Nonmonotonic inferences in neural networks

Christian Balkenius
Cognitive Science
Department of Philosophy
University of Lund
S-223 50 Lund, Sweden
E-mail: Christian.Balkenius@fil.lu.se,

Peter Gärdenfors
Cognitive Science
Department of Philosophy
University of Lund
S-223 50 Lund, Sweden
E-mail: Peter.Gardenfors@fil.lu.se

Abstract

We show that by introducing an appropriate schema concept and exploiting the higher-level features of a resonance function in a neural network it is possible to define a form of nonmonotonic inference relation between the input and the output of the network. This inference relation satisfies some of the most fundamental postulates for nonmonotonic logics. The construction presented in the paper is an example of how symbolic features can emerge from the subsymbolic level of a neural network.

1 INTRODUCTION

Within cognitive science there is a controversy concerning the basic units of cognitive processing. On the one hand there are the so called classical theories (e.g., Fodor and Pylyshyn 1988) where it is argued that the basic units are *symbols* handled by rule-based processes. On the other hand, the connectionist school argues that we should approach cognition at another level and study how *neuronlike* elements interact to produce collectively emerging effects (e.g., Rumelhart et al. 1986).

We believe that it is possible to unify the symbol processing capabilities of the classical theories to the constraint satisfying capabilities of connectionist theories. We want to show that by developing a *high-level* description of the properties of neural networks it is possible to bridge the gap between the symbolic and the subsymbolic levels (see Smolensky 1988). The key concept for this construction will be that of *schemata*. To some extent inspired by earlier schema theories, we will introduce a general schema concept which is appropriate for studying neural networks on levels above the neurons (see Balkenius 1990). Certain operations on schemata will also be presented.

As an application of the schema concept for neural networks, the aim of this paper is to show that certain activities of such networks can be interpreted as *nonmonotonic inferences*. We shall study these inferences in terms of the general postulates for nonmonotonic logics that have recently been introduced in the literature. It seems that a large class of neural networks will generate so-called *cumulative* nonmonotonic inferences (Makinson 1989).

2 A CONCISE DESCRIPTION OF NEURAL NETWORKS

This section outlines a general description of a class of neural networks. The outline will be used as the starting point for the development of a high-level description based on schemata.

We can define a neural network N as a 4-tuple <S,F,C,G>. Here S is the space of all possible *states* of the neural network. The dimensionality of S corresponds to the number of parameters used to describe a state of the system. Usually $S=[a,b]^n$, where [a,b] is the working range of each neuron and n is the number of neurons in the system. We will assume that each neuron can take excitatory levels between 0 and 1. This means that a state in S can be described as a vector $x = \langle x_1,...,x_n \rangle$ where $0 \leq x_i \leq 1$, for all $1 \leq i \leq n$. The network N is said to be binary if $x_i = 0$ or $x_i = 1$ for all i, that is if each neuron can only be in two excitatory levels.

C is the set of possible *configurations* of the network. A configuration $c \in C$ describes for each pair i and j of neurons the connection c_{ij} between i and j. The value of c_{ij} can be positive or negative. When it is positive the connection is *excitatory* and when it is negative it is *inhibitory*. A configuration c is said to be *symmetric* if $c_{ij} = c_{ji}$ for all i and j.

F is a set of *state transition functions* or *activation functions*. For a given configuration $c \in C$, a function $f_c \in F$ describes how the neuron activities spread through that network.

G is a set of *learning functions* that describe how the configurations develop as a result of various inputs to the network. In the sequel, the learning functions will play no significant role.

The two spaces S and C interact by means of the difference equations

$$x(t+1) = f_{c(t)}(x(t))$$

$$c(t+1) = g_{x(t)}(c(t))$$

where $s \in S$, $f \in F$, $c \in C$ and $g \in G$.

This gives us two interacting subsystems in a neural network. First, we have the system <S,F> that governs the *fast* changes in the network, i.e., the transient neural activity. Then, we have the system <C,G> that controls the *slower* changes that correspond to all learning in the system. By changing the behaviour of the functions in the two sets F and G, it is possible to describe a large set of different neural mechanisms. Generally the state transition functions in F have much faster dynamics than the learning functions in G. We will assume that the state in C is fixed while studying the state transitions in S.

Example: In an Interactive Activation network (Rumelhart and McClelland 1986) with four nodes, S is the space $[min,max]^4$, C is the space of all 4×4 matrices, and $F = \{f_c(x)=(1-\theta)x+I(c,x) \mid c \in C, x \in S\}$. $I_i(c,x) = c_i x_i(max-x_i)$ if $c_i x_i > 0$ and $I_i(c,x) = c_i x_i(x_i-min)$ otherwise. Here the constant $1-\theta$ dampens the activation levels of the neurons and I_i describes the change of the activation level of x_i due to the influence from the other neurons.

