To appear in Knowledge, Rationality and Action, issue 2, 2004

Nonmonotonic Inferences and Neural Networks

Reinhard Blutner
University of Amsterdam

Summary

There is a gap between two different modes of computation: the symbolic mode and the subsymbolic (neuron-like) mode. The aim of this paper is to overcome this gap by viewing symbolism as a high-level description of the properties of (a class of) neural networks. Combining methods of algebraic semantics and nonmonotonic logic, the possibility of *integrating* both modes of viewing cognition is demonstrated. The main results are (a) that certain activities of connectionist networks can be interpreted as *nonmonotonic inferences*, and (b) that there is a strict correspondence between the coding of knowledge in Hopfield networks and the knowledge representation in weight-annotated Poole systems. These results show the usefulness of nonmonotonic logic as a descriptive and analytic tool for analyzing emerging properties of connectionist networks. Assuming an exponential development of the weight function, the present account relates to optimality theory – a general framework that aims to integrate insights from symbolism and connectionism. The paper concludes with some speculations about extending the present ideas.

1 Introduction

A puzzle in the philosophy of mind concerns the gap between symbolic and subsymbolic (neuron-like) modes of computation/processing. Complex symbolic systems like those of grammar and logic are essential when we try to understand the general features and the peculiarities of natural language, reasoning and other cognitive domains. On the other hand, most of us believe that cognition resides in the brain and that neuronal activity forms its basis. Yet neuronal computation appears to be numerical, not symbolic; parallel, not serial; distributed over a gigantic number of different elements, not as highly localized as in symbolic systems. Moreover, the brain is an adaptive system that is very sensitive to the statistical character of experience. Hard-edged rule systems are not suitable to deal with this aspect of behavior.

The methodological position pursued in this article is an *integrative* one, which looks for unification. In the case under discussions the point is to assume that symbols and symbol processing are a macro-level description of what is considered as a connectionist system at the micro level. This position is analogous to 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 (e.g. Smolensky 1988, 1996; Balkenius & Gärdenfors 1991; Kokinov 1997).

It should be stresses that the *integrative* methodology is not the only one. Alternatively, some researchers like to play down the neuronal perspective to a pure issue of implementation. Representatives of this position are, *inter alias*, Fodor and Pylyshyn (1988), who insist that the proper role of connectionism in cognitive science is merely to *implement* existing symbolic theory. On the other extreme, there is a school that Pinker & Prince (1988) and Smolensky et al. (1992) call *eliminative* connectionism. The approach to the gap taken by these researchers is simply to ignore it and to deny the existence of the symbolic perspective and higher-level cognition. Many connectionist research falls into this category and some of its representatives make a major virtue out of this denial. Finally, there is a movement towards *hybrid* systems (e.g. Hendler 1989, 1991; Boutsinas & Vrahatis 2001). According to this approach two computationally separate components are assumed, one connectionist and the other symbolic, and an interface component has to be constructed that allows the two components to interact.

It is not unfair to say the hybrid approach is an eclectic one. In particular, it requires extra stipulations to construct the interface. In my opinion, it is not a good idea to develop models where separate modules correspond to separate cognitive processes and are described within separate paradigms, like a connectionist model of perception (and apperception) combined with a symbolic model of reasoning. Instead, both aspects should basically be integrated and contribute at every level to every cognitive process. Qualifying the *hybrid* approach as an eclectic one, I cannot consider the *eliminative* position to be especially helpful, either. It is evident that such a complex object as the human mind (and human reasoning in particular) is too complex to be fully described by a single formal theory or model, and therefore several different and possibly contradicting perspectives are needed.

Downplaying connectionism as a pure issue of implementation is a point that deserves careful attention. I think there are two different reactions that prove this position as untenable. The first one is to show that *non-classical* cognitive architectures may emerge by assuming connectionist ideas at the micro-level. The development of optimality theory (Prince & Smolensky 1993) may be a case in point, and likewise it is the demonstration of the close correspondence between certain kinds of neural networks and belief networks (e.g. Glymour 2001). The second possible reaction is to demonstrate that aspects of possibly old-fashioned, "classical" architectures can be *explained* by assuming an underlying (connectionist) micro-level. If this is achievable we have got much more than what usually is connected with the conception of implementation. In this article, I wish to demonstrate both lines of argumentation.

In a nutshell, the aim of this paper is to demonstrate that the gap between symbolic and neuronal computation can be overcome when we view symbolism as a high-level description of the properties of (a class of) neural networks. The important methodological point is to illustrate that the instruments of model-theoretic (algebraic) semantics and nonmonotonic logic may be very useful in realizing this goal. In this connection it is important to stress that the algebraic perspective is entirely neutral with respect to foundational questions such as whether a "content" is in the head or is a platonic abstract entity (cf. Partee 1997, p. 18). Consequently, the kind of "psycho-logic" we pursue here isn't necessarily in conflict with the general setting of model-theoretic semantics.

Information states are the fundamental entities in the construction of propositions. In the next section a reinterpretation of information states is given as representing states of activations in a connectionist network. In section 3 we consider how activation spreads out and

how it reaches, at least for certain types of networks, asymptotically stable output states. Following and extending ideas of Balkenius & Gärdenfors (1991), it is shown that the *fast dynamics* of the system can be described asymptotically as an non-monotonic inferential relation between information states. Section 4 introduces the notion of weight-annotated Poole systems, and section 5 explains how these systems bring about the correspondence between connectionist and symbolic knowledge bases. Finally, in section 6 we relate the present account to optimality theory (Prince & Smolensky 1993) – a general framework that aims to integrate insights from symbolism and connectionism. The paper concludes with some speculations about some extensions of the integrative story.

2 Information states in Hopfield networks

Connectionist networks are complex systems of simple neuron-like processing units (usually called "nodes") which adapt to their environments. In fact, the 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 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 single units. 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 characteristics 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 network 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 are Hopfield networks (Hopfield 1982). Such networks always stabilize, and 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.

For a good introduction into connectionist networks the reader is referred to the two volumes Rumelhart, McClelland, & the PDP group (1986) which are still a touchstone for a wide variety of work on parallel distributed networks. Other excellent introductions are Bechtel (2002) and Smolensky & Legendre (to appear).

For the following we start with considering neural networks as systems of connected units. Each unit has a certain *working range of activity.* Let this be the set [-1, +1] (+1: maximal firing rate, 0: resting, -1: minimal firing rate). A possible state *s* of the system describes the activities of each neuron: $s \in [-1, +1]^n$, with n = 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 ($-1 \leq w_{ij} \leq +1$, $w_{ij} = w_{ji}$, $w_{ii} = 0$). That means node i has the same effect on node j as node j on node i, and nodes don't affect themselves. The *fast dynamics* describes how neuron 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 linear 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).[1]

For the following it is important to interpret activations as indicating information specification: the activations +1 and -1 indicate maximal specification; the resting activation 0 indicates (complete) underspecification. It is this interpretation of the activation states that allows introducing the notation of information as an observer-dependent notion. Though this interpretation is not an arbitrary one, from a philosophical point of view it is important to stress the idea that information is not a purely objective, observer-independent unit. Instead, the observer of the network decides, at least in part, which aspects of the network are worth our consideration and which abstractions are appropriate for the observer's aims.

