling the behaviors of all neurons simultaneously in the system in a global manner

Many traditional cognitivist concepts need to be given new interpretations. For example, "STM memory capacity" is not a property of some centralized general-purpose processing element or memory structure; now it represents the maximum number of links that the neuron can simultneously instantiate when it is being allocated.

4.6 Neuro-cognitive model

To implement concrete models for evaluating and utilizing the above approaches to smart modeling of data, mathematically more compact representations are needed. The proposed cybernetic cognitive model has practically the same structure as the neural model presented in [12]. It is a streamlined version of that model that was presented in [13] (there the internal system dynamics could be unstable, resulting in emphasized sparsity). As a state-space model, the new structure looks like

$$\frac{d}{dt}\begin{pmatrix} u \\ x \end{pmatrix} = \left(\begin{array}{c|c} -I & \mathbf{0} \\ \hline \mathrm{E}\{\bar{x}f^T(\bar{u})\} & -\mathrm{F}\{\bar{x}f^T(\bar{x})\} \end{array}\right) f\left(\begin{pmatrix} u \\ x \end{pmatrix}\right) + \begin{pmatrix} u_{\mathrm{in}} \\ x_{\mathrm{in}} \end{pmatrix}. \tag{17}$$

There does not exist any *agent mathematics.* Even though the neural/cognitive network is not fully connected, one needs to employ the exhaustive matrix calculus; this means that the matrices become sparse. There is one row for each neuron, and matrix elements reveal how strongly it is connected to other neurons and inputs; the activations of the n concepts (neurons) are collected in the vector x. As compared to the model presented in [12], the main difference here is the interpretation of inputs: The actual inputs u have been included in the state vector, and there is now the augmented input vector with trivial input mapping matrix $B = I$. Inputs u_{in} represent the actual system inputs, signals coming from senses or other subsystems, whereas the role of x_{in} is to implement "handles" to the concepts. The augmented input means that also any of the concepts can be explicitly activated from outside; this makes "learning by being told" possible, in addition to the associative activation that is based on matching with u_{in}. In the recall phase, the inputs can be used for activating associations in the network, so that a *mental image* with its connotations can be waken up: When a concept is activated, those concepts that it is connected to also inherit some agitation, activation thus spreading in the associative structure.

In the model (17), an additional nonlinearity, or "activation function" is included; for the reasons explained in [13], *cut function* is applied. This nonlinearity gives rise also to new problems: For example, there can be no negative values in the output. To restore the expressional power, the output vector is decomposed into a higher-dimensional vector so that complex classifications can better be implemented. Indeed, the nonlinear function f represents *feature extraction:* In this case this means that the *positive and negative outputs have entries of their own* in the vector (see [12]). It is here assumed that there can exist two alternative neuron types: One is assumed to be activated with positive and one with negative signals. When both alternatives are included in the vectors, one can

emulate the whole population of neurons with different strategies in the same framework — only those with the most appropriate strategy will prosper[7].

If there are non-associative cyclic structures, the spread of activation can take a long time. However, only after the network has converged, after the steady-state vectors $\bar{x}$'s and $\bar{u}$'s are found, adaptation of the network weights is carried out. The correlation matrices in (17) are adapted applying the Hebbian and anti-Hebbian principles. The connections between inputs and neurons follow the Hebbian learning:

$$\frac{d\,\hat{\mathrm{E}}\{\bar{x}f^T(\bar{u})\}}{d\,t}(t) = -\lambda\,\hat{\mathrm{E}}\{\bar{x}f^T(\bar{u})\}(t) + \lambda\,\bar{x}f^T(\bar{u}). \tag{18}$$

Here, λ is the learning parameter ("forgetting factor"). To implement the discussions above, the neuron-to-neuron learning part becomes more complicated:

$$\frac{d\,\hat{\mathrm{F}}\{\bar{x}f^T(\bar{x})\}}{d\,t}(t) = -\lambda\,G \odot \hat{\mathrm{F}}\{\bar{x}f^T(\bar{x})\}(t) + \lambda\,K \odot \bar{x}f^T(\bar{x}). \tag{19}$$