The general description of a neural network given here comprises a large class of the systems presented in the literature, for example Hopfield (1984) nets, Boltzmann machines (Ackley et al. 1985), Cohen-Grossberg (1983) models, Interactive Activation models (Rumelhart et al. 1986), the McCulloch-Pitts (1943) model, the BSB model (Anderson et al. 1977), the Harmony networks (Smolen-sky 1986), and the BAM model (Kosko 1987).

3 SCHEMATA

The basic building block for many theories of cognition is the *schema*. Even though the concept seems to have as many definitions as authors, some common core exists in all of them. We will use the term schema as a collective name of the structures as used by Piaget (1952, 1973), Arbib and Hanson (1987), and Rumelhart et al. (1986). We also want to include concepts usually denoted by other names such as 'frames' (Minsky, 1981 1987), 'scripts' (Schank & Abelson, 1977), etc. Among the various proposals we find some common characteristics of schemata:

- Schemata can be used for representing objects, situations, and actions.
- Schemata have variables. There is thus some way of changing the schema to adapt it to different situations. As a consequence, schemata can be embedded. One schema can have another schema as a part or as an instantiation of a variable.
- Schemata support default assumptions about the environment. They are capable of filling in missing information.

Rumelhart et al. (1986) argue that schemata are sets of 'microfeatures' and show how they emerge from the collective behavior of small neuronlike elements. Each microfeature is represented by a neuronlike unit that interacts with the others by activating or deactivating them. This interpretation of schema lends itself to implementation in constraint satisfying networks (Ackley et al. 1985, Hinton and Sejnowski, 1988, Smolensky 1988, Hopfield 1982, 1984).

Arbib, Conklin and Hill (1987) take a higher level view and make a few additional assumptions about schemata. The framework for their notion is described as being "in the style of the brain." Their idea of schemata is similar to that of Rumelhart's in some respects, but ideas from traditional semantic nets are also used. The environment is represented as an assemblage of schemata, each of which corresponds to an object in the environment.

Minsky's notion of frames is yet another instance of a schema theory at a higher level (Minsky, 1987) with inspiration from both neural theory and semantic nets. If we look at other higher level accounts for schemata, such as Schank's 'scripts', we see more rigid structures. A schema is typically considered to be a set of variables and procedures. When we give the variables values we get an *instantiation* of a schema. This is also the definition of schemata usually used in the field of AI (Charniak & McDermott, 1985).

We want to argue that there is a very simple way of defining the notion of a schema within the theory of neural networks that has the desired properties. The definition we propose is that a schema α corresponds to a vector $<\alpha_1,...,\alpha_n>$ in the state space S. That a schema α is currently *represented* in a neural network with an activity vector $x = <x_1,...,x_n>$ means that $x_i \geq \alpha_i$, for all $1 \leq i \leq n$. An equivalent definition is to say that a schema is the *cone* $\alpha = \{z \in S: z_i \geq \alpha_i$, for all $1 \leq i \leq n\}$ generated by the vector $<\alpha_1,...,\alpha_n>$ and where α is represented in a neural network with activity vector x iff $x \in \alpha$. There is a natural way of defining a partial order of 'greater informational content' among schemata by putting $\alpha \geq \beta$ iff $\alpha_i \geq \beta_i$ for all $1 \leq i \leq n$. There is a minimal scheme in this ordering, namely $\mathbf{0} = <0,...,0>$ and a maximal element $\mathbf{1} = <1,...,1>$.

In the light of this definition, let us consider the general desiderata for schemata presented above. Firstly, it is

clear that depending on what the activity patterns in a neural network correspond to, schemata as defined here can be used for representing objects, situations, and actions.

Secondly, if $\alpha \geq \beta$, then β can be considered to be a more *general* schema than α, and α can thus be seen as an *instantiation* of the schema β. The part of α not in β is a *variable* instantiation of the schema β. This implies that all schemata with more information than β can be considered to be an instantiation of β with different variable instantiations. Thus, schemata can have variables even though they do not have any *explicit* representation of variables. Only the *value* of the variable is represented and not the variable as such. The index of the instantiation is identified with the added activity vector α-β.

