Nonmonotonic Reasoning by Inhibition Nets II*

Hannes Leitgeb

In Leitgeb[6] we have shown that certain networks called 'inhibition nets' may be regarded as mechanisms drawing nonmonotonic inferences. The main characteristic of inhibition nets is that there are not only excitatory connections between nodes but also inhibitory connections between nodes and excitatory connections. On the cognitive side, contents of belief are assigned to the patterns of activity in such networks, i.e., distributed representation is employed. An inhibition net together with an interpretation of its net states as belief states is called an 'interpreted inhibition net'. The state transitions which lead from an initial activity pattern to a final stable activity pattern are regarded as nonmonotonic inferences from an initial total belief to a final plausible belief. The nonmonotonicity of the inferences drawn by interpreted inhibition nets is due to the effect of inhibitory connections. In [6] it has been proved that the system CL (introduced by KLM[5], pp.186–189) of nonmonotonic reasoning is sound and complete with respect to the inferences drawn by interpreted finite hierarchical inhibition nets. In this paper the latter result is extended: we characterize further classes of interpreted inhibition networks, s.t. each of the cumulative logical systems studied by KLM[5] may be proved to be sound and complete with respect to one of the classes. Thus, there is an adequate cognitive network semantics for the systems C, CL, P, CM, and M of (nonmonotonic) logic. Inhibition nets are at the same time closely related to (i) logical systems of symbolic nonmonotonic reasoning in the style of KLM[5], (ii) mechanisms like logic programs or truth maintenance systems, and (iii) neural networks. We will briefly indicate some connections to logic programs in our final section 6 (also compare [6], section 6).We omit any discussion of the relationship between inhibition nets and neural nets (but see [6], section 7 for such a discussion). One can show that results similar to the ones that we prove for inhibition nets may also be achieved on the basis of logic programs, or on the basis of artificial neural networks, as long as a network semantics employing distributed representation is used. Balkenius&Gärdenfors [1] and Gärdenfors[4] have studied the relationship between nonmonotonic reasoning and artificial neural networks in a similar way, but without stating any formal results. The main motivation of this study is to show that very simple networks are able to reason according to the rationality constraints expressed in [5], if distributed representation is employed on the cognitive side. This might have consequences for our view of natural cognitive agents as nonmonotonic reasoners (see Leitgeb[7]). This paper is a successor to [6] and an extension of [8], where we have announced the results to be presented without proof and further discussion.

*This paper has been supported by the Austrian Research Fund FWF (SFB F012).

Keywords: Nonmonotonic reasoning; Inhibition networks; Belief and total belief; Network semantics; Cumulativity; KLM90 (Kraus, Lehmann, Magidor); Logic programs.

1. Inhibition Nets

Inhibition nets are directed graphs with two types of edges: (i) edges between nodes, and (ii) edges between nodes and edges of type (i):

Definition 1.1. *(Inhibition Nets)*

1. *Let N be a non-empty set (the set of nodes).*
2. *Let $E \subseteq N \times N$ (the set of excitatory connections).*
3. *Let $I \subseteq N \times E$ (the set of inhibitory connections).*
4. *Let $bias \in N$ be fixed (the bias node).*

Then $\mathcal{I} = \langle N, E, I, bias \rangle$ is an inhibition net(work).

In the following we use 'm' and 'n' (with or without indices) as variables ranging over nodes. We will only consider *finite* inhibition nets.

The nodes may be thought of as formal neurons, the excitatory connections between nodes as excitatory connections between neurons, and the inhibitory connections as presynaptic inhibitory connections. By means of the latter, neurons may inhibit excitatory connections between other neurons without inhibiting the target neurons of such connections themselves. Inhibition nets differ from usual artificial neural networks in having (i) no weigths attached to the connections, (ii) no inhibitory connections from nodes to other nodes, and, concerning dynamics, (iii) no continuous activation states for nodes, no weighted input summation within nodes, and no complex activation functions. On the other hand, if only artificial neural networks are considered where the output of a neuron is a binary signal, the dynamics of a neural network with a fixed set of weights may be shown to coincide with the dynamics of an inhibition network (see Leitgeb[6], section 7).