Generalizing an idea introduced by Balkenius & Gärdenfors (1991), the set $S = [-1, +1]^n$ of activation states can be partially ordered in accordance with their informational content:

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

$s \geq t$ can be read as *s is at least as specific as t*. The poset $\langle S, \geq \rangle$ doesn't form a lattice. 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.[2] It can be shown that the extended poset of activation states $\langle S \cup \perp, >> \rangle$ forms a DeMorgan lattice when we replace the former definition of the informational ordering as follows:

(3) $s \geq t$ iff $s_i = nil$ or $s_i \geq t_i \geq 0$ or $s_i \leq t_i \leq 0$, for all $1 \leq i \leq n$).

The operation $s \circ t = \sup\{s,t\}$ (CONJUNCTION) can be interpreted as the *simultaneous realization* of two activation states; the operation $s \oplus t = \inf\{s,t\}$ (DISJUNCTION) can be interpreted as some kind of generalization of two instances of information states. The COMPLEMENT s^* reflects a *lack* of information. The operations come out as follows:

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

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

[2] Intuitively, *impossible activation states* express clashes between positive and negative activation.

$$(5)\qquad (s\oplus t)_i = \begin{cases} \min(s_i, t_i), \text{ if } s_i,t_i\geq 0 \\ \max(s_i, t_i), \text{ if } s_i,t_i\leq 0 \\ s_i, \text{ if } t_i=\text{nil} \\ t_i, \text{ if } s_i=\text{nil} \\ 0, \text{ elsewhere'} \end{cases}$$

$$(6)\qquad (s^*)_i = \begin{cases} 1-s_i, \text{ if } s_i>0 \\ -1-s_i, \text{ if } s_i<0 \\ \text{nil, if } s_i=0 \\ 0, \text{ if } s_i=\text{nil} \end{cases}$$

The fact that the extended poset of activation states forms a DeMorgan lattice gives the opportunity to interpret these states as propositional objects ("information states").

3 Asymptotic updates and nonmonotonic inference

In general, updating an information state s may result in an information state f...f(s) that doesn't include the information of s. However, in what follows it is important to interpret updating as specification. If we want s to be informationally included in the resulting update, we have to "clamp" s somehow in the network. A technical way to do that has been proposed by Balkenius & Gärdenfors (1991). Let f designate the original update function (1) and $\underline{f}$ the clamped one, which can be defined as follows (including iterations):

(7) $\underline{f}(s) = f(s) \circ s;$
(8) $\underline{f^{n+1}}(s) = f(\underline{f^n}(s)) \circ s$

Hopfield networks (and other so-called *resonance systems*) exhibit a desirable property: for each given input state *s* the system stabilizes in a well-defined output state (it is of no importance here whether the dynamics is clamped or not).

The notion of resonance is a very universal one and can be defined for a dynamic system [S, f] in general (here S denotes the space of possible states and f denotes the dynamics of the system, i.e., the activation function with regard to a specific configuration w of the network).

(9) A state $s \in S$ is called a resonance of a dynamic system [S, f] iff

(i) $f(s) = s$
(Equilibrium)

(ii) For each $\varepsilon>0$ there exists a $0<\delta\leq\varepsilon$ such that for all $n\geq 1$
$|f^n(s')-s| < \varepsilon$ whenever $|s'-s| < \delta$
(Stability)

(iii) For each $\varepsilon>0$ there exists a $0<\delta\leq\varepsilon$ such that $\lim_{n\to\infty} f^n(s') = s$ whenever $|s'-s| < \delta$
(Asymptotic stabilility)

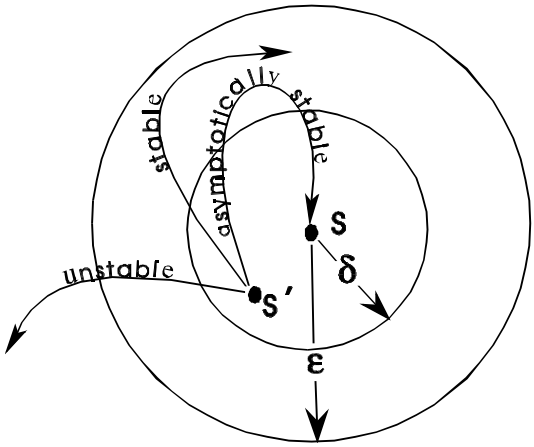

Figure 1: Stable, asymptotically stable, and unstable temporal developments of information states.

The existence of resonances is an emergent collective effect in neural nets. Intuitively, resonances are the stable states of the network and they *attract* other states. When each state develops into a resonance, then the system produces a content-addressable memory. Such memories have emergent collective properties (capacity, error correction, familiarity recognition; for details see Hopfield 1982).

A neural network is called a resonance system iff $\lim_{n\to\infty} (f^n (s))$ exists and is a resonance for each state $s \in S$ and each activation function f (relative to any network configuration $w \in W$). Cohen & Grossberg (1983) were the first who proved that Hopfield networks are resonance systems. The same proves true for a large class of other systems: The McCulloch-Pitts model (McCulloch & Pitts 1943), the Cohen-Grossberg model (1983), Rumelhart's Interactive Activation model (Rumelhart et al. 1986), Smolensky's (1986) Harmony networks etc. (for details see Grossberg (1989).

In the present context, the classical results can be borrowed to establish that the following set of *asymptotic updates* of s is well-defined:

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

In case of asynchronous (non-deterministic) updates, the function $E(s) = -\sum_{i>j} w_{ij} s_i s_j$ is a Ljapunov function (energy function) of the dynamic system (Hopfield 1982). I.e., when the activation state of the network changes, E can either decrease or remain the same. Hence, the output states $\lim_{n\to\infty} \underline{f}^n(s)$ can be characterized as *the local minima* of the Ljapunov-function.

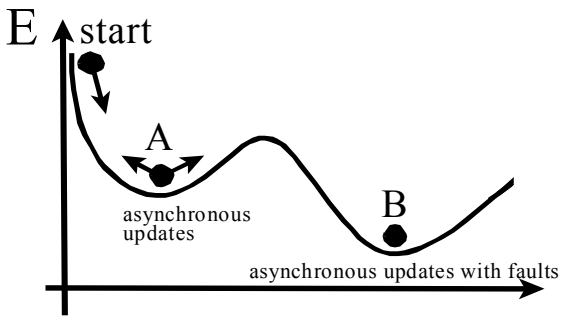

Figure 2: Local and global minima of the Ljapunov function. Local minima can be realized by asynchronous updates with a deterministic activation function while reaching the global minima often requires an *stochastic* activation function.

Usually, the stable state is not the state that would yield the lowest possible values of E (the global minima). The Boltzman machine (Hinton & Sejnowski 1983, 1986) is an adaptation of the Hopfield net that realizes the *global* minima, i.e. their output states $\lim_{n\to\infty} \underline{f}^n(s)$ 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.[3] The latter observation enables us to characterize the asymptotic updates of s as the set of all specifications of s that minimize the energy E of the system:

(11) $\mathrm{ASUP}_w(s) = \min_E(s)$.

The propositional objects called information states are related by the partial ordering ≥. It is obvious that this relation can be interpreted as a strict entailment relation. In any case it satisfies the Tarskian restrictions for such a relation:

(12) (i) $s \geq s$ (REFLEXIVITY)

(ii) if $s \geq t$ and $s \circ t \geq u$, then $s \geq u$ (CUT)

(iii) if $s \geq u$, then $s \circ t \geq u$ (MONOTONICITY)

More interesting, Balkenius & Gärdenfors (1991) have made clear that 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 naturally leads to a nonmonotonic inferential relation between information states:

(13) $s \mid\!\sim_w t$ iff $s' \geq t$ for each $s' \in \mathrm{ASUP}_w(s)$.

[3] The procedure is called "simulated annealing" (based on an analogy from physics).