This representation of variables instantiation avoids the combinatorial explosion of the tensor product representation of variable binding suggested by Smolensky (1991) but is weaker in power since it limits the possible embeddings of schemata. For example a schema cannot be recursively embedded into itself. However, the schema concept presented here can very well be used to represent the 'filler' and 'role' structures of Smolensky's constructions, as long as the sets of fillers and roles are disjoint.

Thirdly, the next sections will be devoted to showing that schemata support default assumptions about the environ-ment. The neural network is thus capable of filling in missing information.

There are some elementary operations on schemata that will be of interest when we consider nonmonotonic inferences in a neural network. The first operator is the *conjunction* $\alpha \bullet \beta$ of two schemata $\alpha = <\alpha_1,...,\alpha_n>$ and $\beta = <\beta_1,...,\beta_n>$ which is defined as $<\gamma_1,...,\gamma_n>$, where $\gamma_i = \max(\alpha_i,\beta_i)$ for all i. In terms of cones, $\alpha \bullet \beta$ is just the intersection of the cones representing α and β.

If we consider schemata as corresponding to observations in an environment, we can interpret $\alpha \bullet \beta$ as the *coincidence* of two schemata, i. e., the simultaneous observation of two schemata.

Secondly, the *complement* α^* of a schema $\alpha = <\alpha_1,...\alpha_n>$ is defined as $<1-\alpha_1,...,1-\alpha_n>$(recall that 1 is assumed to be the maximum activation level of the neurons, and 0 the minimum). In general, the comple-mentation operation does not behave like negation since, for example, if $\alpha = <0.5,...,0.5>$, then $\alpha^* = \alpha$. However, if the neural network is assumed to be binary, that is, if neurons only take activity values 1 or 0, then * will indeed behave as a classical negation on the class of binary-valued schemas.

Furthermore, the interpretation of the complement is different from the classical negation since the activities of the neurons only represent *positive* information about certain features of the environment. The complement α^* reflects a *lack* of positive information about α. It can be interpreted as a schema corresponding to the observation of everything but α. As a consequence of this distinction it is pointless to define implication from conjunction and complement. The intuitive reason is that it is impossible to observe an implication directly. A consequnce is that the ordering $\geq$ only reflects greater *positive* informational content.

However, something similar to classical negation can be constructed in a number of ways. We can let the schema $<0.5,...,0.5>$ represent a total lack of information. Greater activity will correspond to positive information and lesser activity to negative information. The ordering $\geq$ can be changed to reflect this interpretation if we let $\alpha \geq \beta$ iff $|\alpha_i-0.5| \geq |\beta_i-0.5|$ and α_i and β_i both lie on the same side of 0.5, for all i. In this ordering, $<0.5,...,0.5>$ is the minimal schema and **1** and **0** are both maximal.

Yet another way to construct a negation, which forces us to change the network, is to double the number of nodes in the network and let the new nodes represent the negation of the schema represented in the original nodes. This is achieved by letting the schemata obey the condition $x_i = 1-x_{n+i}$ (where n is the number of neurons in the original network). It is also possible to impose weaker conditions on the activity of the new nodes to reflect alternative negations.

These alternative ways of defining negation merit further studies. However, in this paper we will, for simplicity, confine ourselves to the first definition given above.

Finally, the *disjunction* $\alpha \oplus \beta$ of two schemata $\alpha = <\alpha_1,...,\alpha_n>$ and $\beta = <\beta_1,...,\beta_n>$ is defined as $<\gamma_1,...,\gamma_n>$, where $\gamma_i = \min(\alpha_i,\beta_i)$ for all i. The term 'disjunction' is appropriate for this operation only if we consider schemata to represent propositional information. Another interpretation that is more congenial to the standard way of looking at neural networks is to see α and β as two instances of a *variable*. $\alpha \oplus \beta$ can then be interpreted as the *generalization* from these two instances to an underlying variable.

It is trivial to verify that the De Morgan laws $\alpha \oplus \beta = (\alpha^* \bullet \beta^*)^*$ and $\alpha \bullet \beta = (\alpha^* \oplus \beta^*)^*$ hold for these operations. The set of all schemata forms a distributive lattice with zero and unit, as is easily shown. It is a boolean algebra if the underlying neural network is binary. In this way, we have alrady identified something that looks like a *propositional* structure on the set of *vectors* representing schemata.

4 RESONANT SCHEMATA

A desirable property of a network that can be seen as performing *inferences* of some kind is that it, when given a certain input, stabilizes in a state containing the results of the inference. In the theory of neural network such states are called resonant states.

