Neural Networks, Penalty Logic and Optimality Theory

Reinhard Blutner
University of Amsterdam

Ever since the discovery of neural networks, there has been a controversy between two modes of information processing. On the one hand, symbolic systems have proven indispensable for our understanding of higher intelligence, especially when cognitive domains like language and reasoning are examined. On the other hand, it is a matter of fact that intelligence resides in the brain, where computation appears to be organized by numerical and statistical principles and where a parallel distributed architecture is appropriate. The present claim is in line with researchers like Paul Smolensky and Peter Gärdenfors and suggests that this controversy can be resolved by a unified theory of cognition – one that integrates both aspects of cognition and assigns the proper roles to symbolic computation and numerical neural computation.

The overall goal in this contribution is to discuss formal systems that are suitable for grounding the formal basis for such a unified theory. It is suggested that the instruments of modern logic and model theoretic semantics are appropriate for analyzing certain aspects of dynamical systems like inferring and learning in neural networks. Hence, I suggest that an active dialogue between the traditional symbolic approaches to logic, information and language and the connectionist paradigm is possible and fruitful. An essential component of this dialogue refers to Optimality Theory (OT) – taken as a theory that likewise aims to overcome the gap between symbolic and neuronal systems. In the light of the proposed logical analysis notions like recoverability and bidirection are explained, and likewise the problem of founding a strict constraint hierarchy is discussed. Moreover, a claim is made for developing an "embodied" OT closing the gap between symbolic representation and embodied cognition.

1 Introduction

To date, progress in cognitive neuroscience has been hindered by the enormity of the gap between our understanding of some low-level properties of the brain on the one hand, and of some very high-level properties of the mind on the other hand. Research on parallel distributed processing and neural networks (connectionist paradigm) has tried to reduce this gap but was only partially successful. A main characteristic of mainstream connectionism is its *eliminative* character, i.e. the idea that the basic architecture of symbolism (including its

crucial concepts such as representations, rules, compositionality, and modularity) has to be replaced by the concepts of neural networks (cf. Churchland, 1986). In this way, the main advantage of traditional symbolism – the transparency and relative simplicity of descriptions and explanations – are likewise eliminated.

In contrast, there are other researchers who like to play down the neuronal perspective as an issue of implementation. Representatives of this position are, *inter alia*, Fodor and Pylyshyn (1988), who insist that the proper role of connectionism in cognitive science is merely to *implement* existing symbolic theory. According to this view, the *systematicity* of our linguistic competence can be explained only by assuming a classical, symbolist architecture of cognition. If this position reflects an adequate research programme, then the task of overcoming the gap between symbolism and its neural embodiment is not really important for the understanding of our higher-level cognitive abilities.

The methodological position pursued in this article is an *integrative* position. It claims that both modes of computation – symbolic and neural – are theoretically justified and equally important and that there is no need to eliminate one of them. In the case under discussion the point is to assume that symbols and symbol processing are a macro-level description of what is considered as connectionist system at the micro level. This position is not unlike the one taken in theoretical physics, relating, for example, thermodynamics and statistical physics, or, in a slightly different way, Newtonian mechanics and quantum mechanics. Hence, the idea is that the symbolic and the subsymbolic mode of computation can be integrated within a unified theory of cognition. If successful, this theory is able to overcome the gap between the two modes of computation and it assigns the proper roles to symbolic, neural and statistical computation (Balkenius & Gärdenfors, 1991; Smolensky, 1995; Kokinov, 1997; Blutner, 2004; Graben, 2004; Smolensky & Legendre, to appear).

There is a second methodological aspect that relates to the status of theoretical models in integrated research. My primary aim is the demonstration that the tools of logic and algebraic semantics are useful for understanding the emergent properties of neural networks dynamics. However, the dynamics of real neural networks is rather complicated. These systems are perhaps among the most complex known to science. And it is completely unrealistic to understand the emergent properties of such systems by trying to model in detail all what is known about the basic principles of neural operation and causal mechanisms of individual nerve cells. Rather, radical simplification is in order even if these simplifications appear completely unrealistic. These simplifications may lead to different theoretical models which make different views explicit, and this makes it easier to structure the debate for or against a certain position. Theoretical models bring out the hidden assumptions of an approach, particularly with

respect to the elementary neural mechanisms that are required. Moreover, they help to assess the plausibility of certain assumptions, for example with respect to the assumed network architecture. They may invite the construction of new models that make another view and other functional determinants explicit. Even if it is not possible to collect the necessary empirical data to make the model predictions empirically grounded, a lot can still be learned about the causal determinants of certain forms of behavior. Finally, even oversimplified theoretical models may suggest new experiments for empirical data collection.

A third methodological aspect concerns a potential misunderstanding. In the following I will pursue a certain kind of propositional default logic to describe inferences in neural networks. This might suggest that certain logical systems get a deeper justification in terms of neural processing, or it might even suggest that I'm proposing a neural underpinning of certain types of natural reasoning. Hence, it might appear as if we are running in a neuro-cognitive Frege-fallacy by seeing logic as part of cognitive neuroscience. However, such conclusions are unjustified. I only suggest to consider the proposed logical system as a kind of *meta-language* which is useful for modelling certain constraint-based symbolic systems. This is analogous to the use of Prolog as a logical programming language. Without doubt, Prolog can be used for many different applications starting from the modelling of parsing and natural language comprehension and going on to the modelling of planning mechanisms and the abilities of logical inference agents. Nobody would suggest that these applications – if successful – give a deeper justification for Prolog as part of Cognitive Linguistics (at least if we reject the strong view of Artificial Intelligence; see Searle (1980)). In a similar way, the present logical system can be used for many different purposes. This becomes pretty clear when we enlighten the close connection to Optimality Theory (OT) – a general framework which was introduced by Prince & Smolensky (1993/2004) for describing constraint interaction in Generative Grammar.

In the following I will address the issue of formal tools and logical systems which are suitable for grounding the basis for a unified theory of cognition, and I will suggest that an active dialogue between the traditional symbolic approaches to logic, information and language and the connectionist paradigm is possible and fruitful. An essential component of this dialogue refers to OT (Prince & Smolensky, 1993/2004) – taken as a theory that likewise aims to overcome the gap between symbolic and neural systems.

Section 2 introduces symmetric neural networks and explains their basic properties. The idea of inferences in neural networks is explained in Section 3. The developed inferential notion rest on the (non-symbolic) concept of information states and is adequate for describing how neuron activities spread through a symmetric network. Section 4 discusses Penalty Logic – a logic that

was introduced by Pinkas (1995) in order to demonstrate what kind of logical systems symmetric networks can implement. In Section 5 a logic called Penalty/Reward logic is introduced and it is shown that such logic is an adequate tool for dealing with underspecification and conceptual enrichment in symmetric networks. In Section 6 I will discuss the relations to OT, and Section 7 draws some conclusions and shows the connection to recent efforts toward developing an embodied view of cognition.

2 Symmetric networks

Connectionist systems aim at modelling aspects of the nervous system on an abstract computational level. (Good introductions are given in McClelland & Rumelhart, 1986; Rojas, 1996; Bechtel, 2002). The central concept in a connectionist system is the individual unit ('node') which models the functionality of a neuron or a group of neurons. In fact, the units/nodes of most connectionist models are vastly simpler than real neurons. However, such networks can behave with surprising complexity and subtlety. This is because processing is occurring in parallel and usually interactively. In many cases, the way the units are connected is much more important for the behaviour of the complete system than the details of the individual units.

In the following we will assume that the individual units of a connectionist network correspond to larger groups of neurons, sometimes called columns, pools or assemblies (Hebb, 1949; Feldman & Ballard, 1982; Maass, 1999; Wennekers & Palm, 2000). A central idea of the assembly concept is that assemblies can overlap, meaning that one and the same neuron can be part of different assemblies. The organization of assemblies is done according to functional criteria and can be different for different functional contexts. Necessary conditions for constituting an assembly are strong internal couplings within the assembly.

The simplest form of describing the activation dynamics of single units is to assume a nonlinear function that yields the (average) firing rate of the unit given the sum potential of the unit. This sum potential can be calculated by weighted linear combinations of the firing rates of the incoming units. In the present approximation it goes without calculating the full action potentials (spikes). All that is needed are the firing rates of the units, which are directly transferred to the other cells. It has been argued that this method yields a valid approximation of realistic spiking behaviour under certain conditions (for details, see Maass, 1999; Wennekers, 1999; Gerstner & Kistler, 2002). However, it has also been argued that simple rate-based models are not sufficient to model information processing in neuronal systems. There is increasing evidence that the information transferred by a unit consists not only

in the average firing rate but also includes the phase of the spiking functions. This might be relevant for explaining binding by synchronization (e.g. von der Malsburg, 1981; Shastri & Ajjanagadde, 1993; Singer & Gray, 1995). In the following I will simply ignore this complication.[1]

There are different kinds of connectionist architectures. In *multilayer perceptrons*, for instance, we have several layers of nodes (typically an input layer, one or more layers of hidden nodes, and an output layer). A fundamental characteristic of these networks is that they are *feedforward* networks, that means that units at level i may not affect the activity of units at levels lower than i. In typical cases there are only connections from level i to level i+1. In contrast to feedforward networks, *recurrent networks* allow connections in both directions. A nice property of such networks is that they are able to gather and utilize information about a sequence of activations. Further, some types of recurrent nets can be used for modelling associative memories. If we consider how activation spreads out we find that feedforward networks always stabilize. In contrast, there are some recurrent networks that never stabilize. Rather, they behave as chaotic systems that oscillate between different states of activation.

One particular type of recurrent networks is a *symmetric network*, which is also called a *Hopfield network* (Hopfield 1982). Such networks always stabilize. Hopfield proved that by demonstrating the analogy between this sort of networks and the physical system of spin glasses and by showing that one could calculate a very useful measure of the overall state of the network that was equivalent to the measure of energy in the spin glass system. A Hopfield net tends to move toward a state of equilibrium that is equivalent to a state of lowest energy in a thermodynamic system.

As mentioned already, neural networks can be considered systems of connected units. Each unit has a certain *working range of* activity, which can be represented by an interval [*a*, *b*] if an analogous unit is assumed (e.g. Hopfield, 1984; Hopfield & Tank, 1985); *a* indicates the minimal firing rate of the unit and *b* indicates the maximal firing rate. Usual choices for the working range of a node are the interval [0, 1] (e.g. Balkenius & Gärdenfors, 1991; Pinkas, 1995) or the interval [-1, +1] (Blutner, 2004). In the latter case the value 0 can be taken as indicating the resting rate. Though neurons with different working ranges can be assumed to be basically equivalent (supposing the thresholds are adapted appropriately), there may be differences (i) due to the interpretation of the activations, (ii) due to the simplicity of the resulting equations, and (iii) due to the stipulation of different discrete subsets when it comes to the introduction of

[1] Some authors doubt that "binding by synchronization" is really such a realistic solution to the binding problem as it often is suggested. For instance, Palm & Wennekers (1997) argue that also other mechanisms are thinkable based on purely rate-based information.

logical values. The discrete values typically taken are $\{0, 1\} \subset [0, 1]$ in the first case (classical binary logic) and $\{-1, 0, +1\} \subset [-1, +1]$ in the second case (tree-valued logic).

A possible state *s* of the system describes the activities of each node: $s \in [a, b]^n$, with n = the number of units. A possible *configuration* of the network is characterized by a *connection matrix w*. Hopfield networks are defined by symmetric configurations and zero diagonals ($-\infty < w_{ij} < +\infty$, $w_{ij} = w_{ji}$, $w_{ii} = 0$). That means node i has the same effect on node j as node j has on node i, and the nodes don't affect themselves.[2] The *fast dynamics* describes how node activities spread through that network. In the simplest case this is described by the following *update function*:

(1) $f(s)_i = \theta(\Sigma_j \; w_{ij} \, s_j)$ (θ a nonlinear function, typically a step function or a sigmoid function).