The operator $\odot$ stands for elementwise multiplication, and K and G are *masking matrices*. Matrix K makes it possible to implement explicit sparsity, as well as plain principal component analysis. It also makes it possible to switch between Hebbian and anti-Hebbian learning, facilitating hierarchical structures among neurons (see [12]): Negative adaptation resulting in Hebbian learning implement forward shifting of information in a sequential manner, whereas positive adaptation results in anti-Hebbian competitive learning, implementing associative structures. The lerning mode makes a big difference — that is why it is possible to have various neurons taking care of a single specific task, the signs of K being randomly initialized; because of the competition for activation, the best combinations survive, other neurons starving to non-existence.

As compared to [12], the matrix G is something new; its role is to extend the model towards the declarative knowledge representations. Above it was assumed that sequential structures eventually become associative — however, this does not always seem to be the case: Some temporal structures seem to remain there forever unaffected. Zeros in G (and simultaneously in K) mean that in that synapse there is no adaptation whatsoever. This makes it possible to *encapsulate* episodic memory blocks. Note that only internally the adaptation is frozen; links to outside system are adapted consistently. In this sense one could speak of "superneurons", conglomerate neuron blocks with hardwired inner structure facilitating more complex dynamics than what is possible for the basic perceptron. An example of this is given in the following section.

Something needs to be said about how the data structures are maintained. Within the fixed matrix structure it is not very natural to implement brain-like neuron allocation; this can be simulated, however. If there are known to be neurons with special roles, it can be taken care of when initializing the corresponding

[7] Because the matrix A has dimension $(n+m) \times (2n+2m)$, the "identity matrix" in the upper left corner has to be "squeezed" appropriately

rows and columns of the correlation matrix. The new concept neurons are connected to their predecessors by filling in the corresponding slots on that row; typically, the corresponding column will also be non-zero to facilitate bidirectional two-way activations (the column elements must be non-zero and negative to assure stability). If there is plenty of activity available in some neuronal regions, *diffusion* among neurons can be implemented, so that *self-organization* takes place (see [14]). There is no need to explicitly free unused neurons; self-organization "pulls" neurons towards higher activity.

4.7 The way up and the way down

The Heraclitus aporia about the *way up and the way down* can also be given another interpretation in the cognitive setting. It is not only so that the way "up" is the world and its processes, and the way "down" is the perception machinery; the idea can also be seen in an opposite direction: The way up is the process of constructing perceptions, and the way down is how these perceptions are applied to affect the world. This alternative interpretation means that the processing of sensory signals and the construction of motoric signals must be based on the same principles; these processes are, again, mirror images of each other. This vision was studied already in [7].

Following the above neurons-based view up from observation to perception, how could the way down to muscles be explained? This is a direct application of the above mentioned sequential block: The key point is to master sequences of time instances. To analyze this case more closely, one needs to remember that *delay* can be modeled in terms of partial differential equations as

$$\frac{\partial x}{\partial t} = -v \frac{\partial x}{\partial \xi}. \tag{20}$$

Here, t is the time variable, whereas ξ is the spatial variable; $x(t, \xi)$ can be any function of time and location as long as it can be written as a function of only one variable, so that $x(t, \xi) = x'(\xi - vt)$, that is, the function travels at speed v along the ξ axis. This infinite-dimensinal model can be approximated using n state variables as

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \frac{v}{n}\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \frac{v}{n}\begin{pmatrix} u_{\text{in}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{21}$$

If n is large, the signal u_{in} traverses through the grid of variables x_i. Assuming that u_{in} is a start impulse initiating some sequence, any function $y(t)$ can be approximated to arbitrary accuracy using these basis functions:

$$y(t) \approx \begin{pmatrix} f_1 \cdots f_n \end{pmatrix} x(t). \tag{22}$$

This means that any function $y(t)$ can be implemented as a sequence of specially connected neurons (see Fig. 5). The delay structure does not change but the