In order to give a precise definition of this notion, consider a neural network $N = <S,F,C,G>$. Let us assume that the configuration c is fixed (or changes very slowly)

so that we only have to consider one state transitiion function f_c. For a fixed c in C, let $f_c^0(x) = f_c(x)$ and $f_c^{n+1}(x) = f_c \circ f_c^n(x)$. Then a state y in S is called *resonant* if it has the following properties:

(i) $f_c(y) = y$ (equilibrium)

(ii) If for any $x \in S$ and each $\varepsilon > 0$ there exists a $\delta > 0$ such that $|x-y| < \delta$, then $|f_c^n(x)-y| < \varepsilon$ when $n \geq 0$ (stability)

(iii) There exists a δ such that if $|x-y| < \delta$, then $\lim_{n\to\infty} f_c^n(x) = y$ (asymptotic stability).

Here |.| denotes the standard euclidean metric on the state space S. A neural system N is called *resonant* if for each fixed c in C and each x in S there exists a $n > 0$, that depends only on c and x, such that $f_c^n(x)$ is a resonant state.

If $\lim_{n\to\infty} f_c^n(x)$ exists, it is denoted by $[x]_c$, and $[.]_c$ is called the *resonance function* for c. It follows from the definitions above that all resonant systems have a resonance function. For a resonant system, we can then define *resonance equivalence* as $x \sim y$ iff $[x]=[y]$. It follows that $\sim$ is an equivalence relation on S that partitions S into a set of equivalence classes.

A system considered by Cohen and Grossberg (1983) can be written as

(1) $x_i(t+1) - x_i(t) = a_i(x_i)[b_i(x_i) - \Sigma_j c_{ij}d_j(x_j)]$

for a fixed $c \in C$, if $c_{ij} = c_{ji} \geq 0$, $a_i(x_i) \geq 0$ and $d_i'(x_i) \geq 0$. Here a_i which is supposed to be continuous and positive is called the amplification function and b_i the self-signal function. The function d_i which is assumed to be positive and increasing is called the other-signal function and describes how the output from a neuron depends on its activity. It may seem that equation (1) only describes systems where the neurons inhibit each other, but it has been shown that a number of neural models without the restriction $c_{ij} \geq 0$ can be rewritten in the form of (1) by the use of a simple change of coordinates (Grossberg 1989).

Theorem (Cohen and Grossberg 1983): Every trajectory of (1) approaches an equilibrium point.

A consequence of this theorem is that systems that can be described by an equation of the form (1) are resonant systems. Grossberg (1989) shows that the Cohen-Grossberg (1983) model, Hopfield (1984) nets, Boltzmann machines (Ackley et al. 1985), the McCulloch-Pitts (1943) model, the BSB model (Anderson et al. 1977), and the BAM model (Kosko 1987) are resonant systems. Furthermore, it is trival to show that the Harmony networks (Smolensky 1986) also can be described by an equation in the form of (1) and thus are resonant systems too. A common feature of these types of neural networks is that they are based on *symmetrical* configuration functions C, that is, the connections between two neurons are equal in both directions.

The function $[.]_c$ can be interpreted as filling in *default* assumptions about the environment, so that the schema represented by $[\alpha]_c$ contains information about what the network *expects* to hold when given α as input. Even if α only gives a partial description of, for example, an object, the neural network is capable of supplying the missing information in attaining the resonant state $[\alpha]_c$.

5 NONMONOTONIC INFERENCES IN A NEURAL NETWORK

We now turn to the problem of providing an *interpretation* of the activities of a neural network that will show it can perform nomonotonic inferences.

5.1 DEFINITION OF A NONMONOTONIC OPERATION

A first idea for describing the nonmonotonic inferences performed by a neural network N is to say that $[\alpha]_c$ contains the nonmonotonic conclusions to be drawn from α. However, in general we cannot expect the schema α to be included in $[\alpha]_c$, that is, $[\alpha]_c \geq \alpha$ does not always hold. Sometimes a neural network *rejects* parts of the input information – in pictorial terms it does not always believe what it sees.

So if we want α to be included in the resulting resonant state, we have to modify the definition. The most natural solution is to 'clamp' α in the network, that is, to add the *constraint* that the activity levels of all neurons is above α_i, for all i. Formally, we obtain this by first defining a function f_α via the equation $f_\alpha(x) = f(x) \bullet \alpha$ for all $x \in S$. We can then, for any resonant system, introduce the function $[.]_c^\alpha$ for a configuration $c \in C$ as follows:

$$[x]_c^\alpha = \lim_{n\to\infty} f_\alpha^n(x)$$

This function will result in resonant states for the same neural networks as for the function $[.]_c$ above. The reason is that the difference equation for f_α can be written in the form (1) by replacing all occurences of x by $x \bullet \alpha$ and then absorbing the conjunction into the functions a, b and d. Since the conditions of the theorem are also fullfilled for the new a, b and d, the Cohen-Grossberg theorem is thus applicable.