Equation (1) describes a nonlinear threshold unit. This activation rule is the same as that of Rosenblatt's perceptron. It is applied many times to each unit. Hopfield (1982) employed an *asynchronous* update procedure in which each unit, at its own randomly determined times, would update its activation (depending on its current net input).[3]

Using the interval [0, 1] as working range of a unit, Balkenius & Gärdenfors (1991) have argued that the set $S = [0, 1]^n$ of activation states of a network with n units can be partially ordered in accordance with their informational content. Assuming that the vector $\mathbf{0} = \langle 0, 0, \ldots, 0 \rangle$ represents a

[2] It is often mentioned that these assumptions are highly implausible when taking the units of the network as real neurons. It is not clear why real networks should be symmetric and irreflexive. If the assembly idea comes in, we can overcome this problem since it is plausible to assume that the formation of assemblies happens under the pressure of stabilisation, and this might be one of the reasons for symmetry and irreflexivity.

Some people doubt the plausibility of the 'neuron doctrine'. Based on the finding that in the cerebral cortex the majority of neurons have only dendrites and the axons are missing there (this contrasts with the preripheral nervous system system where almost every neuron has an axon) (Jibu & Yasue, 1995, p. 100ff). Hence, it has been argued that the working of the cerebral cortex can be better understood by certain microscopic physical processes taking place in the sophisticated network of dendrites of neurons without axons, that is, in the dendritic network (Pribram, 1991; Jibu & Yasue, 1995). The spin-glass model (or, equivalently, the Hopfield network) can be seen as a first approximation to the dendritic network (Jibu & Yasue, 1995). Hence, Hopfield networks can be seen as a good starting point for modelling brain activity independent of whether we accept the neuron doctrine or not.

[3] The use of asynchronous updates helps to prevent the network from falling into unstable oscillations.

scheme with minimal informational content and that the vector vector **1** = <1, 1, …, 1> represents maximal informational content, then the following ordering can be seen as reflecting greater *positive* informational content:

(2) $s \geq t$ iff $s_i \geq t_i \geq 0$, for all $1 \leq i \leq n$

We call this interpretation of the activation states which is based on the ordering (2) the *Boolean option*.[4]

Sometimes it is useful to assume that both endpoints of the unit's working range carry maximal information and one value in the centre of the scale carries minimal information. The plausibility of such a choice was mentioned by Balkenius & Gärdenfors (1991). These authors suggested to take both **0** and **1** as states of maximal information and to assume that there is a resting state ½ that represents minimal information. Unfortunately, they didn't work out this proposal.

In Blutner (2004) the working range of each unit is stipulated to be [-1, +1]; the activations +1 and -1 indicate maximal specification; the resting activation 0 indicates (complete) underspecification. Generalizing Balkenius & Gärdenfors' (1991) idea, the set $S = [-1, +1]^n$ of activation states can be partially ordered in accordance with their informational content:

(3) $s \geq t$ iff $s_i \geq t_i \geq 0$ or $s_i \leq t_i \leq 0$, for all $1 \leq i \leq n$. (Read $s \geq t$ as *s is at least as specific as t*)

It is a simple exercise to show that the poset $<S, \geq>$ doesn't form a lattice yet. However, it can be extended to a lattice by introducing a set $\perp$ of *impossible activation states*: $\perp = \{s: s_i = nil \text{ for } 1 \leq i \leq n\}$, where *nil* designates the "impossible" activation of an unit, i.e. a clash between positive and negative activation (for details, see Blutner, 2004). Further, it is possible to show that the extended poset of activation states $<S \cup \perp, \geq>$ forms a DeMorgan lattice. This allows us to interpret these activation states as propositional objects ('information states'). It is convenient to call this interpretation of the activation states the *DeMorgan option*.[5]

Symmetric networks may be viewed as searching for the local minima of a quadratic function called an energy function (or Ljapunov function). The

[4] $<S, \geq>$ forms a Boolean algebra if the underlying neural network is binary (cf. Balkenius & Gärdenfors, 1991)

[5] There is another option for modelling activation states: ortho-algebras. The interested reader could consult www.quantum-cognition.de in order to learn more about this alternative option. Unfortunately, space limitation forbids us to discuss the ortho-algebraic approach in the present article.

important fact proven in Hopfield (1982) says that in the case of asynchronous (non-deterministic) updates, the function

(4) $E(s) = -\sum_{i>j} w_{ij} s_i s_j$

is a Ljapunov function of the dynamic system described by the equation in (1)[6]; i.e., when the activation state of the network changes, E can either decrease or remain the same. Hence, the output states $\lim_{n\to\infty} f^n(s)$ can be characterized as *the local minima* of the Ljapunov-function. A consequence of this result is that all states *s* in a symmetric network develop under asynchronous updating into *resonances*, i.e. into stable states of the network that *attract* other states (for details, see Cohen & Grossberg, 1983).

Usually, asynchronous updating results in stable states that are local but not global minima of the energy function E. The Boltzman machine (Hinton & Sejnowski, 1983; Hinton & Sejnowski, 1986) is a modification of the Hopfield network that realizes the *global* minima, i.e. their output states $\lim_{n\to\infty} f^n(s)$ can be characterized as *the global minima* of the Ljapunov-function. Like the Hopfield net, the Boltzman machine updates its units by means of an asynchronous update procedure. However, it employs a stochastic activation function rather than a deterministic one. This activation function can be considered to realize some stochastic noise ("faults") in a decreasing rate during the processing of a single pattern.[7]

Updating an information state s may result in an information state f...f(s) that does not include the information of s. However, if we want to handle logical inferences, it is important to interpret updating as specification. That means we have to make sure that the initial state *s* has to be informationally included in the resulting update. Hence, we have to "clamp" s somehow in the network. A technical way to do that has been proposed by Balkenius & Gärdenfors (1991) making use of an update function $\underline{f}$ that 'clamps' *s* in the network (see also Blutner, 2004).[8] Fortunately, the aforementioned formal results derived for asymptotic updating without clamping also hold for asynchronous updating with clamping.

Hence, the following set of *asymptotic updates* of s is well defined if we use an asynchronous update function $\underline{f}$ with clamping:

---

[6] The simple form of the energy function is due to assuming zero thresholds. We can always mimic the case of non-zero thresholds by assuming bias nodes with a fixed input activation.

[7] The procedure is called 'simulated annealing' (based on an analogy from physics). For details see Hinton & Sejnowski, (1983; 1986).

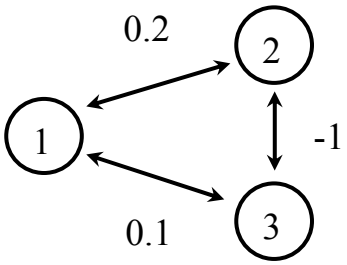
[8] Clamping is not only required if we try to model logical inferences in a connectionist network but also applies in the case of pattern completion (see, e.g. Rumelhart, Hinton, & McClelland, 1986; Smolensky, 1986).

(5) $ASUP_w(s) = \{t: t = \lim_{n\to\infty} \underline{f}^n(s)\}$

Further, in the case of the Boltzman machine, we can characterize the set of asymptotic updates as the set of all specifications of s that minimize the energy E of the system. Using the expression $min_E(s)$ to indicate this set of global energy minima, we have

(6) $ASUP_w(s) = min_E(s)$.

The following example (borrowed from Blutner, 2004) gives an illustration of the basic concepts introduced so far.



$$w = \begin{pmatrix} 0 & 0.2 & 0.1 \\ 0.2 & 0 & -1 \\ 0.1 & -1 & 0 \end{pmatrix}$$

Figure 1: Symmetric network with weight matrix

This figure shows a symmetric network consisting of three units (labelled 1, 2, and 3) and the corresponding connection matrix *w*. The set of activation states is $S = [-1, +1]^3$. Clamping node 1, the fast dynamics yields an output state where node 2 is activated and node 3 is inhibited:

(7) $ASUP_w(<1\ 0\ 0>) = \{<1\ 1\ -1>\}$

The same result is obtained if we consider the energy function on the domain S:

(8) $E(s) = -0.2\ s_1s_2 - 0.1\ s_1s_3 + s_2s_3$

The following table shows the nine possible specifications of the initial state <1 0 0> if we restrict ourselves to the discrete subdomain $S' = \{-1, 0, 1\}^3$:

Table 1: Discrete specifications of <1 0 0> and the energy of all specifications. The energy-minimal state is indicated by ☜. It corresponds to the output state given in (7).

s [state]	*E(s)* [energy]	
<1 0 0>	0	
<1 0 1>	–0.1	
<1 0-1>	0.1	
<1 1 0>	–0.2	
<1 1 1>	0.7	
<1 1-1>	–1.1	☜
<1-1 0>	0.2	
<1-1 1>	–0.9	
<1-1-1>	1.3	

In order to demonstrate that the working range of the nodes of the network is not essential for the dynamic properties of the network, we modify our example so that it relates to an activation space $[0, 1]^3$. The discrete subspace that corresponds to the states in Table 1 is obtained if we consider the map $1 \rightarrow 1$, $0 \rightarrow ½$, and $-1 \rightarrow 0$. Further, we have to adapt the energy function from (8) which is based on zero thresholds. Instead of the zero thresholds we assume thresholds $\theta_i = ½$, which are positioned in the centre of the working range. As a consequence, we have to add an additional term $-\sum_i \theta_i \cdot s_i$, which can also be seen as a consequence of introducing bias nodes with input activity 1 (see footnote 5):

(9) $E(s) = -0.2\, s_1 s_2 - 0.1\, s_1 s_3 + s_2 s_3 - ½\,(s_1 + s_2 + s_3)$

Table 2 shows the energies of the corresponding states of the discrete subspace $\{0, ½, 1\}^3$. As a matter of fact the energy ordering of the states in Table 2 is the same as the energy ordering of the corresponding states in Table1. Hence, the working space of the neurons does not really affect the ordering of the states if the thresholds are adopted accordingly.

Table 2: Corresponding specifications for the activation space $[0, 1]^3$. The energy is calculated according to formula (9) and the energy of all specifications. The energy-minimal state is indicated by ☜.

s [state]	*E(s)* [energy]	
<1 ½ ½>	-0.9	
<1 ½ 1>	-0.95	
<1 ½ 0>	-0.85	
<1 1 ½ >	-1.00	
<1 1 1>	-0.8	
<1 1 0>	-1.2	☜
<1 0 ½ >	-0.8	
<1 0 1>	-1.1	
<1 0 0>	-0.5	

Although the actual working range of a unit is only of marginal interest, the interpretation of the activation values is essential. If we take the interval [0, 1] as working range, for instance, then the interpretation of the value 0 is essential. We can either see 0 as indicating maximal underspecification or as indicating maximal specification (together with the value 1; the value ½ is typically used to indicate underspecification in this case). The former interpretation conforms to the Boolean option; the latter conforms to the DeMorgan option. The consequences of this distinction are discussed in sections 4 and 5.