It is with the help of the equivalence (11) that the usual traits of nonmonotonic consequence relations can be shown:

Theorem 1

Let $\mid\!\sim_w$ be a relation between information states as defined in (13). Then we have

(14)
(i)	if $s \geq t$, then $s \mid\!\sim_w t$	(SUPRACLASSICALITY)
(ii)	$s \mid\!\sim_w s$	(REFLEXIVITY)
(iii)	if $s \mid\!\sim_w t$ and $s \circ t \mid\!\sim_w u$, then $s \mid\!\sim_w u$	(CUT)
(iv)	if $s \mid\!\sim_w t$ and $s \mid\!\sim_w u$, then $s \circ t \mid\!\sim_w u$.	(CAUTIOUS MONOTONICITY)

The proofs for SUPRACLASSICALITY and REFLEXIVITY (clampedness!) are obvious. For CUT, suppose all E-minimal specifications of s are specifications of t and all E-minimal specifications of $s \circ t$ are specifications of u. Let s' be an E-minimal specification of s. s' specifies both s and t; consequently, it specifies $s \circ t$. Since $s \circ t \geq s$, it follows that s' is also an E-minimal specification of $s \circ t$. Consequently, it is a specification of u.

For CAUTIOUS MONOTONICITY, suppose all E-minimal specifications of s are specifications of t and u. We have to prove $\nu \geq u$ for each E-minimal specification ν of $s \circ t$. Let ν be any E-minimal specification of $s \circ t$. Of course, ν is a specification of s. We shall prove now that ν is an E-minimal specification of s. If this were wrong, there would be an E-minimal specification ν' of s such that $E(\nu') < E(\nu)$. But all E-minimal specifications of s are specifications of t, therefore $\nu' \geq t$ and $\nu' \geq s \circ t$. This contradicts the E-minimality of ν with respect to the specifications of $s \circ t$. Therefore ν must be an E-minimal specification of s. Since all E-minimal specifications of s are specifications of u, one concludes that $\nu \geq u$.

The results found so far correspond to the findings of Balkenius & Gärdenfors (1991), who have considered information states for cases where they form a Boolean algebra. The inferential notion that is adequate to describe the *fast dynamics* of the neural system (how neuron activities spread through the network) 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.

A simple example may help to illustrate the ideas introduced so far and to simplify the subsequent explanations. Let's consider a Hopfield network with a set of states $S = [-1, +1]^3$ and the connection matrix (15).

(15)
$$w = \begin{pmatrix} 0 & 0.2 & 0.1 \\ 0.2 & 0 & -1 \\ 0.1 & -1 & 0 \end{pmatrix}$$

Figure 3 shows the activation states of the network before and after updating. For the input state it is assumed that node 1 is activated and the other two nodes are resting (indicating underspecification). Clamping node 1, the fast dynamics yields an output state where node 2 is activated and node 3 is inhibited (minimal firing rate).

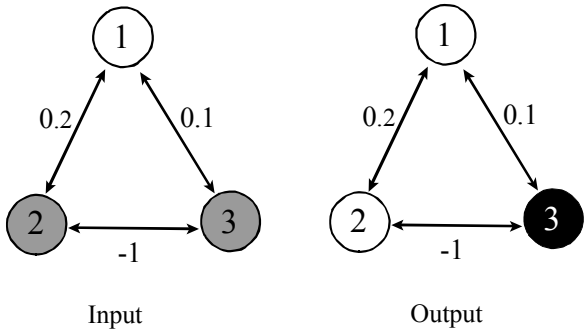

Figure 3: Activation states of a network with connection matrix (15) before and after updating. White, gray, black nodes indicate activation, resting, and inhibition, respectively.

In table 1 nine different possible specifications of the initial state <1 0 0> are shown, and their energy is calculated with regard to the connection matrix (15). The energy-minimal state is indicated by ☜. It corresponds to the output state represented in figure 3.

s [states]	*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	

Table 1: States and their energy in a network with connection matrix (15). ☜ indicates the energy-minimum state.

Using definition (13) and the equivalence (11) it is obvious that the following inferences are valid:

(16) (i) $<1\ 0\ 0> \mid\!\sim_w <1\ 1\text{-}1>$
(ii) $<1\ 0\ 0> \mid\!\sim_w <1\ 1\ 0>$
(iii) $<1\ 0\ 0> \mid\!\sim_w <0\ 1\ 0>$

The latter two inferences can be derived from the first one by taking into account that $<1\ 1\text{-}1> \geq <1\ 1\ 0> \geq <0\ 1\ 0>$.

4 Weight-annotated Poole systems

In connectionist systems “knowledge” is encoded in the connection matrix *w* (or, alternatively,

the energy function *E*). Symbolic systems usually take default logic and represent knowledge as a database consisting of expressions having default status. A prominent example of such a framework has been proposed by Poole (e.g. Poole 1988, 1996). In this section, we introduce a variant of Poole's systems, which we call weight-annotated Poole systems. This variant will be proven to be useful for relating the different types of coding knowledge (see section 5).

Let's consider the language $\mathcal{L}_{At}$ of propositional logic (referring to the alphabet At of atomic symbols). A triple T = <At, Δ, g> is called a *weight-annotated Poole system* iff (i) Δ is a set of consistent sentences built on the basis of At (the possible hypotheses); (ii) g: Δ ⇒ [0,1] (the weight function). *A scenario of a formula* α *in T* is a subset Δ' of Δ such that Δ'∪{α} is consistent. *The weight of a scenario* Δ' is

(17) $G(\Delta') = \sum_{\delta \in \Delta'} g(\delta) - \sum_{\delta \in (\Delta - \Delta')} g(\delta)$.

Hence, the weight 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 included beliefs give a positive contribution to the overall weight, and the missing beliefs give a negative contribution. This is different from a *penalty function* (Pinkas 1995) where only the *missing beliefs* count (i.e. the beliefs that are not included in the scenario/theory).

A *maximal scenario of* α *in T* is a scenario the weight of which is not exceeded by any other scenario (of α in T). With regard to a weight-annotated Poole system T, the following cumulative consequence relation can be defined:

(18) $\alpha \sim\!>_T \beta$ iff β is an ordinary consequence of each maximal scenario of α in T.

It is important to give a preference semantics for weight-annotated Poole systems. This preference semantics may be seen as the decisive link for establishing the correspondence between connectionist and symbolic systems.

Let ν denote an ordinary (total) interpretation for the language $\mathcal{L}_{At}$ (ν: At→{-1,1}). The usual clauses apply for the evaluation 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:

(19) $\mathcal{E}(\nu) = -\sum_{\delta \in \Delta} g(\delta) \, \|\delta\|_\nu$ ($\mathcal{E}$ is called the "energy" of the interpretation, or Poole's system energy)

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 a semantic counterpart to the syntactic notion $\alpha \sim\!>_T \beta$, let's take the following relation:

(20) $\alpha \approx\!>_T \beta$ iff each preferred model of α is a model of β.

As a matter of fact, the syntactic notion (18) and the semantic notion (20) coincide:

Theorem 2

For all formulae α and β of $\mathcal{L}_{At}$: $\alpha \sim\!>_T \beta$ iff $\alpha \approx\!>_T \beta$.

For the proof it is sufficient to show that the following two clauses are equivalent:

(A) There is a maximal scenario Δ' of α in T such that $\{\alpha\}\cup\Delta'\cup\{\neg\beta\}$ is consistent.

(B) There is a preferred model of α such that $\|\beta\|_\nu = -1$.