Since we are working with a fixed configuration c for a given network, we shall suppress the subscript c in the sequel.

The key idea of this paper is then to define a nonmonotonic inference relation $\mid\!\sim$ between schemata in the following way:

$$\alpha \mid\!\sim \beta \text{ iff } [\alpha]^\alpha \geq \beta$$

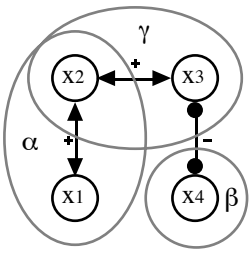
This definition fits very well with the interpretation that nonmonotonic inferences are based on *expectations* as developed in Gärdenfors (1991). Note that α and β in the definition are officially not *propositions* but schemata

that are defined in terms of *neural* activity vectors in a neural network. However, in the definition of $\sim$ they are *treated as* propositions. Thus, in the terminology of Smolensky (1988), we make the transition from the subsymbolic level to the symbolic simply by giving a different *interpretation* of the structure of a neural network. We do this without assuming two different systems as Smolensky does, but the symbolic level *emerges* from the subsymbolic in one and the same system (cf. Woodfield and Morton 1988). Just like a person may bid at an auction *by* raising his hand, a neural network may carry out symbolic inferences *by* performing subsymbolic operations. This kind of double interpretation of an information processing system is also discussed in Gärdenfors (1984).

The symbolic interpretation of neural networks that is based on the schema concept presented here can be applied to *all* neural networks. This is in contrast to some symbol-handling neural networks, like e.g., the μKLONE system of Derthick (1991), where the propositional structure is a starting point for the construction of the network.

Before turning to an investigation of the general properties of $\sim$ generated by the definition, we want to illustrate it by showing how it operates for a simple neural network.

Example: The network consists of four neurons with activities $x_1,...,x_4$. Neurons that interact are connected by lines. Arrows at the ends of the lines indicate that the neurons excite each other; dots indicate that they inhibit each other. If we consider only schemata corresponding to binary activity vectors, it is possible to identify schemata with *sets* of active neurons. Let three schemata α,β,γ correspond to the following activity vectors α=<1 1 0 0>, β=<0 0 0 1>, γ=<0 1 1 0>. Assume that x_4 inhibits x_3 more than x_2 excites x_3. Given α as input the network will activate γ, thus $\alpha \sim \gamma$. Extending the input to $\alpha\bullet\beta$ causes the network to withdraw γ since the activity x_4 inhibits x_3. In formal terms $\alpha\bullet\beta \not\sim \gamma$.



5.2 GENERAL PROPERTIES OF NONMONOTONIC OPERATIONS

One way of characterizing the nonmonotonic inferences generated by a neural network is to study them in terms of the general postulates for nonmonotonic logics that have recently been introduced in the literature (Gabbay 1985, Makinson 1989, 1991, Kraus, Lehmann, and Magidor 1990, Makinson and Gärdenfors 1990, Gärdenfors 1991). We shall present some of these postulates and determine whether they are satisfied for a function $[.]^\alpha$ determined by a transition function in a neural network.

It follows immediately from the definition of $[.]^\alpha$ that $\sim$ satisfies the property of *Reflexivity*:

$\alpha \sim \alpha$

If we say that a schema β follows logically from α, in symbols $\alpha \vdash \beta$, just when $\alpha \geq \beta$, then it is also trival to verify that $\sim$ satisfies *Supraclassicality*:

If $\alpha \vdash \beta$, then $\alpha \sim \beta$

In words, this property means that immediate consequences of a schema are also nonmonotonic consequences of the schema.