3 Examples

In the previous section we have seen that the propositional objects called information states are related by a partial ordering ≥. It is obvious that this relation can be interpreted as a strict (monotonic) entailment relation since it satisfies the Tarskian restrictions for such a relation:

(10) a. $s \geq s$ (*Reflexivity*)

b. if $s \geq t$ and $s \circ t \geq u$, then $s \geq u$ (*Cut*)

c. if $s \geq u$, then $s \circ t \geq u$ (*Monotonicity*)

Here we have to make use of the operation $s \circ t =_{def} \sup\{s,t\}$, which is called *conjunction.* This operation expresses the *simultaneous realization* of two activation states. In the case where ≥ expresses the positive informational

content with regard to the state set $[0, 1]^n$ (*Boolean option*) the explicit form of the conjunction operation is given in (11a); in the second case where $\geq$ expresses specificity with regard to the state set $[-1, 1]^n$ (*DeMorgan option*) the conjunction operation is given in (11b):

(11) a. $(s \circ t)_i = \max(s_i, t_i)$

b. $$(s \circ t)_i = \begin{cases} \max(s_i, t_i), & \text{if } s_i, t_i \geq 0 \\ \min(s_i, t_i), & \text{if } s_i, t_i \leq 0 \\ \text{nil}, & \text{elswhere} \end{cases}$$

As shown by Balkenius & Gärdenfors (1991), Blutner (2004), and in a somewhat different sense by Hölldobler (1991), Pinkas (1995), and others, it is possible to define a nonmonotonic inference relation that reflects asymptotic updating of information states. Let $<S, \geq>$ be a poset of activation states, and *w* the connection matrix. Then the notion of asymptotic updates as given in (5) naturally leads to a nonmonotonic inferential relation between information states defined as follows (cf. Blutner, 2004):

(12) $s \mid\!\approx_w t$ iff $s' \geq t$ for each $s' \in \mathrm{ASUP}_w(s)$

Of course, there is an equivalent formulation in terms of energy minimization:[9]

(13) $s \mid\!\approx_E t$ iff $s' \geq t$ for each $s' \in \min_E(s)$

We also call the inferential relation between information states *subsymbolic inferential relation* and the inferences themselves *subsymbolic inferences.*

Following Balkenius & Gärdenfors (1991), the inferential notion that is adequate to describe how neuron activities spread through the network (i.e. the *fast dynamics* of a neural system) can be characterized in terms of the general postulates that Gabbay (1985) and Kraus, Lehmann, and Magidor (1990) have seen as constituting a *cumulative* (nonmonotonic) consequence relation. This holds independently of the particular working range that is chosen for the nodes of the network and it rests on the equivalence of the two inferential notions defined in (12) and (13). In (14) the relevant properties are listed.

(14) a. if $s \geq t$, then $s \mid\!\approx_w t$ (*Supraclassicality*)
b. $s \mid\!\approx_w s$ (*Reflexivity*)

[9] We simply have to use of the equivalence (6) that holds in the case of the Boltzman machine.

c. if s $|\approx_w$ t and s ∘ t $|\approx_w$ u, then s $|\approx_w$ u (*Cut*)
d. if s $|\approx_w$ t and s $|\approx_w$ u, then s ∘ t $|\approx_w$ u (*Cautious Monotonicity*)

For a proof of the validity of these properties in the case of a symmetric network, see Blutner (2004).

Going back to the earlier example introduced in Figure 1, it is a simple exercise to show that the following inferences are valid:

(15) a. <1 0 0> $|\approx_w$ <1 1 –1>
b. <1 0 0> $|\approx_w$ <1 1 0>
c. <1 0 0> $|\approx_w$ <0 1 0>

The latter two inferences can be derived from the first one by taking into account that <1 1 –1> ≥ <1 1 0> ≥ <0 1 0>.

In connectionist systems (domain) knowledge is encoded in the connection matrix *w* (or, alternatively, the energy function *E*). In the following two sections I want to discuss the close correspondence to certain symbolic systems that represent knowledge in a database consisting of expressions with default status.

4 Penalty Logic

According to Pinkas (1992, 1995), domain knowledge can be represented by a logic-based scheme, the *Penalty Logic*. This logic associates to each formula of a knowledge base the price to pay if this formula is violated. In this section I will give a concise introduction into Penalty Logic following in part the exposition in de Saint-Cyr, Lang, & Schiex (1994). Further, I will make clear that we have to adopt the Boolean option of interpreting activation states in order to reconstruct Pinkas' claim of the equivalence between inferences in Penalty Logic and inferences in symmetric networks.

Let's consider the language $\mathcal{L}_{At}$ of propositional logic (referring to the alphabet At of atomic symbols). A triple <At, Δ, k> is called a *penalty knowledge base* (PK) iff (i) Δ is a set of consistent sentences built on the basis of At (the possible hypotheses); (ii) k: Δ ⇒ (0, ∞)[10] (the penalty function). Intuitively, the penalty of an expression δ represents what we should pay in order to get rid of δ. If we pay the requested price we no longer have to satisfy δ. Hence, the larger k(δ) is, the more important δ is.

Let α be a formula of our propositional language $\mathcal{L}_{At}$. A *scenario*[11] *of* α *in PK* is a subset Δ' of Δ such that Δ'∪{α} is consistent. The cost K_{PK}(Δ') of a

[10] The notation (0, ∞) refers to the positive real numbers (excluding 0).
[11] I borrow this expression from Poole (1988).

scenario Δ' in PK is the sum of the penalties of the formulas of PK that are not in Δ':

(16) $K_{PK}(\Delta') = \sum_{\delta \in (\Delta - \Delta')} k(\delta)$

A *optimal scenario of* α *in PK* is a scenario the cost of which is not exceeded by any other scenario (of α in PK), so it is a penalty minimizing scenario. With regard to a penalty knowledge base PK, the following cumulative consequence relation can be defined:

(17) $|\sim_{PK}$ β iff β is an ordinary consequence of each optimal scenario of α in PK.

Hence, penalties may be used as a criterion for selecting preferred consistent subsets in an inconsistent knowledge base, thus inducing a non-monotonic inference relation.

To illustrate the approach I consider an example from Asimov (1950). Isaac Asimov described what became the most famous view of the ethical rules for robot behaviour in his "three laws of robotics"[12]:

First Law
A robot may not injure a human being.[13]
Second Law
A robot must follow (obey) the orders given it by human beings, except where such orders would conflict with the First Law.
Third Law
A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Now assume some human X says to the robot 'kill my wife'. The relevant knowledge base can be formalized by five propositional formulae, where I, F, P have the obvious intended meaning in connection with the three laws, S expresses that some human X gives this shocking order to the robot, and K expresses the content of the order. The first three formulae in (18) express the three laws, the last two formulae express very strong meaning postulates:

[12] Thanks to Bart Geurts for drawing my attention to this example.
[13] I am simplifying a bit. The original clause is more complicated: "A robot may not injure a human being, or, through inaction, allow a human being to come to harm."

(18)

$\neg$I	5
F	2
P	1
$(S \wedge F) \rightarrow K$	1000
$K \rightarrow I$	1000

The positive real numbers associated with the formulae are the penalties. Consider now the following two scenarios for S:

(19) $\Delta_1 = \{\neg I, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$
$\Delta_2 = \{F, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$

The cost of these two scenarios with regard to the PK given in (19) are $K_{PK}(\Delta_1) = 2$ and $K_{PK}(\Delta_2) = 5$, respectively. Since the cost of all other possible scenarios is higher, we can conclude that Δ_1 is the optimal scenario of S. Hence, according to the ethical rules, our robot should not injure anybody, neither X's wife nor himself.

Now we come to the semantic interpretation of the Penalty Logic introduced so far. Let ν denote an ordinary (total) interpretation for the language $\mathcal{L}_{At}$ (ν: At$\rightarrow${0,1}). The usual clauses apply for the evaluation $[\![.]\!]_\nu$ of the formulas of $\mathcal{L}_{At}$ relative to ν. The following function indicates how strongly an interpretation ν conflicts with the space of hypotheses Δ of a penalty knowledge base PK:

(20) $\mathcal{E}_{PK}(\nu) =_{def} \sum_{\delta \in \Delta} k(\delta) [\![\neg\delta]\!]_\nu$ ($\mathcal{E}$ is called the *system energy* of the interpretation)[14]

An interpretation ν is called a *model* of α just in case $[\![\alpha]\!]_\nu = 1$. A *preferred model* of α is a model of α with minimal energy $\mathcal{E}$ (with regard to the other models of α). As the semantic counterpart to the syntactic notion $\alpha \mid\sim_{PK} \beta$ given in (17) we can define the following relation:

(21) $\alpha \mid\approx_{PK} \beta$ iff each preferred model of α is a model of β.

As a matter of fact, the syntactic notion (17) and the semantic notion (21) coincide. Hence, the logic is sound and complete. A proof can be found in Pinkas (1995).

[14] What I call the system energy of an interpretation (with regard to a PK) is called *violation rank* for the interpretation in Pinkas (1995); deSaint-Cyr et al. (1994) call it the *cost of interpretation*.

With regard to the integration of neural networks and symbolic systems, Pinkas (1992, 1995) made a breakthrough. On the one hand he was able to demonstrate that the problem of finding preferred models for a given set of assumptions can be reduced to the minimization problem of an energy function in symmetric networks. On the other hand he showed that the minimization problem of an energy function of a symmetric network can be reduced to the problem of finding preferred models for a given set of assumptions representing domain knowledge

In the following I will give a concise description of Pinkas' basic results. I start with sketching the transformation that enables one to construct a symmetric network that is *strongly equivalent* with a given knowledge base PK. Strong equivalence means that the energy function of the neural network and the system energy of the knowledge base in Penalty Logic are the same (up to a constant c). I will sketch the basic elements of this transformation only; the reader is referred to Pinkas (1992; 1995) for a fuller description.

For each logical expression α a characteristic function $B(\alpha)$: $[0, 1]^n \rightarrow [0,1]$ is defined. The letter B for the translation operation indicates that the translation relates to the *Boolean option* of interpreting activation states. The characteristic function $B(\alpha)$ is defined in its analytical form making use of variables x_i which refer to real numbers in the interval [0, 1].

(22) a. $B(p_i) = x_i$, where p_i designates the i^{th} atomic formula of $\mathcal{L}_{At}$
b. $B(\neg\alpha) = 1-B(\alpha)$
c. $B(\alpha\wedge\beta) = B(\alpha)\cdot B(\beta)$

It is simple to see the characteristic function $B(\alpha)$ has its maximum value(s) exactly when α has a value of true (supposing the integer values of x_i are the values of the interpretations of p_i). For example, $B(p_1\wedge p_2) = x_1\cdot x_2$.[15] The maximization of $x_1\cdot x_2$ yields $x_1\rightarrow 1$, $x_2\rightarrow 1$. Further, $B(p_1\rightarrow p_2) = B(\neg(p_1\wedge\neg p_2)) = x_1\cdot x_2 - x_1 + 1$ and the maximization of the resulting term gives three solutions corresponding to the three interpretations that make the material implication true. Finally, $B(p_1\vee p_2) = B(\neg(\neg p_1\wedge\neg p_2)) = x_1\cdot x_2 - x_1 - x_2$; the maximization again gives three solutions. Figure 2 provides a graphical representation of the three characteristic functions.

[15] The same function is sometimes used in fuzzy logic. It is called product t-norm (cf. Hajek, 1998).