In order to prove this equivalence we have to state some simple facts which are immediate consequences from the corresponding definitions. Let T = <At, Δ, g> be a weight-annotated Poole system and let ν be any model. Then we have

(21) $G(sc(\Delta,\nu)) = -\mathcal{E}(\nu)$, where $sc(\Delta,\nu) =_{def} \{\delta\in\Delta: \|\delta\|_\nu = 1\}$ (the scenario associated with the model ν).

(22) $sc(\Delta,\nu) = \Delta'$ in case Δ' is a maximal scenario of α in T, and ν is a model of $\{\alpha\}\cup\Delta'$.

Now we are ready to prove the equivalence between (A) and (B).

(A) ⇒ (B): Let's assume that Δ' is a maximal scenario of α in T and ν is a model of $\{\alpha\}\cup\Delta'\cup\{\neg\beta\}$. We have to show that ν is a preferred model of α in T, i.e., we have to show that for each model ν' of α in T, $\mathcal{E}(\nu') \geq \mathcal{E}(\nu)$. From fact (21) it follows that $\mathcal{E}(\nu') = -G(sc(\Delta,\nu'))$, and the facts (21) & (22) necessitate $\mathcal{E}(\nu) = -G(\Delta')$. Since $sc(\Delta,\nu')$ is a scenario of α in T and Δ' is a maximal scenario, it follows that $\mathcal{E}(\nu') \geq \mathcal{E}(\nu)$.

(B) ⇒ (A): We assume a preferred model ν of α and assume $\|\beta\|_\nu = -1$. Obviously, the set $sc(\Delta,\nu)\cup\{\alpha\}\cup\{\neg\beta\}$ is consistent (ν is a model of it). We have to show now that the scenario $sc(\Delta,\nu)$ is a maximal scenario of α in T. If it were not, then there would exist a maximal scenario Δ' with $G(\Delta') > G(sc(\Delta,\nu))$. Because we have $G(\Delta') = -\mathcal{E}(\nu')$ for any model ν' of $\{\alpha\}\cup\Delta'$ and $G(sc(\Delta,\nu) = -\mathcal{E}(\nu)$ [facts (21) & (22)], this would contradict the assumption that ν is a preferred model of α in T. End of proof.

5 Relating connectionism and symbolism

In investigating the correspondence between connectionist and symbolic knowledge bases, we have first to look for a symbolic representation of information states. Let's consider again the propositional language $\mathcal{L}_{At}$, but now let's take this language as a symbolic means to speak about information states. Following usual practice in algebraic semantics, we can do this formally by interpreting (some subset of the) expressions of the propositional language by the corresponding elements of the DeMorgan algebra $<S\cup\bot, \geq>$. More precisely, let's call the triple $<S\cup\bot, \geq, \downharpoonright\ \downharpoonleft>$ a *Hopfield model* (for $\mathcal{L}_{At}$) iff $\downharpoonright\ \downharpoonleft$ is a function assigning some element of $S\cup\bot$ to each atomic symbol and obtaining the following conditions:

(23) (i) $\downharpoonright\alpha\wedge\beta\downharpoonleft = \downharpoonright\alpha\downharpoonleft\circ\downharpoonright\beta\downharpoonleft$

(ii) $\downharpoonright\sim\beta\downharpoonleft = -\downharpoonright\beta\downharpoonleft$ ("~" converts positive into negative activation and *vice versa*).

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

(24) $\uparrow p_1 \downarrow = \langle 1\ 0\ ...\ 0 \rangle$
$\uparrow p_2 \downarrow = \langle 0\ 1\ ...\ 0 \rangle$
...
$\uparrow p_n \downarrow = \langle 0\ 0\ ...\ 1 \rangle$

An information state s is said to be *represented* by a formula α of $\mathcal{L}_{At}$ (relative to a Hopfield model M) iff $\uparrow \alpha \downarrow = s$. With regard to our earlier example, the following formulae *represent* proper activation states: p_1 represents $\langle 1\ 0\ 0 \rangle$, p_2 represents $\langle 0\ 1\ 0 \rangle$, p_3 represents $\langle 0\ 0\ 1 \rangle$, $p_1 \wedge p_2$ represents $\langle 1\ 1\ 0 \rangle$, $\sim p_1$ represents $\langle -1\ 0\ 0 \rangle$, and $p_1 \wedge p_2 \wedge \sim p_3$ represents $\langle 1\ 1\ -1 \rangle$. With regard to local Hopfield models it is obvious that each state can be represented by a conjunction of literals (atoms or their inner negation). In other words, for local models each information state can be considered symbolic.

Local Hopfield models give us the opportunity to relate connectionist and symbolic knowledge bases in a straightforward way and to represent nonmonotonic inferences in neural (Hopfield) networks by inferences in weight-annotated Poole systems. The crucial point is the translation of the connection matrix w into an associated Poole system T_w. Let's consider a Hopfield system (n neurons) with connection matrix w, and let $At = \{p_1, ..., p_n\}$ be the set of atomic symbols. Take the following formulae α_{ij} of $\mathcal{L}_{At}$:

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

For each connection matrix w the *associated Poole system* is defined as $T_w = \langle At, \Delta_w, g_w \rangle$, where the following two clauses apply:

(26) (i) $\Delta_w = \{\alpha_{ij}: 1 \leq i < j \leq n\}$
(ii) $g_w(\alpha_{ij}) = |w_{ij}|$

Updating information states was treated as a kind of specification in section 3. Under certain conditions (viz., where there are no isolated nodes) it can be shown that each (partial) information state is completed asymptotically; i.e., in the asymptotic state the corresponding node activities are either +1 or –1. Consequently, $ASUP_w(s)$ contains only total information states. As a matter of fact, each total information state t corresponds to a total propositional interpretation function ν/t where $\nu/t\ (p_i) = t_i$. Now we have the following facts:

(27) (i) $\|p_i\|_{\nu/t} = t_i$
(ii) $\|\sim\alpha\|_{\nu/t} = -\|\alpha\|_{\nu/t}$, in case α is a literal
(iii) $\|\alpha \leftrightarrow \beta\|_{\nu/t} = \|\alpha\|_{\nu/t} \cdot \|\beta\|_{\nu/t}$
(iv) $t \geq \|\alpha\|$ iff $\|\alpha\|_{\nu/t} = 1$, in case α is a conjunction of literals
(v) $\mathcal{E}(\nu/t) = E(t)$, where $E(t) = -\sum_{i>j} w_{ij}\ t_i\ t_j$ is the energy function of a Hopfield network with the connection matrix w and $\mathcal{E}$ is Poole's system energy for the weight-annotated Poole-system T_w; cf. (19)

(27)(i-iv) are direct consequences of the corresponding definitions. (27)(v) expresses the equivalence between Poole's system energy and the Hopfield energy of an information state. In order to prove this equivalence we start with the definitions (25) and (26) and we get $\Delta_w = \{(p_i \leftrightarrow \text{sign}(w_{ij})\ p_j\colon 1 \leq i < j \leq n\}$. Next we use the definition (19) for $\mathcal{E}$ with regard to Δ_w and we employ the results (27)(i-iii). The equality (27)(v) is obvious if we take into account the equation $|w_{ij}| = g_w(p_i \leftrightarrow \text{sign}(w_{ij})\ p_j)$, which results from (26)(ii).