If we turn to the operations on schemata, the following postulate for conjunction is also trivial:

If $\alpha \sim \beta$ and $\alpha \sim \gamma$, then $\alpha \sim \beta \bullet \gamma$
(*And*)

More interesting are the following two properties:

If $\alpha \sim \beta$ and $\alpha \bullet \beta \sim \gamma$, then $\alpha \sim \gamma$
(*Cut*)

If $\alpha \sim \beta$ and $\alpha \sim \gamma$, then $\alpha \bullet \beta \sim \gamma$
(*Cautious Monotony*)

Together Cut and Cautious Monotony are equivalent to each of the following postulates:

If $\alpha \sim \beta$ and $\beta \vdash \alpha$, then $\alpha \sim \gamma$ iff $\beta \sim \gamma$

(*Cumulativity*)

If $\alpha \sim \beta$ and $\beta \sim \alpha$, then $\alpha \sim \gamma$ iff $\beta \sim \gamma$

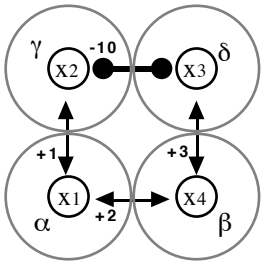
(*Reciprocity*)

Cumulativity has become an important touchstone for nonmonotonic systems (Gabbay 1985, Makinson 1989, 1991). It is therefore interesting to see that the inference operation defined here seems to satisfy Cumulativity for almost all neural networks where it is defined. However, it is possible to find cases where it is not satisfied:

Counterexample to Reciprocity: The network illustrated below is a simple example of a network that does not satisfy Reciprocity (or Cumulativity). This network is supposed to be *linear* in the sense that the total impact on the activity of a neuron can be written as a *weighted sum* of the excitatory and inhibitory inputs to the neuron, with the c_{ij}'s as weights (represented as +1, +2, +3, and -10 in the figure below).

If we assume that there is an excitatory connection between α and β, it follows that $\alpha \sim \beta$ and $\beta \sim \alpha$ since α and β do not receive any inhibitory inputs. Suppose that α

= <1 0 0 0> is given as input. Since we have assumed that the inputs to x_3 interact additively, it follows that γ receives a larger input than δ, because of the time delay before δ gets activated. If the inhibitory connection between γ and δ is large, the excitatory input from β can never effect the activity of x_3. We then have $\alpha \mathrel{|\!\sim} \gamma$ and $\alpha \not\mathrel{|\!\sim} \delta$. If instead β=<0 0 0 1> is given as input, the situation is the opposite, and so δ gets excited but not γ, and consequently $\beta \not\mathrel{|\!\sim} \gamma$ and $\beta \mathrel{|\!\sim} \delta$. Thus, the network does not satisfy Reciprocity.



A critical factor here seems to be the *linear* summation of inputs that locks x_2 and x_3 to inputs from the outside because the inhibitory connection between them is large.

We have performed extensive computer simulations of networks wich obey *shunting* rather than linear summation of excitatory and inhibitory inputs. The simulations suggest that Reciprocity is satisfied in all networks of this kind.

Shunting interaction of inputs is used in many biologically inspired neural network models (e.g., Cohen and Grossberg 1983, Grossberg 1989, Hodgkin 1964, Katz 1966, and Kuffer and Nichols 1976) and is an approximation of the membrane equations of neurons. A simple example of such a network can be described by the following equation:

$$(2) \qquad x_i(t+1) = x_i(t) + \delta(1-x_i(t))\Sigma_j d(x_i(t))c_{ji}^{+} + \\ + \delta x_i(t)\Sigma_j d(x_i(t))c_{ji}^{-}$$

Here δ is a small constant, c_{ij}^{+} and c_{ij}^{-} are matrices with all $c_{ij}^{+}=c_{ji}^{+}\geq 0$, and all $c_{ij}^{-}=c_{ji}^{-}\leq 0$; $d(x)\geq 0$ and $d'(x)>0$. The positive inputs to neuron x_i are shunted by the term $(1-x_i(t))$ and the negative inputs by $x_i(t)$. As a consequence, the situation where one input locks another of opposite sign cannot occur, in contrast to the linear case above. In other words, a change of input, that is a change in $\Sigma_j d(x_i(t))c_{ji}^{+}$ or $\Sigma_j d(x_i(t))c_{ji}^{-}$, will always change the equilibrium of x_i.

We believe that the fact that one input never locks another of opposite sign is the reason why all the simulated shunting networks satisfy Reciprocity.

Simulation example: As an example of a simulation we start from the network illustrated above. The matrix for the values of c_{ij}^{+} and c_{ij}^{-} in equation (2) is as follows:

Table 1: Connection Matrix

0	1	0	2
1	0	-10	0
0	-10	0	3
2	0	3	0

The function d in the equation was chosen as $d(x_i) = = (\max(x_i-\theta,0))^2/((\max(x_i-\theta,0))^2 + 0.2)$, where θ is an output threshold which was chosen to be 0.1. Finally, the constant δ was set to 0.005. (The values of the constants are not crucial for the outcome of the simulation).