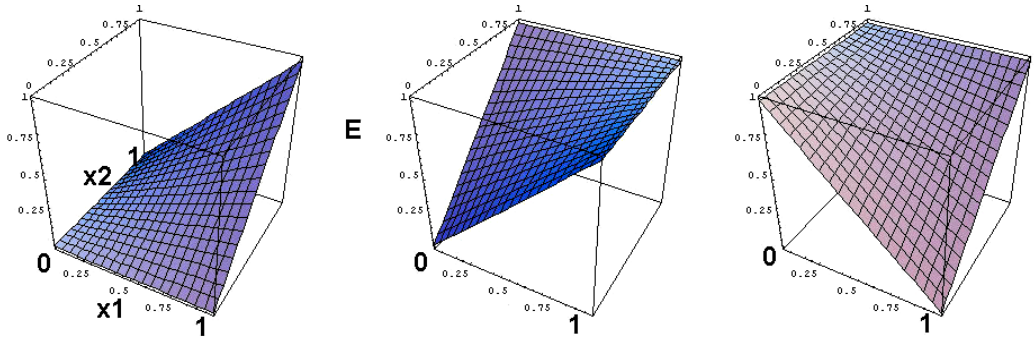


Figure 2: graphical representation of the characteristic functions for conjunction, disjunction, and material implication, respectively (from left to right)

Now we are ready to introduce a translation of a *penalty knowledge base* <At, Δ, k> into a symmetric network. We simply construct a network with the following energy function using the characteristic function B for translating propositional formulas into numerical functions:

(23) $E(x_1, \ldots, x_n) = \sum_{\delta \in \Delta} k(\delta) \cdot B(\neg\delta)$

It can be shown that the constructed symmetric network is *strongly equivalent* with the given knowledge base PK. In other words, we have the following fact:

Fact 1:

For each knowledge base PK with the assigned energy function E:
$\mathcal{E}_{PK}(\nu) = E(x_1, \ldots x_n)$ for each interpretation ν provided $\nu(p_i) = x_i$

The proof is a simple consequence of the observation that the value of a propositional formula δ for a given interpretation ν is the same as the value of the corresponding characteristic function B(δ) provided $\nu(p_i) = x_i$, i.e.

(24) $[\![\delta]\!]_\nu = B(\delta)\ [\nu(p_1)/\ x_1, \ldots, \nu(p_n)/\ x_n]$

Fact 1 then immediately follows from the definition of $\mathcal{E}$ given in (20). The proof of (24) is by induction using the translation provided in (22). Taking up the earlier example about the robot's ethics (18), we come to the following energy calculation (instead of the variables x_i we use the names of the atomic formulas as names for the variables):

Table 3: Calculation of the energy function for the PK given in (19)

Penalty	Expression in PK	Energy function
5	$\neg$I	5 I
2	F	–2F
1	P	–P
1000	(S $\wedge$ F) $\rightarrow$ K	1000(S·F–S·F·I)
1000	K $\rightarrow$ I	1000(K–K·I)
	E = 5I–2F–P+1000K+1000S·F–1000K·I–1000S·F·I	

The energy function contains a cubic term *–1000S·F·I* that goes beyond the simple quadratic energy functions introduced in (4). Such higher order energy functions refer to connectionist networks having sigma-pi units with multiplicative connections (Rumelhart et al., 1986). In the case under discussion, the following network results:

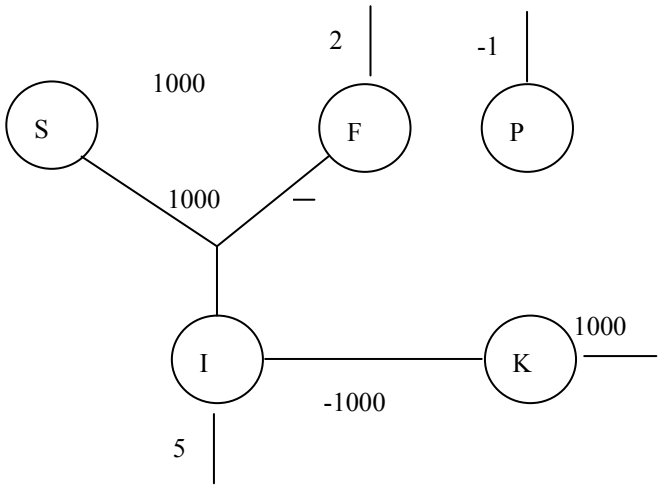


Figure 3: Higher order network representing the energy function calculated in Table 3

Pinkas (1992) has shown that higher order terms can be eliminated by introducing *hidden units*. In the case of the cubic terms const·X·Y·Z here is the relevant elimination rule, where the variable T refers to the hidden unit:

$$(25)\quad w\cdot X\cdot Y\cdot Z = \begin{cases} 2w\cdot X\cdot T+2w\cdot Y\cdot T+2w\cdot Z\cdot T-5w\cdot T, & \text{if } w<0 \\ w\cdot X\cdot Y-2w\cdot X\cdot T-2w\cdot Y\cdot T+2w\cdot Z\cdot T+3w\cdot T, & \text{if } w>0 \end{cases}$$

In the present case the coefficient is negative and the final quadratic energy function is

(26) E = 5I-2F-P+1000SF-2000ST-2000FT-2000IT+5000T+1000K-1000KI

The final network with quadratic the energy function and the hidden node T is shown in Figure 4.

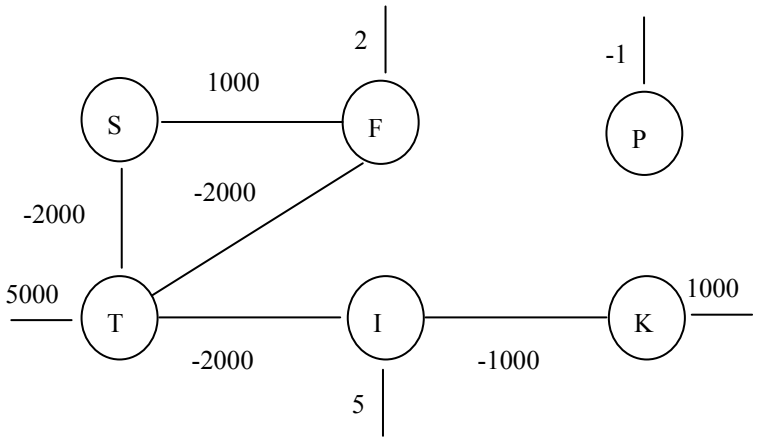


Figure 4: First order network with one hidden unit T

This was my brief sketch of how to translate any knowledge base PK into a strongly equivalent symmetric network supposed the Boolean option of interpreting activation states has been adopted.

There is also a reverse procedure that translates any symmetric network into a PK. I will outline this translation now. For simplicity, I exclude higher order units and/or hidden units. We consider a Hopfield system with connection matrix w (n units), and we assume At = $\{p_1, ..., p_n\}$ to be a set of atomic symbols. Then we consider the following formulae β_{ij} of $\mathcal{L}_{At}$:

(27) $\beta_{ij} =_{def} \text{sign}(w_{ij})(p_i \wedge p_j)$, for $1 \leq i < j \leq n$ [16]

For each connection matrix w the *associated penalty knowledge base* is defined as $PK_w = <At, \Delta_w, k_w>$, where the following two clauses apply:

[16] Sign(x) is an operator that introduces a negation sign "¬" for x<0 and it leaves the expression in its scope unchanged if x≥0. For instance, sign(0.2) (α) = α and sign(-0.2) (α) = $\neg\alpha$.

(28) a. $\Delta_w = \{\beta_{ij}: 1 \leq i < j \leq n\}$
b. $k_w(\beta_{ij}) = |w_{ij}|$

With these notations at hand we can state the following fact:

Fact 2:
For each connection matrix w the energy function $E(s) = -\sum_{i>j} w_{ij}\, s_i\, s_j$ is strongly equivalent with the associated penalty knowledge base PK_w; i.e. $\mathcal{E}_{PK}(\nu) = E(s_1, \ldots, s_n) + \text{constant}$, provided $\nu(p_i) = s_i$

For the proof we notice first that

$$[\![\beta_{ij}]\!]_\nu = [\![\text{sign}(w_{ij})(p_i \wedge p_j)]\!]_\nu = \begin{cases} \nu(p_i)\cdot\nu(p_i), & \text{if } w_{ij} \geq 0 \\ 1-\nu(p_i)\cdot\nu(p_i), & \text{if } w_{ij} < 0 \end{cases}$$

Then we have the following equivalences: $\mathcal{E}_{PK}(\nu) =_{def} \sum_{\delta\in\Delta} k(\delta)\, [\![\neg\delta]\!]_\nu = \sum_{i>j} k(\beta_{ij})\, [\![\neg\beta_{ij}]\!]_\nu = \text{const} - \sum_{i>j} w_{ij}\cdot \nu(p_i)\cdot\nu(p_i)) = \text{const} + E(s) + \text{constant}$ (provided $\nu(p_i) = s_i$). Hence, $\mathcal{E}_{PK}(\nu)$ and $E(s)$ differ only by a term $\text{const} = \tfrac{1}{2}\sum_{i>j} (w_{ij} + |w_{ij}|)$ and are therefore strongly equivalent.

For the example introduced in Figure 1 the energy function (9) was associated assuming bias nodes with fixed activity 1 that mimic thresholds $\theta_i = \tfrac{1}{2}$. This expression is repeated here for convenience:

(9) $E(s) = -0.2\, s_1s_2 - 0.1\, s_1s_3 + s_2s_3 - 0.5\, (s_1+s_2+s_3)$

The associated penalty knowledge base then comes out as follows:

(29)

$p_1 \wedge p_2$	0.2
$p_1 \wedge p_3$	0.1
$\neg(p_2 \wedge p_3)$	1
p_1	0.5
p_2	0.5
p_3	0.5

With regard to this PK it is not difficult to show that

(30) a. $p_1 \mid\sim_{PK} p_2$
b. $p_1 \mid\sim_{PK} \neg p_3$

It would be nice to have a possibility to express such inferences directly as subsymbolic inferences in the corresponding network. Unfortunately, this is possible only for inferences between positive literals such as considered in (30a):

(31) $\langle 1\ 0\ 0\rangle \mid\!\approx_E \langle 1\ 1\ 0\rangle$

Here the state <1 0 0> indicates an activation of the first node that corresponds to the atom p_1, and <1 1 0> indicates that, in addition, the second node is activated (corresponding to p_2). Unfortunately, the zero elements cannot be interpreted as negations. The reason is that in the Boolean option of interpreting node activities the vector $\mathbf{0} = \langle 0, 0, \ldots, 0\rangle$ indicates a scheme with minimal informational content. Hence, 0 indicates maximum underspecification, not a negative truth-value. As a consequence, we have no direct way to express the inferences (30b) in the subsymbolic mode.[17] In the next section we overcome this shortcoming by adopting the DeMorgan option of interpreting activation states.

5 Penalty/Reward Logic

The DeMorgan option of interpreting activation states means that we explicitly consider a resting state in the *centre* of the unit's working range in order to represent minimal information (complete underspecification). For reasons of symmetry and parsimony I choose the interval [−1, +1] as working range of a unit; the activations +1 and −1 indicate maximal specification (corresponding to the truth values T and F); the activation 0 indicates underspecification (see Section 2).

Assuming a symmetric network with n nodes it is possible now to express *all* elements of the discrete subspace $\{-1, 0, +1\}^n \subset [-1, 0, +1]^n$ by symbolic expressions. Following Blutner (2004), we can do this formally by interpreting the conjunction of literals in $\mathcal{L}_{At}$ by the corresponding elements of the DeMorgan algebra $\langle S\cup\bot, \geq\rangle$. More precisely, we call the triple $\langle S\cup\bot, \geq, 1\!\upharpoonright\rangle$ a *Hopfield model* for $\mathcal{L}_{At}$ if and only if $1\!\upharpoonright$ is a function assigning some element of $S\cup\bot$ to each atomic symbol and obtaining the following conditions:

[17] Of course, we can introduce a hard rule $\neg p_3 \leftrightarrow p_4$ in the knowledge base PK, and correspondingly an additional node that corresponds to p_4 into the network. Then we have $p_1 \mid\!\sim_{PK} p_4$ instead of (30b) and this corresponds to $\langle 1\ 0\ 0\ 0\rangle \mid\!\approx_E \langle 0\ 0\ 0\ 1\rangle$ in the extended space.