The bias node *bias* is the only node which is active in every state of the network. Thus we may assume that there is no $n \in N$ s.t. $n\ E\ bias$, since excitatory connections to *bias* would be without use. In [6] we have concentrated on a particular subclass of inhibition nets:

Definition 1.2. *(Hierarchical Inhibition Nets)*
An inhibition net $\mathcal{I}$ is hierarchical iff it does not have cycles (where paths are along excitatory connections, or *along inhibitory connections towards the target node of the inhibited connection).*

By an E-path we mean a path in an inhibition net which is along excitatory connections only.

In fig.1 you can see a finite hierarchical inhibition net (we abbreviate in the following by 'FHIN'); we have omitted the bias node graphically since it is assumed to have no influence on the other nodes in this example. Fig.2 depicts a non-hierarchical inhibition network:

Example 1.3. *Let* $\mathcal{I}_1 = \langle N_1, E_1, I_1, bias \rangle$, *s.t.* $N_1 = \{bias, n_1, n_2, n_3, n_4\}$, $n_1\ E_1\ n_2$, $n_1\ E_1\ n_3$, $n_4\ I_1\ \langle n_1, n_2 \rangle$, *and there are no other connections.*

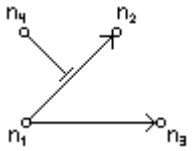

Fig. 1: $\mathcal{I}_1$

Example 1.4. *Let* $\mathcal{I}_2 = \langle N_2, E_2, I_2, bias \rangle$, *s.t.* $N_2 = \{bias, n_1, n_2, n_3\}$, $bias\ E_2\ n_2$, $bias\ E_2\ n_3$, $n_2\ E_2\ n_3$, $n_3\ E_2\ n_2$, $n_1\ I_2\ \langle n_2, n_3 \rangle$, $n_1\ I_2\ \langle n_3, n_2 \rangle$, $n_2\ I_2\ \langle bias, n_3 \rangle$, $n_3\ I_2$ $\langle bias, n_2 \rangle$, *and there are no other connections.*

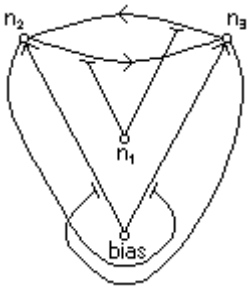

Fig. 2: $\mathcal{I}_2$

2. Inhibition Nets as Dynamical Systems

Inhibition nets may be considered as simple dynamical systems. We postulate that the nodes of inhibition nets have a certain binary state of activity, i.e. they are "on" (1) or "off" (0). We assume that nets receive external ("sensory") inputs which dictate certain nodes to fire independently of the current net state. We allow such inputs to affect every node in the network and not just the nodes of a distinguished layer of input nodes.

The internal causal dynamics of inhibition nets is defined by the evolution of states determined by the input and the topology of the network. The rule governing the state transitions within inhibition nets is as follows: a node n is excited if and only if (i) it is directly excited by the input, or (ii) there is an excitatory connection e from a further node m to n, s.t. m is itself active and e is not inhibited by yet another active node which is inhibitorily connected to e.

Put formally, this amounts to:

Definition 2.1. *(Dynamics of Inhibition Nets)*

Let $\mathcal{I} = \langle N, E, I, bias\rangle$ be an inhibition net.

Let $S = \{s \,|s : N \to \{0,1\}$ with $s(bias) = 1\}$ be the space of states of the net $\mathcal{I}$ (we omit the reference to $\mathcal{I}$ and just use 'S' instead of the more adequate '$S_{\mathcal{I}}$' for simplicity).

Let $s^* \in S$ be an arbitrary state of $\mathcal{I}$ (the "input"):

let $F_{s^*} : S \to S$, s.t. for all $n \in N \setminus \{bias\}$: $F_{s^*}(s)(n) = 1$ iff

1. *$s^*(n) = 1$, or*
2. *$\exists n_1 \in N\,(s(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(s(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$.*

F_{s^*} is the state transition function given relative to the input s^* and the net $\mathcal{I}$ (again we omit the reference to $\mathcal{I}$ and just say 'F_{s^*}' instead of '$F^{\mathcal{I}}_{s^*}$').