Together with theorem 2 (expressing that the proof procedure in weight-annotated Poole systems is sound and complete) and the fact that in a local Hopfield model each state is symbolic and can be represented by a conjunction of literals, the statement (27) (v) allows us to prove that nonmonotonic inferences based upon asymptotic updates can be represented by inferences in weight-annotated Poole systems:

Theorem 3

Let α and β be formulas that are conjunctions of literals. Assume further that the Poole system T is *associated* with the connection matrix *w*. Then

(28) $\lceil\alpha\rfloor \mid\sim_w \lceil\beta\rfloor$ iff $\alpha \approx\!>_T \beta$ (iff $\alpha \sim\!>_T \beta$).

For the simple proof we start with (27)(v) – the equivalence between Poole's system energy and the Hopfield energy of an information state. Then it is straightforward that the semantic notion $\alpha \approx\!>_T \beta$ (entailment in preferred models, cf. definition (20)) coincides with the nonmonotonic inferential relation $s \mid\sim_w t$ between information states (energy-minimal specifications, cf. equation (11) and definition (13)) assumed we have a local Hopfield model that realizes the correspondences $\lceil\alpha\rfloor = s$ and $\lceil\beta\rfloor = t$.

The result expressed by theorem 3 shows that we can use nonmonotonic logic to characterize asymptotically how neuron activities spread through the connectionist network. In particular, a weighted variant of Poole's logical framework for default reasoning has proven to be essential. Hence, the usefulness of nonmonotonic logic as a descriptive and analytic tool for analyzing emerging properties of connectionist networks has been illustrated.

Going back to our earlier example, figure 4 illustrates the close relation between the connection matrix in Hopfield systems and the corresponding default system (weight-annotated Poole system). Using the connection matrix *w* as given in (15), the corresponding Poole system is given by the following weight-annotated defaults.[4]

(29) $T_w = \langle \text{At}, \Delta_w, g_w \rangle$, where
(i) $\text{At} = \{p_1, p_2, p_3\}$
(ii) $\Delta_w = \{p_1 \leftrightarrow p_2,\ p_1 \leftrightarrow p_3;\ p_2 \leftrightarrow \sim p_3\}$
(iii) $g_w = \{[p_1 \leftrightarrow p_2, 0.2], [p_1 \leftrightarrow p_3, 0.1], [p_2 \leftrightarrow \sim p_3, 1]\}$

The translation mechanism can be read out from figure 4. It simply 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 $\sim$ of one of

[4] In Figure 4 the weights are represented as indexing the biconditionals in the corresponding defauls.

its arguments. Furthermore, the weights of the defaults have to be taken as the absolute value of the corresponding matrix elements.

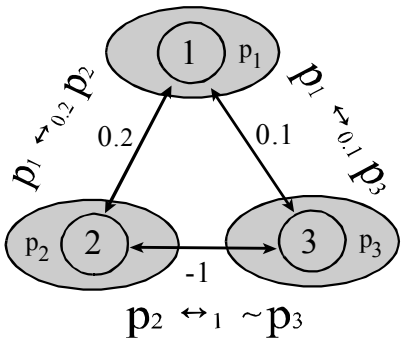

Figure 4: The correspondence between the connection matrix and a weight-annotated Poole system.

There is another perspective from which one may look at Theorem 3. One possible application of the Theorem 3 may be the use of connectionist techniques (such as "simulated annealing") to implement nonmonotonic inferences.[5] Seen in isolation, the latter point would favour the position of implementative connectionism – that connectionist ideas may be used only for the implementation of established symbolist systems. In the next section it is argued that already by considering local representations (where symbols correspond to single nodes in the network) connectionism provides remarkable insights that go beyond what usually is associated with the conception of implementation: connectionism is able to *explain* some peculiarities of symbolist systems.

6 Exponential weights and optimality theory

Non-trivial examples of weight-annotated Poole systems may be extracted from intrasegmental phonology. Intrasegmental phonology has been a source of inspiration for developing theories of markedness (e. g. Chomsky & Halle 1968, Kean 1975, 1981). In the present context it is used for demonstrating some aspects of integrative connectionism.

Table 2 presents a fragment of the vowel system of English (adapted from Kean 1975), which is a bit simplified for the present purpose. It contains a classification of the vowels in terms of the binary phonemic features *back*, *high* and *low*. The feature *round* has to be added in order to distinguish the segment /ɔ/ (rounded) from the segment /a/ (not rounded).

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

Table 2: Fragment of the vowel system of English

[5] For details cf. Derthick 1990.

For the purpose of formalization, 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 explicitly as in table 3 (left hand part) – a list that enumerates the feature specifications for each vowel segment.

	/a/	/i/	/o/	/u/	/ɔ/	/e/	/æ/
BACK	+	−	+	+	+	−	−
LOW	+	−	−	−	+	−	+
HIGH	−	+	−	+	−	−	−
ROUND	−	−	+	+	+	−	−

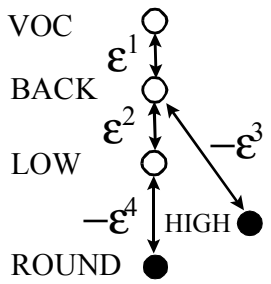

Table 3: Left: Feature specifications for a vowel fragment. Right: Hopfield network with exponential weights representing the generic knowledge of a phonological agent.

It is evident that this list contains strong (absolute) and weak (or "probabilistic") redundancies. For example, *all* segments that are classified as +HIGH are correlated with the specification –LOW (strong redundancy) and *most* segments that are classified as +BACK are correlated with the specification +LOW (weak redundancy).[6] Let's assume the following two hard constraints:

(30) (i) LOW → ~HIGH

(ii) ROUND → BACK.[7]

The generic knowledge of the phonological agent concerning this fragment may be expressed with regard to the hierarchy of features:

(31) <VOC, BACK, LOW, HIGH, ROUND>

Markedness conventions in the sense of Kean (1975, 1981), are rules for determining the (un)marked value of a feature *F* for a segment *x* given its value for a feature *G* preceding *F* in the feature hierarchy.

(32) (i) BACK is an unmarked property of the class VOC of vowels

(ii) LOW is an unmarked property of BACK (the class of back vowels)

[6] From the perspective of Universal Grammar, absolute constraints are universal in the sense that they are actually inviolate in every language. Weak constraints, on the other hand, embody universality in a *soft* sense. The idea is that each feature has two values or specifications, one of which is marked, the other unmarked. Unmarked values are crosslinguistically preferred and basic in all grammars, while marked values are crosslinguistically avoided and used by grammars only to create contrasts.

[7] The first constraint is universally valid for obvious reasons. Although the second constraint is satisfied for the present vowel fragment, it is not valid generally. Considering it as a hard constraint despite this fact is a rough stipulation that makes sense only in order to simplify the subsequent discussion.

(iii) ~HIGH is an unmarked property of BACK

(iv) ROUND is an unmarked property of ~LOW.[8]

There is a very general principle that allows to calculate the unmarked values of the *complements*:

(33) If αF is an unmarked property of βG, then $\sim\alpha F$ is an unmarked property of $\sim\beta G$, where "~" turns the value of α from + into –, and *vice versa*

For example, (33) allows us to determine +HIGH as an unmarked property of –BACK. It is intuitively clear now how the feature specifications in table 3 can be calculated if the shaded elements are taken to be given (marked) explicitly. For instance, in order to calculate the specifications for the unmarked vowel /a/ we go stepwise through the feature hierarchy and calculate +BACK by applying (32)(i), +LOW by applying(32)(ii), –HIGH by applying the hard constraint (30)(i), –ROUND by applying (32)(iv).

Taking the feature hierarchy (31) into account, the markedness conventions (32) can be represented as the elementary Hopfield network shown on the right-hand site of table 3. The technical means of expressing the hierarchy is the use of *exponential weights* with basis $0 < \varepsilon \leq 0.5$.[9] It should be mentioned that the implementation of the hard constraints (30) is not possible within a localist Hopfield network, and we will simply assume that these constraints are external and restrict the possible activations of the nodes in the network in this way.