(32) a. $\downarrow\alpha\wedge\beta\downarrow = \downarrow\alpha\downarrow \circ \downarrow\beta\downarrow$

b. $\downarrow\neg\beta\downarrow = -\downarrow\beta\downarrow$ ("$\neg$" converts positive into negative activation and *vice versa*).

A Hopfield model is called *local* (for $\mathcal{L}_{At}$) iff it realizes the following assignments:

(33) $\downarrow p_1\downarrow = <1\ 0\ ...\ 0>$

$\downarrow p_2\downarrow = <0\ 1\ ...\ 0>$

...

$\downarrow p_n\downarrow = <0\ 0\ ...\ 1>$

An information state s is said to be *represented* by a formula α of $\mathcal{L}_{At}$ (relative to a Hopfield model M) iff $\downarrow\alpha\downarrow = s$. It is obvious that each discrete state $s\in\{-1, 0, +1\}^n$ can be represented by a conjunction of literals in $\mathcal{L}_{At}$ using the local Hopfield model M given in (33). For instance, if we take n=3, the following formulae *represent* proper activation states: (i) p_1 represents <1 0 0>, (ii) p_2 represents <0 1 0>, (iii) p_3 represents <0 0 1>, (iv) $p_1\wedge p_2$ represents <1 1 0>, (v) $\neg p_1$ represents <–1 0 0>, and (vi) $p_1\wedge p_2\wedge\sim p_3$ represents <1 1 –1>. Hence, for local Hopfield models each discrete activation state can be considered symbolic.

Now the following important question arises: can each connection matrix be translated into domain knowledge such that all subsymbolic inferences between information states correspond to inferences in a certain symbolic system (perhaps a Penalty Logic or a modification of it)? And, conversely: can we translate domain knowledge into a connection matrix such that all symbolic inferences of our logical system correspond to subsymbolic inferences of he connectionist system? The answer to both these question is *yes* if we use a variant of Pinkas' Penalty Logic – a variant I will call *Penalty/Reward Logic*. I will proceed as follows: first I introduce Penalty/Reward Logic, next I explain the transformation that encodes domain knowledge expressed in this logical system into a connection matrix of a symbolic network, after that I present the reverse transformation, and finally I discuss the advantages of the present approach in comparison with Pinkas' approach.

The syntax of Penalty/Reward Logic is the same as the syntax of Penalty Logic. Hence, we consider the language $\mathcal{L}_{At}$ of propositional logic (referring to the alphabet At of atomic symbols) and take a triple <At, Δ, k> as a *penalty/reward knowledge base* (PRK) where (i) Δ is a set of consistent sentences built on the basis of At and (ii) k: $\Delta \Rightarrow (0, \infty)$ is our cost function. The

idea that is connected with the cost function is that it penalizes an expression of Δ if it is not satisfied with regard to given circumstances and it rewards an expression of Δ if it is satisfied. Hence, for a *scenario of* α *in PRK* (i.e. a subset Δ' of Δ such that $\Delta' \cup \{\alpha\}$ is consistent) the cost $K_{PRK}(\Delta')$ of the scenario Δ' is defined as follows:

(34) $K_{PRK}(\Delta') =_{def} \sum_{\delta \in (\Delta - \Delta')} k(\delta) - \sum_{\delta \in \Delta'} k(\delta)$

Hence, the cost of a scenario takes into account both the beliefs that are included in the scenario Δ' and the beliefs that are not included in Δ'. The missing beliefs give a positive contribution to the overall cost and the included beliefs give a negative contribution. This contrasts with the Penalty Logic correspondence (16) where only the missing beliefs count.

However, this contrast is not really striking since we can show that Penalty Logic and Penalty/Reward Logic are weakly equivalent in the terminology of Pinkas (1995); that means they are connected by a linear transformation:

(35) $K_{PRK}(\Delta') = 2\, K_{PK}(\Delta') - \sum_{\delta \in \Delta} k(\delta)$

The last term can be seen as constant. As a consequence, Penalty Logic and Penalty/Reward Logic produce the same orderings of scenarios. However, there are differences in the probability distributions that can be calculated by using standard statistical techniques (Boltzman machine: cf. Hinton & Sejnowski, 1983; Hinton & Sejnowski, 1986).

I will define now the system energy $\mathcal{E}_{PRK}(\nu)$ which indicates how strongly an interpretation ν conflicts with the space of hypotheses Δ of the knowledge base PRK:

(36) $\mathcal{E}_{PRK}(\nu) =_{def} -\sum_{\delta \in \Delta} k(\delta)\, [\![\delta]\!]_\nu$

This definition appears to be identical with the earlier definition (20). However, we are working with the DeMorgan option now and an interpretation ν according to this option denotes a function ν: $At \rightarrow \{-1,1\}$). The usual clauses apply for the evaluation $[\![.]\!]_\nu$ of the formulas of $\mathcal{L}_{At}$ relative to ν if we take into account that –1 stands for *false* now instead of 0 in the Boolean case.

The definition (17) for a syntactic consequence relation and (21) for its semantic pendant can be taken over from the Boolean to the DeMorgan option:

(37) $|\sim_{PRK} \beta$ iff β is an ordinary consequence of each optimal scenario of α in PRK (minimizing the cost K_{PRK})

(38) $\alpha \mid\!\approx_{PRK} \beta$ iff each preferred model of α (minimizing the system energy $\mathcal{E}_{PRK}$) is a model of β.

As in the former case, the syntactic notion (37) and the semantic notion (38) coincide. Hence, the logic is sound and complete. A proof can be found in Blutner (2004).

Now I come to the transformation that enables one to construct a symmetric network that is *strongly equivalent* with a given knowledge base. Given a logical expression α a characteristic function $M(\alpha)$: $[-1, 1]^n \rightarrow [-1,1]$ is defined. The letter M indicates that the translation relates to the *DeMorgan option* of interpreting activation states. In the present case the generated variables x_i refer to real numbers in the interval [–1, 1].

(39) a. $M(p_i) = x_i$, where p_i designates the i^{th} atomic formula of $\mathcal{L}_{At}$
b. $M(\neg\alpha) = -M(\alpha)$
c. $M(\alpha\wedge\beta) = ½\ (M(\alpha)\cdot M(\beta)+M(\alpha)+M(\beta)-1)$

As a matter of fact the amount of the characteristic function $M(\alpha)$ has its *maximum* value exactly when α has a value of true (supposing the integer values of x_i are the values of the interpretations of p_i). For example, $M(p_1\wedge p_2) = ½\ (x_1\cdot x_2+x_1+x_2-1)$. The maximization of x_1 $x_1\cdot x_2+x_1+x_2-1$ yields $x_1\rightarrow 1$, $x_2\rightarrow 1$. Further, $M(p_1\rightarrow p_2) = M(\neg(p_1\wedge\neg p_2)) = ½\ (x_1\cdot x_2+ x_2-x_1+1)$ and the maximization of the resulting term gives three solutions corresponding to the three interpretations that make the material implication true. For the disjunction we get $M(p_1\vee p_2) = M(\neg(\neg p_1\wedge\neg p_2)) = ½\ (x_1+x_2-x_1\cdot x_2+1)$; the maximization again gives three solutions. The shape of these functions is precisely as in Figure 2 but with axis values running from –1 to +1 instead of from 0 to 1. It is further obvious that the characteristic function $M(\alpha)$ has its *minimum* value(s) exactly when α has a value of false. Now $\mathbf{0} = \langle 0\ 0\ 0\rangle$ builds the centre of the three dimensional cube and it conforms to the point of maximum underspecification.

The translation that transforms a *penalty/reward knowledge base* $\langle$At, Δ, k$\rangle$ into a symmetric network is straightforward. We simply construct a network with the following energy function using the characteristic function M for translating propositional formulas into numerical functions:

(40) $E(x_1, \ldots, x_n) = -\sum_{\delta\in\Delta} k(\delta)\cdot M(\delta)$

It can be shown that the constructed symmetric network is *strongly equivalent* with the given knowledge base PK. In other words, we have the following fact:

Fact 3:

For each knowledge base PRK with the assigned energy function E:
$\mathcal{E}_{PRK}(\nu) = E(x_1, \ldots x_n)$ for each interpretation ν provided $\nu(p_i) = x_i$

As in the Boolean case, the proof is a consequence of the observation that the value of a propositional formula δ for a given interpretation ν is the same as the value of the corresponding characteristic function M(δ) provided $\nu(p_i) = x_i$, i.e.

(41) $[\![\delta]\!]_\nu = M(\delta)\ [\nu(p_1)/\ x_1, \ldots, \nu(p_n)/\ x_n]$

Fact 3 then immediately follows from the definition of $\mathcal{E}_{PRK}$ given in (36). The constructed network can contain higher order units. These units can be eliminated in the same way as discussed in section 4 by introducing hidden units. The main advantage of the DeMorgan option relates to the procedure that translates a symmetric network into a symbolic knowledge PRK. As in the Boolean case discussed before, I exclude higher order units and/or hidden units.

A connection between two nodes i and j contributes a term $w_{ij} \cdot x_i \cdot x_j$ to the energy function. Now we can ask what expression α translates to the product $x_i \cdot x_j$. The answer is the biconditional: $M(p_i \leftrightarrow p_j) = M((p_i \rightarrow p_j) \wedge (p_j \rightarrow p_i)) = x_i \cdot x_j + 1/8(x_i^2 \cdot x_j^2 - x_i^2 - x_j^2 + 1)$. The last term $1/8(x_i^2 \cdot x_j^2 - x_i^2 - x_j^2 + 1)$ can be neglected since it always gives the constant 1/8 for the discrete values {−1, 0, 1}. Hence, I propose to consider the following expressions γ_{ij} as a translation of a single connection:

(42) $\gamma_{ij} =_{def} \text{sign}(w_{ij})(p_i \leftrightarrow p_j)$, for $1 \leq i < j \leq n$

For each connection matrix *w* the *associated penalty/reward knowledge base* is defined as $PRK_w = \langle At, \Delta_w, k_w \rangle$, where the following two clauses apply:

(43) a. $\Delta_w = \{\gamma_{ij}: 1 \leq i < j \leq n\}$
b. $k_w(\gamma_{ij}) = |w_{ij}|$

Corresponding to fact 2 in the Boolean case, we can prove now the following fact (cf. Blutner 2004):

Fact 4:

For each connection matrix *w* the every energy function $E(s) = -\sum_{i>j} w_{ij}\, s_i\, s_j$ is strongly equivalent with the associated knowledge base PRK_w , i.e.
$\mathcal{E}_{PK}(\nu) = E(s_1, \ldots, s_n)$ + constant, provided $\nu(p_i) = s_i$

For the proof we notice first that $[\![\gamma_{ij}]\!]_\nu = [\![\text{sign}(w_{ij})\ (p_i \leftrightarrow p_j)]\!]_\nu = \text{Sign}(w_{ij}) \cdot \nu(p_i) \cdot \nu(p_i)$, where Sign(x) equals x if $x \geq 0$ and equals –x if $x<0$. Then we have the following equivalences: $\mathcal{E}_{PRK}(\nu) =_{def} -\sum_{\delta\in\Delta} k(\delta)\ [\![\delta]\!]_\nu = -\sum_{i>j} k(\gamma_{ij})\ [\![\gamma_{ij}]\!]_\nu = -\sum_{i>j} |w_{ij}| \cdot \text{Sign}(w_{ij}) \cdot \nu(p_i) \cdot \nu(p_i) = -\sum_{i>j} w_{ij} \cdot \nu(p_i) \cdot \nu(p_i)) = E(s)$. Hence, $\mathcal{E}_{PK}(\nu)$ and E(s) are identical provided $\nu(p_i) = s_i$. Thus, they are strongly equivalent.