In the simulation, the atomic schemata correspond to the activities x_1, ..., x_4 of the single neurons. For these neurons there are 16 different combinations of *binary* input schemata. In the table below these input schemata are indicated by bold face values. This represents clamped activities of the neurons according to the definition of the function $[x]_c^{\alpha}$.

The vectors in the table describe the resonant states induced by the different input vectors. Except for the limiting case **0**, there are four different resonant states for this network. It is easy to check that the nonmonotonic inference relation generated by this set of resonant states satisfies Reciprocity. For example, in case 2 we have $\gamma \mathrel{|\!\sim} \alpha$ and in case 3 we have $\gamma \bullet \alpha \mathrel{|\!\sim} \beta$ and in accordance with Reciprocity (or Cut), we also find in case 2 that $\alpha \mathrel{|\!\sim} \beta$.

Table 2: Resonant States

run	x_1	x_2	x_3	x_4
0:	0.00	0.00	0.00	0.00
1:	**1.00**	0.12	0.71	1.00
2:	1.00	**1.00**	0.23	1.00
3:	**1.00**	**1.00**	0.23	1.00
4:	1.00	0.09	**1.00**	1.00
5:	**1.00**	0.09	**1.00**	1.00
6:	1.00	**1.00**	**1.00**	1.00
7:	**1.00**	**1.00**	**1.00**	1.00
8:	1.00	0.12	0.71	**1.00**
9:	**1.00**	0.12	0.71	**1.00**
10:	1.00	**1.00**	0.23	**1.00**
11:	**1.00**	**1.00**	0.23	**1.00**

12:	1.00	0.09	**1.00**	**1.00**
13:	**1.00**	0.09	**1.00**	**1.00**
14:	1.00	**1.00**	**1.00**	**1.00**
15:	**1.00**	**1.00**	**1.00**	**1.00**

For the disjunction operation it does not seem possible to show that any genuinely new postulates for nonmontonic inferences are fulfilled. The following special form of transitivity is a consequence of Cumulativity (cf. Kraus, Lehmann, and Magidor (1990), p. 179):

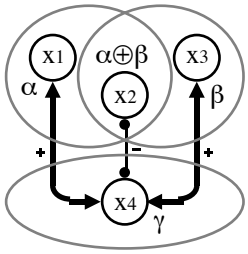
If $\alpha \oplus \beta \mathrel{|\!\sim} \alpha$ and $\alpha \mathrel{|\!\sim} \gamma$, then $\alpha \oplus \beta \mathrel{|\!\sim} \gamma$

This principle is thus satisfied whenever Cumulativity is.

The general form of Transitivity, i.e., if $\alpha \mathrel{|\!\sim} \beta$ and $\beta \mathrel{|\!\sim} \gamma$, then $\alpha \mathrel{|\!\sim} \gamma$, is not valid for all α, β, and γ, as can be shown by the first example above. Nor is the following principle generally valid:

If $\alpha \mathrel{|\!\sim} \gamma$ and $\beta \mathrel{|\!\sim} \gamma$, then $\alpha \oplus \beta \mathrel{|\!\sim} \gamma$
(Distribution)

Counterexample to Distribution: The following network is a simple counterexample: x_1 excites x_4 more than x_2 inhibits x_4. The same is true for x_3 and x_2. Giving α=<1 1 0 0> or β=<0 1 1 0> as input activates x_4, thus $\alpha \mathrel{|\!\sim} \gamma$ and $\beta \mathrel{|\!\sim} \gamma$. The neuron x_2 which represents schema $\alpha \oplus \beta$ on the other hand has only inhibitory connections to x_4. As a consequence $\alpha \oplus \beta \not\mathrel{|\!\sim} \gamma$.



6 CONCLUSION

We have shown that by introducing an appropriate schema concept and exploiting the higher-level features of a resonance function in a neural network it is possible to define a form of nonmonotonic inference relation. It has also been established that this inference relation satisfies some of the most fundamental postulates for nonmonotonic logics. The construction presented in this paper is an example of how symbolic features can emerge from the subsymbolic level of a neural network.

Acknowledgements

Research for this article has been supported by the Swedish Council for Research in the Humanities and Social Sciences. We want to thank the referees and the participants of the Cognitive Science seminar in Lund for helpful comments.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985): "A learning algorithm for Bolzmann machines," *Cognitive Science 9*, 147-169.

Anderson, J. A., Silverstein, J. W., Ritz, S. R. & Jones, R. S. (1977): "Distinctive features, categorial perception, and probability learning: Some applications of a neural model," *Psychological Review 84*, 413-451.

Arbib, M. A., Conklin, E. J., and Hill, J. C. (1987): *From Schema Theory to Language*. New York: Oxford University Press.