The assigned Poole-system in the case under discussion has also to make use of exponential weights:

(34) (i) VOC $\leftrightarrow\varepsilon^1$ BACK
(ii) BACK $\leftrightarrow\varepsilon^2$ LOW
(iii) BACK $\leftrightarrow\varepsilon^3$ ~HIGH
(iv) LOW $\leftrightarrow\varepsilon^4$ ~ROUND

Theorem 3 ensures the equivalence between the connectionist and the symbolic treatment. Obviously, in the symbolic case the exponential ranking corresponds to a linear ranking of the defaults – with the hard constraints (30) on top. Using finite state transducers, the computational consequences of this view have been investigated by Frank & Satta (1998) and Karttunen (1998).

In cognitive psychology, the distinction between *automatic processing* and *controlled processing* has been demonstrated to be very useful (e.g. Shiffrin & Schneider 1977). *Automatic processing* is highly parallel, but limited in power. *Controlled processing* has powerful operations, but is limited in capacity. The distinction between these two types of

---

[8] Kean (1975, 1981) considers the more general case where more than one feature preceding F in the feature hierarchy is necessary to determine the (un)marked value of the feature *F*. An example is

(iv') ROUND is an unmarked property of BACK $\wedge$ ~LOW

The expression (iv) comes close to the formulation of (iv') if the hard constraint (30)(ii) is taken into account (see footnote 7).

[9] $\varepsilon = ½$ or smaller is a proper base in case of binary features which can applied only one time.

processing has been used, for example, for modelling lexical access, visual perception, problem solving and parsing strategies in natural language processing. In the present context it may be helpful to assume that the distinction correlates with the shape of the Ljapunov function. In a case where the network is built with exponential weights, the energy landscape is structured like high mountains with an obvious path to the valley of the global energy minimum. Moreover, the stochastic update procedure yields robust results given some fluctuation of the parameters corresponding to attention. Hence we have the tentative characterization of automatic processing. In the other case, however, the energy landscape is flat and there are many walls of comparable hight to cross before we see the valley of minimum energy. In this case, it is much more difficult for the adiabatic freezing mechanism to find the global optimum. Processing is slow and the attentional parameters may become much more influential. That corresponds to the traits of controlled processing. Although the present story has several elements of speculation it is obvious that the assumption of exponential weights has important consequences for the crucial characteristics of processing.

Theorem 2 opens a third way to calculate the consequences of activating a structure α, namely by determining the preferred (or *optimal*) model(s) of α. Table 4 shows a so-called OT tableau for an *input* α = Voc, which precisely illustrates this calculation.

Input: + VOC

	BACK	LOW	HIGH	ROUND	VOC ↕ BACK	BACK ↕ LOW	BACK ↕ ~HIGH	LOW ↕ ~ROUND
	+	+	–	+				*
☞	+	+	–	–				
	–	–	+	–	*			*
	+	–	+	+		*	*	
	+	–	+	–		*	*	*
	+	–	–	+		*		
	+	–	–	–		*		*
	–	+	–	–	*	*	*	
	–	–	–	–	*		*	*

Table 4: OT tableau for calculating the optimal vowels

As with weight-annotated Poole systems, OT looks for an optimal satisfaction of a system of conflicting constraints. Most importantly, the exponential weights of the constraints result in a *strict ranking* of the constraints, meaning that violations of many lower ranked constraints invariably count less than one violation of a higher ranked constraint (Prince & Smolensky 1993). The candidates can be seen as information (activation) states (left hand site of table 4). The Harmony (or NegEnergy H = -E) can be recognized immediately from the violations of the (strictly ranked) constraints. Violations are marked by * (right hand site of table4) and the optimal candidate is indicated by ☞. The nonmonotonicity of the OT framework corresponds to the fact that the optimal candidate(s) for a subset may be different from he optimal candidate(s) of the original set. This is demonstrated in table 5.

Input: + VOC ∧ +HIGH

	–	–	+	–	*			*
☞	+	–	+	+		*	*	
	+	–	+	–		*	*	*
	BACK	LOW	HIGH	ROUND	VOC ↕ BACK	BACK ↕ LOW	BACK ↕ ~HIGH	LOW ↕ ~ROUND

Table 5: OT tableau for calculating the optimal high vowels

Whereas in table (4) the segment /a/ comes out as the optimal vowel, in table 5 the segment /u/ comes out as the optimal high vowel.

Table 4 shows that *all* constraints are satisfied for the optimal candidate. Table 5 demonstrates a case where two constraints conflict. The conflicting constraints are: VOC ↔ BACK and BACK↔LOW. Whereas the first candidate /i/ violates the first constraint and satisfies the second constraint, the second candidate /u/ satisfies the first constraint and violates the second. The constraint conflict is resolved via a notion of differential strength. The second candidate wins because the first constraint is stronger (ranked higher) than the second.[10]

Optimality theory was originally proposed by Prince & Smolensky (1993) as a kind of symbolic approximation to the patterns of activation constituting mental representation in the domain of language. It relates to *Harmony Theory* (Smolensky 1986), a mathematical framework for the theory of connectionist computation. This framework makes it possible to abstract away from the details of connectionist networks.

The present study contrasts with this approach. It is based on a *particular* type of network, which exhibits simple attractive mathematical properties: Hopfield networks. Hence, it cannot be our aim to contribute to the explanation of the *general* correspondence between connectionism and optimality theory. Instead, the present study brings about the correspondence in a very singular case only. Using the idea of local representations (Hopfield models), we were able to provide an *explicit* connection between Hopfield networks and particular system of constraints and their ranking (Poole systems).

The present study gives a concrete model that may be helpful in deciding which defining principles of OT are derivable from connectionism and which are not. Following the general exposition in Smolensky (2000), it is obvious that the following principles can be derived from the connectionist setting:

- ***Optimality***: The correct output representation is the one that maximizes Harmony. This can be derived by taking harmony as H = - E, where E is the Ljapunov-function of the connectionist system.

[10] The third candidate violates the same constraints as the second one plus the additional constraint LOW ↔ ROUND. It can be seen as irrelevant for determining the optimal candidate.

- ***Containment***: Competition for optimality is between outputs that include the given input. This point is derived from the idea that clamping the input units restricts the optimization in a network to those patterns including the input.

- ***Parallelism***: Harmony measures the degree of simultaneous satisfaction of constraints. This is clearly expressed by definition (19).

- ***Conflict***: Constraints conflict: it is typically impossible to simultaneously satisfy them all. This point is derived from the fact that positive and negative connections typically put conflicting pressures on a unit-s activity.

- ***Domination***: Constraint conflict is resolved via a notion of differential strength: stronger constraints prevail over weaker ones in cases of conflict (consider the expression (19) again).

As it is made clear in Smolensky (2000), all these principles hold independently of the particular connectionist networks that have been considered.

The present treatment demonstrates that there is at least one principle that is derivable from the peculiarities of Hopfield networks only. This principle concerns the observation that all weak (violable) constraints have the form of bi-conditionals. As already mentioned, this principle follows from the symmetry of the weight matrix (section 2) and from the idea of local representations (section 5). Keans (1975, 1981) general principle (33) is a consequence of this fact. This principle is crucial for giving Kean's markedness theory its restrictive power, and it is amazing that it is motivated by the general structure of the underlying neural network.

Finally, there are at least two principles that are not derivable from connectionism but need some extra motivation.