At the beginning of this section we introduced local Hopfield models that allow one to represent each discrete information state by a conjunction of literals of the propositional language $\mathcal{L}_{At}$. Now we can state that each subsymbolic inference between information states corresponds to an inference in Penalty/Reward Logic (and vice versa). This is an immediate consequence of Facts 3 and 4.

Fact 5:

Let α and β be formulas that are conjunctions of literals. Assume further that a penalty/reward knowledge base PRK is associated with the connection matrix *w* – by using either the transformation PRK → *w* (40) or the transformation *w* → PRK (43). Then we have: $1\alpha\downarrow \mid\approx_w 1\beta\downarrow$ iff $\alpha \mid\approx_{PRK} \beta$ (iff $\alpha \mid\sim_{PRK} \beta$)

The equivalence between subsymbolic inferences in Hopfield networks and symbolic inferences in Penalty/Reward Logic can be applied in two different ways. First, this outcome of the integrative methodology can help the symbolist to find more efficient implementations of solving optimization problems and constraint satisfaction problems. Second, the results can help the connectionist to better understand their networks and to solve the so-called *extraction problem*, i.e the extraction of symbolic knowledge from connectionist networks. The latter approach was stressed by d'Avila Garcez, Broda, & Gabbay (2001) *inter alia*, the former was pioneered by Pinkas (1992, 1995).

In our example from Figure 1 the energy function (8) was calculated in case of the DeMorgan option, repeated here.

(8) $E(s) = -0.2\ s_1s_2 - 0.1\ s_1s_3 + s_2s_3$

The corresponding knowledge base is given by the following weight-annotated defaults.

(44)

$p_1 \leftrightarrow p_2$	0.2
$p_1 \leftrightarrow p_3$	0.1
$p_2 \leftrightarrow \neg p_3$	1

The translation mechanism is very simple and transparent: it translates a node *i* into the atomic symbol p_i, translates an activating link in the network into the logical biconditional $\leftrightarrow$, and translates an inhibitory link into the biconditional $\leftrightarrow$ plus an internal negation $\neg$ of one of its arguments. Furthermore, the weights of the defaults have to be taken as the absolute value of the corresponding matrix elements.

Is the difference between choosing the Boolean option and choosing the DeMorgan option really essential? A first hint for an essential difference is obtained if we look at Figure 5 which presents the energy function (8) as function of s_2 and s_3 with a fixed value s_1=1, i.e. the first node is clamped with its maximum activity. We are interested in calculating the minimum value of the energy regarding the s_2–s_3 plane. Of course, the starting point for the minimization route is important. The De Morgan option allows us to take the starting point as expressing maximum underspecification. This corresponds to the vector <1 0 0> in the full three dimensional activation space or to the two dimensional projection <0 0>. This point is called B in Figure 5. B contrasts with the point A, which is <–1 –1>. A is the starting point in a corresponding picture using the *Boolean option*.

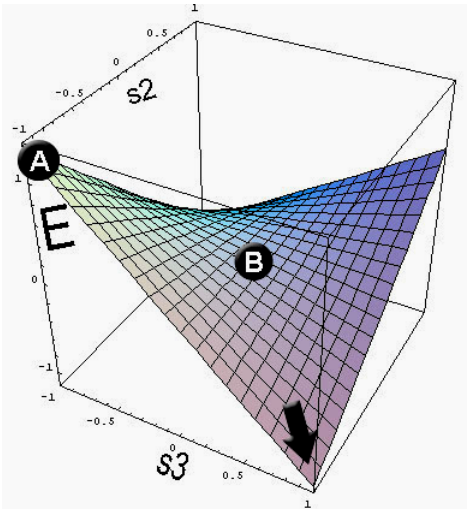


Figure 5: Energy landscape for calculating the asymptotic updates of <1 0 0>. Starting points for energy minimization: **A** for the Boolean option, **B** for the DeMorgan option.

By beginning near the centre of the cube (B) and searching using gradient descent, the network has better chances of finding a global minimum than by beginning on the top position A. Hence, the De Morgan option bears a real advantage of improving the performance of the system. In Hopfield and Tank

networks (Hopfield & Tank, 1985) this advantage is regularly exploited, and the preferred option is to start the search from the centre of the cube.

Another advantage of the De Morgan option concerns the conceptual simplicity and naturalness of solving the extraction problem. Of especial importance is that the thresholds can be assumed to be zero in cases where the De Morgan option is chosen (with a working space [-1, 1] of a unit). Hence, the additional term $-\sum_i \theta_i \cdot s_i$ can be dropped, which leads to a considerable simplification of the translation that transforms symmetric networks into symbolic knowledge bases.

A third advantage has to do with the explanation of recoverability (*bidirectionality*). In natural language theories this trait refers to a general characteristic of the form-meaning relation realized in understanding/production: *what we produce we are able to understand adequately and what we understand we are able to produce adequately.* Using the DeMorgan option of interpreting activation states, this picture will make the explanation much more transparent than the Boolean option.

In the abstract framework of pattern association patterns at a level A are associated with patterns at level B. Recoverability/bidirectionality can now be formulated as follows. We assume a simple experimental situation where a subject is presented with a (repeated) series of pairs $[a_i, b_i]$ of pattern from $A \times B$. The subject has to learn to produce the associated element, say b_i when the first member a_i of the pair is presented. Hence, in this paradigm the subject has to learn a predefined relation between a set of input patterns a_i and a set of output patterns b_i. For instance, an input pattern can be a lexigram (e.g. senseless syllable), and an output pattern can be a picture of a fruit. We assume a 1–1 correspondence between inputs and outputs.

If subjects are qualified to match stimulus a_i to b_i and then, without further training, match b_i to a_i, they have passed a **test of symmetry**. Passing this test, thus conforms to the characteristic of recoverability or bidirectionality in the domain of natural language computation. The test of symmetry plays an important role in research on the acquisition of functional symbol usage in apes and children. The important empirical finding is that children as young as 2 years pass the symmetry test (e.g. Green, 1990). In contrast, chimps did not show symmetry: having learned to match lexigram comparisons to object samples, the chimps were not able, without further training, to match the same objects now presented as comparisons to the corresponding lexigrams, now presented as samples (cf. Savage-Rumbaugh, 1984; Dugdale & Lowe, 2000).[18]

[18] A possible exception is Kanzi, the bonobo monkey. Kanzi's knowledge was reciprocal. There was no need to teach her separately to produce and to comprehend (Savage-Rumbaugh & Lewin, 1994).

Using symmetric networks it is very simple to account for recoverability (passing the symmetry test) after learning the association $a_i \rightarrow b_i$ (assuming a 1-1 correspondence). For simplicity, we adopt a localist model with two levels of nodes such that the nodes correspond to the pattern a_i and b_i, respectively. Using the DeMorgan option, this corresponds to a system of weighted constraints $\{[a_i \leftrightarrow b_j: w_{ij}], 1 \leq i,j \leq N\}$ plus strict inhibitory links within the level A and B, respectively: $\{[a_i \leftrightarrow \neg a_j: \infty], i \neq j\} \cup \{[b_i \leftrightarrow \neg b_j: \infty], i \neq j\}$. Now it is not difficult to show that we can reproduce the list $a_i \rightarrow b_i$ for all i if and only if $w_{ii} > \sum_{1 \leq j \leq N} w_{ij}$ for each $1 \leq i \leq N$. That conforms to getting the inferences $a_i \mid\approx_{PRK} b_i$ with the corresponding knowledge base PRK. Because of the symmetry of the knowledge base it can be concluded that the list can be reproduced in reverse order: $b_i \rightarrow a_i$ (i.e. $b_i \mid\approx_{PRK} a_i$).

Concluding this section we can say that the DeMorgan option has a series of advantages if compared to the Boolean option: (i) it accounts to the idea of underspecification and inferential completion; (ii) it helps to improve the performance of the optimization procedure; (iii) it provides a conceptually simple and natural solution to the extraction problem; (iv) it makes the feature of recoverability transparent.

6 Optimality Theory and Symmetric Networks

Optimality theory (OT) was initiated by Prince & Smolensky (1993/2004) as a new phonological framework that deals with the interaction of violable constraints. In recent years, OT was the subject of lively interest also outside phonology. Students of morphology, syntax and natural language interpretation became sensitive to the opportunities and challenges of the new framework (e.g. Blutner & Zeevat, 2004). The reasons for linking scientists into this new research paradigm is manifold: (i) the aim to decrease the gap between competence and performance, (ii) interest in an architecture that is closer to neural networks than to the standard symbolist architecture, (iii) the aim to overcome the gap between probabilistic models of language and speech and the standard symbolic models, (iv) the logical problem of language acquisition, (v) the aim to integrate the synchronic with the diachronic view of language.

In the present context we emphasize the second motive. OT is deeply rooted in the connectionism paradigm of information processing. As a consequence, OT does not assume a strict distinction between representation and processing. The development of OT demonstrates a new and exciting research strategy: augmenting and modifying symbolist architecture by integrating

insights from connectionism. The development of Penalty Logic is another illustration of this strategy.

It's not possible to give a systematic introduction into OT here.[19] The primary aim of this section is to draw attention to the close similarities between OT and the logical approach proposed in Sections 4 and 5, but also to point out some significant differences. The main difference between OT and numerical theories like *Penalty Logic* and *Harmonic Grammar* (Smolensky, 1986, 1995) is the shift from numerical to non-numerical constraint satisfaction. Why Prince and Smolensky (1993) proposed this shift, is explained by Smolensky (Smolensky, 1995: 266) as follows: "Phonological applications of Harmonic Grammar led Alan Prince and myself to a remarkable discovery: in a broad set of cases, at least, the relative strengths of constraints of constraints *need not be specified numerically*. For if the numerically weighted constraints needed in these cases are ranked from strongest to weakest, it turns out that each constraint is stronger than all the weaker constraints *combined*." In other words, the shift from Harmonic Grammar to Optimality Theory, that means the realization of what is called *strict dominance of the OT constraints* appears to be mainly motivated by empirical findings in the domain of phonology.

A possible advantage of strict dominance lies in the robustness of processing. Following a suggestion of David Rumelhart the following argument was put forward: "Suppose it is important for communication that language processing computes global harmony maxima fairly reliably, so different speakers are not constantly computing idiosyncratic parses which are various local Harmony maxima. Then this puts a (meta-)constraint on the Harmony function: it must be such that local maximization algorithms give global maxima with reasonably high probability. Strict domination of grammatical constraints appears to satisfy this (meta-)constraint." (Smolensky 1995, note 38: 286).

In concord with this argument it is not implausible to assume that the theoretical explanation for differences between automatic and controlled psychological processes (Schneider & Shiffrin, 1977) can also be seen as an emergent effect of the underlying neural computations (cf. Blutner, 2004). Whereas controlled processing relates to the capacity-limited processing when the global harmony maxima (= global energy minima) are difficult to grasp, automatic processing relates to a mode of processing where most local harmony maxima are global ones.

In order to illustrate the strictness of domination of grammatical constraints I consider a small fragment of the vowel system of English (cf. Kean, 1995),

[19] For good introductions the reader is referred to the literature (e.g. Archangeli & Langendoen, 1997; Kager, 1999; Smolensky & Legendre, to appear).

which is roughly simplified for the present purpose.[20] The example rests on a classification of the vowels in terms of the binary phonemic features as illustrated in Table 4.