Arbib, M. A. and Hanson, A. R. (1987): *Vision, Brain, and Cooperative Computation*. Cambridge, MA: MIT Press.

Balkenius, C. (1990): "Neural mechanisms for self-organization of emergent schemata, dynamical schema processing, and semantic constraint satisfaction," manuscript, Cognitive Science, Department of Philosophy, Lund University.

Charniak, E., and McDermott, D. (1985): *Introduction to Artificial Intelligence*. New York: Addison-Wesley.

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

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

Fodor, J. and Pylyshyn, Z. (1988): "Connectionism and cognitive architecture: A Critical Analysis," in. S. Pinker & J. Mehler (eds.) *Connections and Symbols*. Cambridge, MA: MIT Press.

Gabbay, D. (1985): "Theoretical foundations for non-monotonic reasoning in expert systems," in *Logic and Models of Concurrent Systems*, K. Apt ed., Berlin: Springer-Verlag.

Gärdenfors P. (1984): "The dynamics of belief as a basis for logic," *British Journal for the Philosophy of Science 35*, 1-10.

Gärdenfors P. (1991): "Nonmonotonic inferences based on expectations: A preliminary report," these *Proceedings*.

Grossberg, S. (1978): "A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans," *Progress in Theoretical Biology, Vol. 5*, Academic Press, 233-374.

Grossberg, S. (1989): "Nonlinear Neural Networks: Principles, Mechanisms, and Architectures," in *Neural Networks*, Vol. 1, pp. 17-66.

Hinton, G. E. and Sejnowski, T. J. (1988): "Learning and relearning in Bolzmann machines," in Rumelhart, D.E., *Parallel Distributed Processing*, *Vol 1*, 282-317, Cambridge, MA: MIT Press.

Hodgkin, A. L. (1964): *The conduction of the nervous impulse*, Liverpool: Liverpool University Press.

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

Hopfield, J. J. (1984): "Neurons with graded response have collective computational properties like those of two-state neurons," *Proceedings of the National Academy of Sciences 81*.

Katz, B. (1966): *Nerve, muscle, and synapse,* New York: McGraw-Hill.

Kosko, B. (1987): "Constructing an associative memory," *BYTE Magazine 12*, 137-144.

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

Kuffer S. W. and Nicholls, J. G. (1976): *From neuron to brain,* Sunderland, MA: Sinauser Associates, 1976.

Makinson, D. (1989): "General theory of cumulative inference," in *Non-Monotonic Reasoning,* M. Reinfrank, J. de Kleer, M.L. Ginsberg, and E. Sandewall, eds., Berlin: Springer Verlag, Lecture Notes on Artificial Intelligence nº 346.

Makinson, D. (1991): "General patterns in nonmonotonic reasoning," to appear as Chapter 2 of *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume II: Non-Monotonic and Uncertain Reasoning*. Oxford: Oxford University Press.

Makinson, D. and Gärdenfors, P. (1990): "Relations between the logic of theory change and nonmonotonic logic," to appear in *Proceedings of the Konstanz Workshop on Theory Change*, A. Fuhrmann and M. Morreau, eds., Berlin: Springer Verlag.

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

Minsky, M. (1981): "A framework for representing knowledge," in J. Haugeland (Ed.), *Mind Design*, Cambridge, MA: The MIT Press.

Minsky, M. (1987): *The Society of Mind*. London: Heinemann.

Piaget, J. (1952): *The Origins of Intelligence in Children*. New York: International University Press.

Piaget, J. and Inhelder, B. (1973): *Memory and Intelligence*. London: Routledge & Kegan Paul.

Rumelhart, D. E. (1988): *Parallel distributed processing*, Vols I and II. Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L. and Hinton, G. E (1986): "Schemata and sequential thought processes in PDP models," in Rumelhart, D.E., *Parallel Distributed Processing*, *Vol 2*, 7-57, Cambridge, MA: MIT Press.

Schank, R. and Abelson, R. P. (1977): *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Smolensky, P. (1986): "Information processing in dynamical systems: foundations of harmony theory," in Rumelhart, D.E., *Parallel Distributed Processing*, Vol 1, 194-281, Cambridge, MA: MIT Press.

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

Smolensky, P. (1991): "Tensor product variable binding and the representation of symbolic structures in connectionist systems," *Artificial Intelligence 46*, 159-216.

Woodfield, A. and Morton, A. (1988): "The reality of the symbolic and subsymbolic systems," *Behavioral and Brain Sciences 11*, 58.