- ***Strictness of domination***: Each constraint is stronger than all weaker constraints combined. This principle makes it possible to determine optimality without numerical computation. As we have seen this principle may be motivated by the assumption of an automatic processing mode.

- ***Universality***: The constraints are the same in all human grammars. This principle corresponds to a strong restriction on the content of the constraints. At the moment, it expresses an empirical generalization and it is absolutely unclear how to explain it (Smolensky 2000).

In section 1 the implementationalist position was mentioned – the position that downplays connectionism as a pure issue of the implementation of existing symbolic systems. Without doubt, connectionism can be used to implement existing symbolic systems. In the present context, for example, the Boltzman machine can be used in order to implement nonmonotonic inferences in weight-annotated Poole systems. However, connectionism is much more than a device to implement existing systems. Assuming connectionist ideas at the micro-level, the

development of optimality theory (Prince & Smolensky 1993) has shown that *non-classical* cognitive architectures may emerge with broad applications in linguistics.[11]

Another important point is the demonstration that aspects of possibly old-fashioned, "classical" architectures can be *explained* by assuming an underlying (connectionist) micro-level. In one case we have demonstrated that the underlying symmetric architecture of the Hopfield networks helps to motivate the general form of the symbolic system (principle (33)). Furthermore, an underlying connectionist system may help motivate the connection between the automatic processing mode and the strictness assumption of domination in ranked default systems). Hence, there are good reasons to accept the *integrative* methodology elucidated in section 1.

7 Related work

In the introduction to this paper I stressed the integrative methodology the present account is pursuing. Due to the pioneering work by Balkenius & Gärdenfors (1991) the results of such a programme can be useful both for traditional connectionist and for traditional symbolists. On the one hand, the results can help the connectionist better to understand their networks and to solve the so-called *extraction problem* (e.g. d'Avila Garcez, Broda & Gabbay 2001). On the other hand, the outcomes of the integrative methodology can help the symbolist to find more efficient implementations, for instance by using connectionist methods for solving "hard" problems such as optimization problems and constraint satisfaction problems.

Most of the authors that aim to bridge the gap between connectionism and symbolism are concentrated either on the extraction problem or on the problem of implementation. An example for the first approach is d'Avila Garcez, Broda & Gabbay (2001). These authors investigate feedforward networks and provide a complete and sound algorithm for extracting the underlying logical function that maps each vector of input activation to the corresponding output vector.

The second approach is pursued by Derthick (1990) and Pinkas (1995), *inter alias*. The work by Pinkas (1995) is particularly relevant. Very similarly to the present account, it maps preferred models into the minima of an energy function. However, there are also important differences. The first difference concerns the symbolic reasoning systems. Pinkas introduced *penalty logic*. Likewise to knowledge representation in weight-annotated Poole systems, a positive real number is assigned (the penalty) to every propositional formula representing domain knowledge. Importantly, these weighted beliefs are used differently in the two reasoning systems. In the case of weight-annotated Poole systems both the beliefs that are included in the scenario and the beliefs that are not included count, cf. the expression (17) repeated here:

(17) $G(\Delta') = \sum_{\delta \in \Delta'} g(\delta) - \sum_{\delta \in (\Delta - \Delta')} g(\delta)$.

In contrast, in penalty logic, a *penalty function* F is constructed that counts the *missing beliefs* only (i.e. the beliefs that are **not** included in the scenario/theory):

[11] Obviously, optimality theory has markedness theory as one of its predecessors. However, OT goes far beyond classical markedness theories. It includes a powerful learning theory (Tesar & Smolensky 2001) and applies to the phenomenon of gradedness (e.g. Boersma and Hayes 2001).

(35) $F(\Delta') = \sum_{\delta \in (\Delta - \Delta')} g(\delta)$.

Another important difference concerns the mapping between the expressions of the propositional default logic and the assigned symmetric (Hopfield) network. Pinkas seeks a connectionist implementation of his penalty logic. Most importantly, he is able to define a function that translates every set of standard propositional formulas (paired with penalties) into a strongly equivalent symmetric network. However, this function is not one-to-one since different logical systems can be connected with the same network. For instance, the two penalty logical systems $\psi_1 = \{<1: p \rightarrow q>\}$ and $\psi_2 = \{<1: q \rightarrow p>, <1: q>, <1: \neg p>\}$ realize the same (two nodes) network – a network that can be characterized by the energy function $E = p - pq$.

Surely, this fact isn't relevant when it comes to the connectionist *implementation* of penalty logic since in this case a unique symmetric network can be constructed for each set of expressions in penalty logic. The fact, however, becomes highly relevant when it comes to deal with the extraction problem and we have to extract the corresponding system of expressions in penalty logical for a given symmetric net. Usually, there is no unique solution to this task, and this leads to a complication of the extraction procedure since it isn't clear now which system should be extracted. This contrasts with the unique translation procedure proposed in the present article. In this case the proposed mapping between Hopfield nets and (a subset of) weight-annotated Poole systems is one-to-one. The mapping is a very transparent one and simply translates the links in the network to biconditionals in the logic.

From the view of implementation, of course, it is a disadvantage that the present default system is restricted to simple biconditionals $\alpha \leftrightarrow \beta$ where α and β are literals. However, such expressions seem to have a privileged status in certain cognitive systems, for example in theories of markedness in intrasegmental phonology (see section 6). This raises two important questions: the first asking to justify the special status of such a restricted systems; the second asking to overcome the restrictions by introducing hidden units and/or distributed representations.

8 Conclusions

Hopfield networks are, for (integrative) connectionists, to some extent what harmonic oscillators are for physicists and what propositional logic is for logicians. They are simple to study; their mathematical properties are pretty clear, but they have very restricted applications. The main advantage of concentrating on this simple type of network is a methodological one: it helps to clarify the important notions, it sharpens the mathematical instruments, and it provides a starting point for extending and modifying the simplistic framework.

In this vein, it has to be stressed that the primary aim of the present investigation is a methodological one: the demonstration that model-theoretic semantics may be very useful for analyzing (emerging properties of) connectionist networks. The main finding was that certain activities of connectionist networks can be interpreted as *nonmonotonic inferences*. In particular, there is a strict correspondence between Hopfield networks and particular nonmonotonic inferential systems (Poole systems). The relation between nonmonotonic inferences and neural computation was established to be of the type that holds between higher level and lower level systems of analysis in the physical sciences. (For example, statistical

mechanics explains significant parts of thermodynamics from the hypothesis that matter is composed of molecules, but the concepts of thermodynamic theory, like "temperature" and "entropy," involve no reference whatever to molecules.) Hence, our approach is a reductionist one – understanding reductionism in a way that sees *unification* as the primary aim and not *elimination* (cf. Dennett (1995) for a philosophical rehabilitation of reductionism).

Admittedly, the results found so far are much too simplistic to count as a real contribution for closing the gap between symbolism and connectionism. There are two main limitations. The first concerns the consideration of local representations only, where symbols correspond to single nodes in the network. Further, the present paper didn't make use of hidden nodes, which considerably restricts the capacity of knowledge representation. The second limitation concerns the rather static conception of node activity studied here, which precludes the opportunity to exploit the idea of coherent firing of neurons (temporal synchrony). As a consequence, we are confronted with a very pure symbolic system that fails to express constituent structure, variable binding, quantification, and the realization that consciousness and intentionality are prerequisites for cognition and knowledge (cf. Bartsch 2002).