Table 4: Fragment of the vowel system of English and the phonological feature specifications

	/a/	/i/	/o/	/u/	/ɔ/	*/e/*	/æ/
back	+	–	+	+	+	–	–
low	+	–	–	–	+	–	+
high	–	+	–	+	–	–	–
round	–	–	+	+	+	–	–

For the purpose of applying propositional Penalty Logic, the phonological features may be represented by the atomic symbols BACK, LOW, HIGH, ROUND. The knowledge of the phonological agent concerning this fragment may be represented by the following violable constraints (usually called *markedness conventions*)[21]:

(45) a. VOC $\leftrightarrow$BACK ε^1
b. BACK $\leftrightarrow$LOW ε^2
c. BACK $\leftrightarrow\sim$HIGH ε^3
d. LOW $\leftrightarrow\sim$ROUND ε^4

With regard to the agent's knowledge, the feature specifications in Table 4 are highly redundant. It can be shown that only the feature specifications in the grey fields must be given, the specification in the remaining fields can be calculated by the agent's knowledge. For the proper working of the constraint system in (45) it is required that the constraints are ordered in a hierarchical way, with (45a) at top and (45d) at bottom. This hierarchy corresponds to a relation of strict domination: one violation of a higher ordered constraint cannot be overpowered by arbitrary many violations of lower ordered constraints. The technical means of expressing the hierarchy is the use of *exponential penalties* with a basis $0 < \varepsilon \leq 0.5$. In the present case, $\varepsilon = ½$ or smaller is a proper base since we are concerned with binary features which can be applied only once in each case.

[20] I borrow this example from Blutner (2004).
[21] Further, two hard constraints are needed to express strong redundancies: LOW$\rightarrow\neg$HIGH; ROUND$\rightarrow$BACK.

Table 5 illustrates a sample calculation using an OT tableau. As usual in the OT literature a violation of a constraint is indicated by * and the *small hand* icon is used to mark the optimal candidate. In the present case we have only two candidates that satisfy the input's conditions for a non-high front vowel. The only free feature corresponds to ±LOW. It resolves to –LOW because of the second constraint, which is the highest ranked constraint that discriminates the two candidates: it is satisfied for the optimal candidate but violated for the other candidate. The optimal candidate distinctively characterizes the vowel /e/. In the last column penalties are calculated from the constraint violations assuming penalties ε^n for constraints of rank n (with $\varepsilon = ½$).

Table 5: OT tableau for calculating the optimal non-high front vowel: /e/

Input: + VOC ∧ –BACK ∧ –HIGH

	BACK	LOW	HIGH	ROUND	VOC ↕ BACK	BACK ↕ LOW	BACK ↕ ~HIGH	LOW ↕ ~ROUND	*Penalty* ($\varepsilon = ½$)
	–	+	–	–	*	*	*		0.5550
☞	–	–	–	–	*		*	*	0.5055

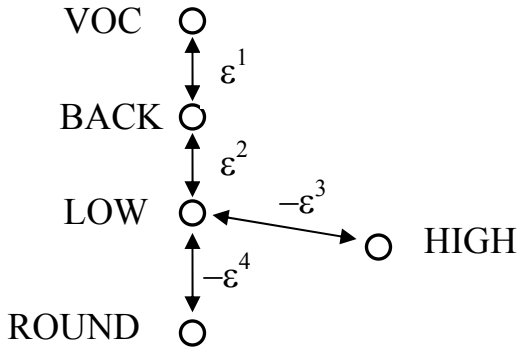


Figure 6: Hopfield network with exponential weights representing the generic knowledge[22] of a phonological agent

[22] The hard constraints mentioned in footnote 20 are not represented in this network. We leave it as an exercise for the interested reader to perform the corresponding modifications using the techniques explained in section 5.

We conclude that both Penalty Logic and Optimality Theory look for an optimal satisfaction of a system of conflicting constraints. Most importantly, the exponential form of the penalty function results in *strict domination* of the constraints, meaning that violations of many lower ranked constraints invariably count less than one violation of a higher ranked constraint. Moreover, we have seen how constraints that conform to formulae of propositional logic can be translated into a symmetric connectionist network by assuming a localist interpretation of the atomic symbols.

Early proposals to ground OT in connectionist architecture made use of (non-symmetric) feedforward networks (cf. Smolensky, 1986; Prince & Smolensky, 1993/2004). However, Smolensky & Legendre (to appear) also acknowledged the relevance and power of symmetric networks for developing an integrated connectionist/symbolic cognitive architecture. One important advantage of symmetric networks is that they give a natural account of the emergence of recoverability and bidirectionality.

There are two shortcomings with the presented account of reducing OT to connectionist networks. The first one concerns the use of a localist interpretation. Though a localist interpretation generates a fairly transparent relationship between symbols and node activities, this idea is much too naïve to be taken seriously as a promising programme in Cognitive Neuroscience. In realistic examples, the relation between the symbolic expressions as used in Penalty Logic and the elements of the pre-symbolic product space is much less direct than localist Hopfield models suggest. In an outstanding dissertation, Martinez (2004) proposed ideas for simultaneously using discrete symbolic means and non-discrete numerical means, and she developed tools of relating the two different realms in a much less direct way than strictly localist accounts suggest (see also Barwise & Seligman, 1997; Martinez, 2003). I think these ideas have a big potential for future accounts for an integrated connectionist/symbolic cognitive architecture.

The second shortcoming relates to the fact that the constraints we used in the example from intrasegmental phonology are *micro-constraints* in the sense that they are in direct correspondence to a very small fragment of the network. In fact, in the case under discussion each constraint corresponds to a pair of two linked nodes in the network. It is also indispensable to have constraints that correspond to larger parts of a network even when a localist interpretation is used. The whole idea of assemblies we mentioned in section 2 suggests that constraints are distributed over significant parts of the network. Hence, it is opportune to propose an extended scheme. In this connection I will introduce the notion of *macro-constraints*. In a first approximation, macro-constraints can be

seen as an organized congregation of micro-constraints, and they can be considered to constitute *innate structure*. The idea of macro-constraints is closely related to the idea of an *abstract genome* as developed by Smolensky & Legendre (to appear: Chapter 21). In detail, the idea has been worked out for basic CV syllable theory.

Macro-constraints can be defined as collections of micro-constraints with identical penalties. The idea of associating micro-constraints with identical penalties becomes appealing when we translate the set of micro-constraints into a neural net. Then identical penalties correspond to fixed relationships between certain connection weights in the symmetric network. For instance, let's assume a weighted macro-constraint $\mathbf{C} = \{p_i \leftrightarrow p_j, i \neq j\}: w$, where w is the penalty associated with all the micro-constraints $p_i \leftrightarrow p_j$ in **C.** Hence, all weights between the nodes i and j in the network are required to be identical and to have the value w. Though the penalties can be changed by learning it is assumed that the identity of the corresponding weights is not lost over the course of learning. Thus this relationship is maintained during learning, although the absolute magnitude of the weights changes as particular knowledge is acquired. As a consequence, the relationship between connection weights can be considered to constitute the innate knowledge provided by a constraint (cf. Smolensky & Legendre, to appear).

Concluding, macro-constraints are essential for two related reasons: (i) they correspond to larger parts of the network and constitute assemblies, (ii) they express an innate a relationship, which is not influenced by learning.

In sections 4 and 5 we have formalized a penalty (penalty/reward) knowledge base as a triple <At, Δ, k> where Δ was a system of propositional expressions (=micro-constraints). Now we consider macro-constraints as (non-empty) sets of micro-constraints, and a macro-knowledge base MK can be defined as a corresponding triple $<At, {}^{M}\Delta, {}^{M}k>$, where (i) ${}^{M}\Delta$ is a set of nonempty sets of consistent sentences built on the basis of At; (ii) ${}^{M}k: {}^{M}\Delta \Rightarrow (0, \infty)$, the penalty function that associates penalties with each macro-constraint. Now the system energy of an interpretation ν with regard to a macro-knowledge base MK is defined as follows:

$$(46) \quad \mathcal{E}_{MK}(\nu) =_{def} -\sum_{\mu \in {}^{M}\Delta} {}^{M}k(\mu) \sum_{\delta \in \mu} [\![\delta]\!]_{\nu}$$

For each macro-knowledge base $MK = <At, {}^{M}\Delta, {}^{M}k>$ we can construct the associated ordinary knowledge base $K = <At, \Delta, k>$, where $\Delta = \cup {}^{M}\Delta$ and $k(\delta) = {}^{M}k(\mu)$ if $\delta \in \mu$. It is obvious that the system energy (47) of an interpretation with regard to a macro-knowledge base is identical to the system energy of an interpretation with regard to the associated ordinary knowledge base: $\mathcal{E}_{MK}(\nu) = \mathcal{E}_{K}(\nu)$. The crucial point is that the penalties $k(\delta)$ for all micro-constraints δ that

constitute the macro-constraint μ are identical. Further, it is obvious how to construct the symmetric network that corresponds to a macro-knowledge base: build the associated ordinary (micro-) knowledge base and translate it into the network using the technique explained in sections 4 and 5.

7 Conclusions: Logic and embodied theories of cognition

The present contribution can be seen as part of recent efforts to develop an embodied view of cognition. The emerging viewpoint of embodied cognition holds that cognitive processes are deeply rooted in the body's interactions with the world (cf. Brooks (1999); Anderson (2003); Lakoff & Johnson (1999); Varela, Thompson, & Rosch (1993)). The idea of embodiment has diverse aspects. Several philosophers and cognitive scientist agree that at least the following three aspects are of special importance (cf. Anderson, 2003):

- Reductionist aspect: The system must be realised in a coherent, integral physical/biological structure. As an immediate consequence, certain features of the symbolic system (e.g. the OT Grammar) must be reducible to plausible neural models.
- Evolutionary aspect: The explanation of the behaviour must include reference to cultural evolution. This derives from the observation that intelligence lies less in the individual brain and more in the dynamic interaction of brains with the wider world, including especially the social and cultural worlds.[23]
- Grounding aspect: Symbol-manipulation has to be grounded in non-symbolic function. OT constraints are embodied, not disembodied. A symbol is grounded if it has its meaning or content by virtue of its causal properties and relations to the referent of the symbol. Hence, symbols have to be grounded ultimately in the sensory-motor system or other bodily systems or are appropriately defined in terms of grounded symbols.

The research program of embodied cognition is a continuation of the program of *situated* cognition. It is the centrality of the *physical grounding project* in

[23] In the domain of linguistics, Jackendoff (2002)makes the following remarkable claim stressing the influence of cultural interaction in understanding language: "If some aspects of linguistic behaviour can be predicted from more general considerations of the dynamics of communication in a community, rather than from the linguistic capacities of individual speakers, then they should be." (Jackendoff 2002:101).

embodied cognition that differentiates these two research programs (cf. Anderson, 2003).

Taking up the view of embodiment, the present article builds mainly around the reductionist aspect of embodiment. What are the central general principles of computation in connectionist – abstract neural – networks? How can these principles be reconciled with those of symbolic computation? Which basic assumptions of OT can be reduced to connectionist computation? And in what case alternate explanations are required? In a nutshell, we can state the following main results:

- To overcome the gap between symbolism and connectionism it is useful to view symbolism as a high-level description of the properties of (a class of) neural networks. The application of algebraic and model-theoretic techniques for a higher-level analysis of neural networks (e.g. Balkenius & Gärdenfors, 1991; Pinkas, 1995; Blutner, 1997, 2004) and their development in the present paper proves especially valuable when it comes to study the concrete link between inferences in symmetric networks and inferences in nonmonotonic logic.
- The foundational issue of OT: The general shape of symbolic OT systems proves to be conforming to the penalty-logical treatment proposed in sections 4 and 5. Because of the close relations between Penalty Logic and symmetric networks, certain features of standard OT appear to be reducible to the basic traits of neural network models. This concern first at all the idea of domination: constraint conflict is resolved via a notion of differential strength: stronger constraints prevail over weaker ones in cases of conflict.
- Strictness of domination (hierarchical encoding of constraint strengths): This problem matters both from a theoretical and an empirical perspective. In the words of Bechtel, the solution to this problem "may create a rapprochement between network models and symbolic accounts that triggers an era of dramatic progress in which alignments are found and used all the way from the neural level to the cognitive/linguistic level (Bechtel, 2002, p.17). Presently, there are only vague ideas about how to account for the strictness of domination and the entailed idea that Grammar (usually) does not count. Moreover, it is rather unclear how to give a theoretically satisfying account for explaining under which conditions the strict domination of constraints applies and under which conditions it does not.
- The idea of macro-constraints is essential for matching larger parts of a network (assembly formation). Further, macro-constraints can be used to express innate relationships on symmetric networks – i.e. relationships that aren't controlled by learning.