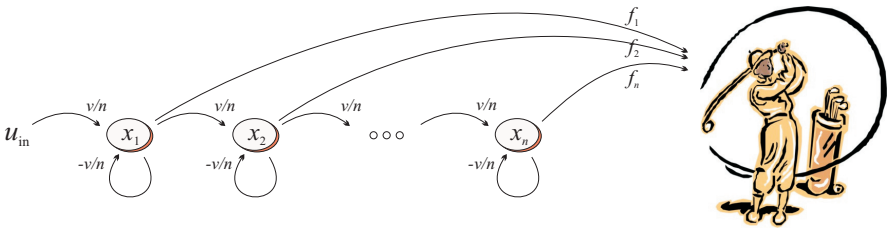

Fig. 5. Any pattern of excitation can be implemented to arbitrary accuracy

connections from this sequential block to the outside world, that is, the weights f_i, can be adapted. This adaptation can be based on Hebbian learning: If a muscle follows some activation pattern, the neurons learn that behavior.

5 Conclusion

If a tree falls in the forest and nobody is around to hear it, does it make a sound? Does an objective reality exist, are there systems outside human minds?

In tomorrow's smart systems, systems must become *mature,* and become independent of humans. Clever adaptation according to the observations is necessary — this means autonomous modeling of the environment in such agents. Extending the views of Nietsche, one could say that anthropocentric modeling practices may soon become obsolete: "Oversystems" need no more humans.

The discussions above are far from conclusive; however, it can be claimed that such studies are a *Prolegomena to any future metaphysics that will be able to present itself as a Complex Systems Science.*

References

1. Alander, J.: Evolutionary algorithms in cybernetics — a bibliographical overview. Elsewhere in this *Proceedings of the Finnish Artificial Intelligence Conference STeP 2004,* Vantaa, Finland, September 1–3, 2004.
2. Anderson, J.R.: *The Architecture of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1983.
3. Chase, W. G. and Simon H. A.: The minds eye in chess. In W. Chase (ed.), Visual information processing. Academic Press, New York, 1973.
4. Chomsky, N.: *Syntactic Structures.* Mouton, The Hague, 1957 (Reprint Berlin and New York, 1985)
5. von Foerster, H.: The Logical Structure of Environment and its Internal Representation. *International Design Conference,* Aspen 1962.
6. Goldstein, H.: *Classical Mechanics.* Addison-Wesley, 1980 (second edition).
7. Haavisto, O. and Hyötyniemi, H.: Life-Like Modeling and Control of a Biped Robot — Lessons Learned. Elsewhere in this *Proceedings of the Finnish Artificial Intelligence Conference STeP 2004,* Vantaa, Finland, September 1–3, 2004.
8. Hannon, B. and Ruth, M.: *Dynamic modeling.* Springer-Verlag, New York, 1994.

9. Haykin, S.: *Neural Networks. A Comprehensive Foundation.* Macmillan College Publishing, New York, 1994.
10. Hyötyniemi, H. and Saariluoma, P.: Chess — Beyond the Rules. In Timo Honkela (ed.): *Games, Computers and People (Pelit, tietokone ja ihminen),* Finnish Artificial Intelligence Society, Helsinki, Finland, 1999, pp. 100-112.
11. Hyötyniemi, H.: HUTCH Model in Information Structuring. *Proceedings of the Finnish Artificial Intelligence Conference STeP'02,* December 16–17, 2002, Oulu, Finland, pp. 241-255.
12. Hyötyniemi, H.: Cybernetics — Towards a Unified Model? Submitted to the *Finnish Artificial Intelligence Conference (STeP'04),* Vantaa, Finland (September 2004).
13. H. Hyötyniemi: Hebbian and Anti-Hebbian Learning: System Theoretic Approach. Submitted to *Neural Networks* (2004).

14. *Additional material on cybernetics will be available in public domain in near future at* `http://www.control.hut.fi/cybernetics`.