There are several possible ways to overcome these shortcomings. First, a straightforward extension is to incorporate Pinkas' ideas concerning the role of hidden units into the present framework. Another aspect is to introduce distributed representations for realizing constituent structures and binding (e.g. Smolensky's (1990) tensor product representation[12]). An alternative possibility is to adopt the so-called *coherence view*, where a dynamic binding is realized by some sort of coherence (e.g., coherent firing of neurons – temporal synchrony: Shastri & Ajjanagadde 1993, Shastri & Wendelken 2000; circuit activity organized by attractors: Grossberg 1996, Bartsch 2002).

Optimality Theory (OT) has proposed a new computational architecture for cognition which claims to integrate connectionist and symbolic computation. Though too simple to give a full justification of OT's basic principles, the present account of unifying connectionism and symbolism can help to understand them. Especially, it may help to understand the hierarchical encoding of constraint strengths in OT. The solution to this particular 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)

Concluding, it is important to get an active dialogue between the traditional symbolic approaches to logic, information and language and the connectionist paradigm. Perhaps, this dialogue may stimulate the present discussion of founding the basic principles of Optimality Theory, and likewise it may shed new light on the old notions like partiality, underspecification, learning, genericity, probabilistic logic, and prototypicality.

Acknowledgement

My special thanks go to Johan van Benthem, Michiel van Lambalgen and Larry Moss who have encouraged me to pursue this line of research and gave valuable impulses and stimulation. Furthermore, I have to thank Anton Benz, Paul Doherty, Jason Mattausch, Oren Schwarz, Paul Smolensky, Anatoli Strigin, and Henk Zeevat. I am grateful to two anonymous referees for very helpful comments.

[12] For a critical review of this approach the reader is referred to Bartsch (2002), section 2.3.3.

References

d'Avila Garcez, A., Broda, K., & Gabbay; D. (2001): "Symbolic knowledge extraction from trained neural networks: A sound approach". *Artificial Intelligence* 125, 153-205.

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.

Boutsinas, B. & Vrahatis, M. (2001): "Artificial nonmonotonic neural networks". *Artificial Intelligence* 132, 1-38.

Bartsch, R. (2002): *Consciousness emerging*, John Benjamins, Amsterdam & Philadelphia.

Bechtel, W. (2002). *Connectionism and the mind*. Blackwell Publishers, Oxford.

Boersma, P. & Hayes, B. (2001), "Empirical tests of the gradual learning algorithm". *Linguistic Inquiry* 32, 45-86.

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.

Chomsky, N., & Halle, M. (1968): *The sound pattern of English*, Harper and Row, New York.

Dennett, D.C. (1995): *Darwin's dangerous idea*. Simon & Schuster, New York.

Derthick, M. (1990): "Mundane reasoning by settling on a plausible model". *Artificial Intelligence* 46, 107-157.

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

Frank, R. & Satta, G. (1998). "Optimality theory and the generative complexity of constraint violability". *Computational Linguistics* 24, 307-315.

Gabbay, D. (1985): "Theoretical foundations for non-monotonic reasoning in expert systems". In K. Apt (ed.) *Logics and models of concurrent systems*, Springer-Verlag, Berlin, 439-459.

Glymour, C. (2001): *The mind's arrows.* The MIT Press, Cambridge & London.

Grossberg, S. (1989): "Nonlinear neural networks: principles, mechanisms, and architectures". *Neural Networks* 1, 17-66.

Grossberg, S. (1996): "The attentive brain". *American Scientist* 83, 438-449.

Hendler, J.A. (1989): "Special issue: Hybrid systems (symbolic/connectionist)". *Connection Science* 1, 227-342.

Hendler, J.A (1991). "Developing hybrid symbolic/connectionist models". In J. Barnden & J. Pollack (eds.) *High-level connectionist models, advances in connectionist and neural computation theory, vol. 1*, Norwood, NJ: Ablex Publ. Corp.

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

Hinton, G.E. & Sejnowski, T.J. (1986): "Learning and relearning in Boltzman machines". In D.E. Rumelhart, J.L. McClelland, and the PDP research group, 282-317.

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

Karttunen, L. (1998). The proper treatment of optimality in computational phonology. Manuscript. Xerox Research Centre Europe.

Kean, M. L. (1975): *The theory of markedness in generative grammar*, Ph.D. thesis, MIT, Cambridge, Mass.

Kean, M. L. (1981): "On a theory of markedness". In R. Bandi, A. Belletti, and L. Rizzi (Eds.),

Theory of markedness in Generative Grammar. Estratto, Pisa, 559-604.

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). Lawrence Erlbaum Associates, Hilsdale, NJ.

Kraus, S., Lehmann, D. & Magidor, M. (1990): "Nonmonotonic reasoning, preferential models and cumulative logics". *Artificial Intelligence* 44, 167-207.

McCulloch, W.S. & Pitts, W. (1943): "A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics* 5, 115-133.

Partee, B. with Hendriks, H.L.W. (1997): "Montague Grammar". In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language*. MIT Press, Cambridge. Pp. 5-91.

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

Pinker, S., & Prince, A. (1988). "On language and connectionism: Analysis of a parallel distributed processing model of language acquisition". *Cognition* 28**,** 73–193.

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

Poole, D. (1996): "Who chooses the assumptions?" In P. O'Rorke (Ed.), *Abductive reasoning.* MIT Press, Cambridge.

Prince, A. & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar.* Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ.

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

Rumelhart, D.E., Smolensky, P. , McClelland, J.L. & Hinton, G.E. (1986) "Schemata and sequential thought processes in PDP models". In D.E. Rumelhart, J.L. McClelland, and the PDP research group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* MIT Press/Bradford Books, Cambridge, MA. Pp. 7-57.

Shastri, L. & Ajjanagadde V. (1993): "From simple associations to systematic reasoning". *Behavioral and Brain Sciences* 16, 417-494.

Shastri, L. & Wendelken, C. (2000): "Seeking coherent explanations – a fusion of structured connectionism, temporal synchrony, and evidential reasoning. Proceedings of Cognitive Science, Philadelphia, PA.

Shiffrin, R. M. & Schneider, W. (1977): "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory", *Psychological Review* 84, 127–190.

Smolensky, P. (1986): "Information processing in dynamical systems: Foundations of harmony theory". In D.E. Rumelhart, J.L. McClelland, and the PDP research group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations.* MIT Press/Bradford Books, Cambridge, MA. Pp. 194-281.

Smolensky, P. (1988): "On the proper treatment of connectionism". *Behavioral and Brain Sciences 11*, 1-23.

Smolensky, P. (1990): "Tensor product variable binding and the representation of symbolic structures in connectionist networks". *Artificial Intelligence* 46, 159-216.

Smolensky, P. (1996): "Computational, dynamical, and statistical perspectives on the

processing and learning problems in neural network theory". In P. Smolensky, M.C. Mozer, & D.E. Rumelhart (Eds.), *Mathematical Perspectives on Neural Networks.* Lawrance Erlbaum Publishers, Mahwah, NJ. Pp. 1-13.

Smolensky, P. (2000): Grammar-based connectionist approaches to language. *Cognitive Science* 23, 589-613.

Smolensky, P., Legendre, G., & Miyata, Y. (1992): "Principles for an integrated connectionist/symbolic theory of higher cognition". Technical Report CU-CS-600-92, Department of Computer Science, Institute of Cognitive Science, University of Colorado at Boulder.

Smolensky, P., & Legendre, G. (to appear). *The harmonic mind: From neural computation to optimality-theoretic grammar*. Blackwell, Oxford.

Tesar, B. & Smolensky, P. (2000), *Learnability in optimality theory*. MIT Press, Cambridge, Mass.