Standard OT respects the generative legacy in assuming that the universal features of language can be explained by assuming a Universal Grammar (UG). UG describes the innate knowledge of language that is shared by individual humans. In standard OT, the innate knowledge of language consists (a) of a generative device that generates the admissible input-output pairs and (b) the set of constraints. Language-particular aspects refer to the possible rankings of the constraints (e.g. Prince & Smolensky, 1993/2004). Hence, the suggestion of an abstract genome (Smolensky & Legendre, to appear) as well as the suggestion of macro-constraints and the way they constrain symmetric networks nicely fits into this picture.

However, recent effort on the problem of the evolution of language in humans (e.g. Hurford, 1998; Steels, 1998; Kirby, 2002; Zeevat & Jäger, 2002) made clear that a thorough explanation of the universal properties of language cannot be exclusively based on an individual's cognitive capacity which is taken to be biologically determined. So, if we want to know how and where the universal features of language are specified, it is not sufficient to consider only an individual's competence and how it is derived from primary linguistic data via the Language Acquisition Device (LAD). Rather, it is essential to focus on how certain hallmarks of human language can arise in the absence of biological change by assuming the force of *cultural evolution*. In explaining the universal properties of language, the evolutionary approach is in line with the claims made by proponents of embodied cognitive science. Hence, it is our central task to investigate the interaction between biological and cultural substrates. The paradigm of iterated learning (e.g. Kirby & Hurford, 1997; Kirby, 2002) has proven as especially useful in investigating the emerging effects from this interaction.

Taking the dimension of cultural evolutionary into account suggest that at least some principles of OT van be explained as emergent factors of cultural exchange. This concerns, first at all the explanation of bias constraints (Zeevat & Jäger, 2002) and the principle of constructional iconicity[24], which is related to the feature of weak bidirection (Mattausch, 2004). Hence, naïve OT with its assumption of inborn constraints has to be overcome by an embodied OT, which respects the role of grounding constraints by iterated learning. In this regard it is important that the mechanism of grounding is directed by mechanisms that are

[24] Constructional iconicity states that there is a harmonic linking between complex semantic contents and complex (surface) forms on the one hand and less complex semantic contents and simple forms on the other hand. Both in pragmatics and in (natural) morphology the principle plays an important role in describing the *direction of language change*. In formal semantics, this principle is called *division of pragmatic labour* (Horn, 1984); in the school of „natural morphology“ it is called *constructional iconicity* (Wurzel, 1998).

very close to those used in modelling evolutionary change (e.g. Hayes, 1996; Boersma, 1998).

In this article I have concentrated on the reductionist aspect of embodied cognition – certain features of a symbolic system (e.g. the OT Grammar) must be reducible to plausible neural models. Though the reductionist programme is an integral part of the embodied paradigm it is not the whole story. The evolutionary aspect and the aspect of grounding likewise deserve attention. Once more, the feature of situatedness, i.e. dynamic interaction of brains with the wider world, including especially the social and cultural worlds, should prove promising for future research.

8 References

Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence, 149*, 91–130.

Archangeli, D., & Langendoen, D. T. (1997). *Optimality theory: An overview*. Malden, MA/Oxford, UK: Blackwell.

Asimov, I. (1950). *I, Robot*: Gnome Press.

Balkenius, C., & Gärdenfors, P. (1991). Nonmonotonic inferences in neural networks. In J. A. Allen & R. Fikes & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning*. San Mateo, CA: Morgan Kaufmann.

Barwise, J., & Seligman, J. (1997). *Information flow: the logic of distributed systems*. New York: Cambridge University Press.

Bechtel, W. (2002). *Connectionism and the Mind*. Oxford: Blackwell.

Blutner, R. (1997). Nonmonotonic logic and neural networks. In P. Dekker & M. Stokhof & Y. Venema (Eds.), *Proceedings of the eleventh Amsterdam Colloquium* (pp. 79-84): ILLC/Department of Philosophy, University of Amsterdam.

Blutner, R. (2004). Nonmonotonic inferences and neural networks. *Synthese (Special issue Knowledge, Rationality and Action), 142*, 143-174.

Blutner, R., & Zeevat, H. (Eds.). (2004). *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.

Boersma, P. (1998). *Functional phonology*. The Hague: Holland Academic Graphics.

Brooks, R. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* Cambridge, MA: MIT Press.

Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competetive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 815-826.

d'Avila Garcez, A. S., Broda, K., & Gabbay, D. M. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence, 125*, 155-207.

deSaint-Cyr, F. D., Lang, J., & Schiex, T. (1994). Penalty logic and its link with Dempster-Shafer theory, *Proceedings of the 10th Int. Conf. on Uncertainty in Artificial Intelligence (UAI'94)* (pp. 204-211).

Dugdale, N., & Lowe, C. F. (2000). Testing for symmetry in the conditional discriminations of language-trained chimpanzees. *Journal of the Experimental Analysis of Behavior, 73*, 5-22.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205-254.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition, 28*, 3-71.

Gerstner, W., & Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, Mass.: Cambridge University Press.

Graben, P. b. (2004). Incompatible Implementations of Physical Symbol Systems. *Mind and Matter, 2*, 29–51.

Green, G. (1990). Differences in development of visual and auditory-visual equivalence relations. *Journal of the Experimental Analysis of Behavior, 51*, 385–392.

Hajek, P. (1998). *Metamathematics of fuzzy logic*. Dordrecht: Kluwer.

Hayes, B. P. (1996). Phonetically Driven Phonology: The Role of Optimality Theory and Inductive Grounding.

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

Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference, *Proceedings of the Institute of Electronic and Electrical Engineers Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 448-453). Washington, DC: IEEE.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzman machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I and II.* (pp. 282-317). Cambridge, MA: The MIT Press/Bradford Books.

Hölldobler, S. (1991). Towards a connectionist inference system. In N. Cercone & F. Gardin & G. Valle (Eds.), *Computational Intelligence III* (pp. 25-38): North-Holland.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective comput-ational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554-2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci., 81*, 3088-3092.

Hopfield, J. J., & Tank, D. W. (1985). Neural computation of decisions in optimization problems. *Biol. Cybern., 52*, 144-152.

Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.

Hurford, J. R. (1998). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua, 77*, 187–222.

Jackendoff, R. (2002). *Foundations of Language*. Oxford: Oxford Univeristy Press.

Jibu, M., & Yasue, K. (1995). *Quantum Brain Dynamics and Consciousness*. Amsterdam/Philadelphia: John Benjamins.

Kager, R. (1999). *Optimality theory*. Cambridge: Cambridge University Press.

Kean, M. L. (1995). *The theory of markedness in generative grammar.* Unpublished Ph.D. thesis, MIT, Cambridge, Mass.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life, 8*, 185–215.

Kirby, S., & Hurford, J. (1997). *The evolution of incremental learning: language, development and critical periods*. Edinburgh: University of Edinburgh.

Kokinov, B. (1997). Micro-level hybridization in the cognitive architecture DUAL. In R. Sun & F. Alexander (Eds.), *Connectionist-symbolic integration: From unified to hybrid approaches* (pp. 197-208). Hilsdale, NJ: Lawrence Erlbaum Associates.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

Maass, W. (1999). *Pulsed Neural Networks*. Cambridge, Mass.: MIT Press.

Martinez, M. (2003). Towards a model of heterogeneous commonsense

reasoning. In J. Baldwin & R. d. Queiroz & E. H. Hauesler (Eds.), *Proceedings of WoLLIC'2001. Matemtica Contempornea, V 24*: Sociedade Brasileira de Matemtica.

Martinez, M. (2004). *Commonsense reasoning via product state spaces.* Unpublished PhD, Indiana University, Bloomington.

Mattausch, J. (2004). *On the Optimization & Grammaticalization of Anaphora.* Unpublished Ph.D. Thesis, Humboldt University, Berlin.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I and II*. Cambridge, MA: The MIT Press/Bradford Books.

Palm, G., & Wennekers, T. (1997). Synchronicity and its use in the brain. *Behavioral and Brain Sciences, 20*, 295-296.

Pinkas, G. (1992). *Logical inference in symmetric connectionist networks.* Unpublished Doctoral thesis, Washington University, St Louis, Missouri.

Pinkas, G. (1995). Reasoning, connectionist nonmonotonicity and learning in networks that capture propositional knowledge. *Artificial Intelligence, 77*, 203-247.

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence, 36*, 27-47.

Pribram, K. H. (1991). *Brain and Perception*. New Jersey: Lawrence Erlbaum.

Prince, A., & Smolensky, P. (1993). *Optimality theory*. Rutgers Center for Cognitive Science: Technical Report RuCCSTR-2.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Rutgers University and University of Colorado at Boulder: Technical Report RuCCSTR-2, available as ROA 537-0802. Revised version published by Blackwell, 2004.

Rojas, R. (1996). *Neural Networks - A Systematic Introduction*. Berlin, New-York: Springer-Verlag.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In J. L. McClelland & D. E. Rumelhart & the-PDP-Research-Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1*. Cambridge, MA: MIT Press.

Savage-Rumbaugh, E. S. (1984). Acquisition of functional symbol usage in apes and children. In H. L. Roitblat & T. G. Bever & H. S. Terrace (Eds.), *Animal Cognition* (pp. 291-310). Hillsdale, NJ: Lawrence Erlbaum Associates.

Savage-Rumbaugh, S., & Lewin, R. (1994). *Kanzi : The Ape at the Brink of the Human Mind*: John Wiley & Sons.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. *Psychological Review, 84*, 1-66.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences, 3*, 417-457.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rule, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16*, 417-494.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypotheses. *Ann. Rev. Neuroscience, 18*, 555-586.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. volume 1: Foundations* (pp. 194-281). Cambridge, Mass., London: MIT.

Smolensky, P. (1995). Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation* (pp. 221-290). Oxford: Blackwell.

Smolensky, P., & Legendre, G. (to appear). *The Harmonic Mind: From neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.

Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence, 103*, 133–156.

Varela, F. J., Thompson, E., & Rosch, E. (1993). *The Embodied Mind*.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report 81-2). Göttingen: Max-Planck-Institut für Biophysikalische Chemie.

Wennekers, T. (1999). *Synchronisation und Assoziation in neuronalen Netzen*. Aachen: Shaker Verlag.

Wennekers, T., & Palm, G. (2000). Cell assemblies, associative memory and temporal structure in brain signals. In R. Miller (Ed.), *Time and the Brain. Conceptual Advances in Brain Research, vol II*: Harwood Academic Publishers.

Wurzel, W. U. (1998). On markedness. *Theoretical Linguistics, 24*, 53-71.

Zeevat, H., & Jäger, G. (2002). *A statistical reinterpretation of harmonic alignment.* Paper presented at the 4the Tblisi Symposium on Logic, Language and Linguistics, Tblisi.