If s is a state of $\mathcal{I}$ and $s(n) = 1$, we say that n fires or that n is active (in s). A set of nodes is called 'active' if each of its members is active. We often identify a state (which is a mapping) with the set of neurons active in the very state: e.g., if we say that $s_1 \subseteq s_2$ we actually mean that for all $n \in N$: if $s_1(n) = 1$ then $s_2(n) = 1$; vice versa, we often identify sets of neurons with their characteristic functions.

The 'if' direction of the clause for F_{s^*} above says that if a node is caused to fire, it indeed fires; the 'only if' direction states that a node should only fire if it is also caused to fire. The inhibition of an excitatory connection is always dominant over any simultaneous impulse within the very excitatory connection. The bias node fires in every state s and commits the net to a certain preferred state of minimal energy which the net always reaches in the case of lacking input.

For each $s \in S$ (and each given input $s^* \in S$) the iterated application of F_{s^*} defines a trajectory s, $F_{s^*}(s)$, $F^2_{s^*}(s) = F_{s^*}(F_{s^*}(s))$, $F^3_{s^*}(s) = F_{s^*}(F_{s^*}(F_{s^*}(s)))$,... of states. $F^k_{s^*}(s)$ is the net state at time k given that s has been the initial state at time 0, and given the input s^* which is considered to be constant for a sufficient amount of time. $\langle S, F_{s^*}\rangle$ is a discrete dynamical system which is associated with the input s^* and the net $\mathcal{I}$. $(\langle S, F_{s^*}\rangle)_{s^* \in S}$ is a family of discrete dynamical systems associated with $\mathcal{I}$.

E.g., consider $\mathcal{I}_1$: if n_1 is the only node that fires at time 0, n_2 is caused to fire at time 1, but if both n_1 *and* n_4 fire initially, then n_2 does *not* fire at the next step due to inhibition. This is going to be the reason for the potential nonmonotonicity of the inferences drawn by inhibition nets. If $\{n_1\}$ is the constant input for $\mathcal{I}_1$, then the network reaches the stable state $\{n_1, n_2, n_3\}$, and this is the only stable state under the input $\{n_1\}$; if $\{n_1, n_4\}$ is the constant input for $\mathcal{I}_1$, then $\{n_1, n_3, n_4\}$ is the unique stable state. Here we use the following notion of a stable state:

Definition 2.2. *(Stable States)*

s is a stable state under input s^* iff $F_{s^*}(s) = s$, i.e. if s is a fixed point of F_{s^*}.

The existence of uniquely defined stable states in $\mathcal{I}_1$ under arbitrary inputs is a consequence of the fact that $\mathcal{I}_1$ is an FHIN (see Leitgeb[6], p.170):

Theorem 2.3. *(Stability Property for FHINs)*
For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$, for every $s^* \in S$ there is exactly one stable state s of $\mathcal{I}$ under the input s^*.

This justifies the following definition:

Definition 2.4. *(Closure Operator for FHINs)*
For every FHIN $\mathcal{I}$ let $Cl : S \to S$, s.t. $Cl(s^*)$ is the unique stable state under input s^* (again actually $Cl = Cl_{\mathcal{I}}$, but we drop the index '$\mathcal{I}$'). Cl is the closure operator of $\mathcal{I}$, $Cl(s^*)$ is the closure of s^*.

Stable (resonant, equilibrium) states play an excellent role in the literature on neural networks. Often they are considered to be the "answers" of neural networks to inputs ("questions"), and this is also our motivation for studying such states. FHINs do not only possess unique stable states $Cl(s^*)$ for all input s^*, but the states of an FHIN may also be shown to finally *converge* to $Cl(s^*)$ under the constant input s^*, where the selection of the initial state s is irrelevant (see Leitgeb[6], pp.170f):

Theorem 2.5. *(Convergence Property for FHINs)*
For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$, every input s^* and every initial state s:
there is an $i \in \mathbb{N}$ with $F^i_{s^*}(s) = Cl(s^*)$.

Now let us generalize the notion of closure to finite inhibition nets which are not necessarily hierarchical. $Cl(s^*)$ may be defined as the unique stable state to which every state converges under the constant input s^*, *given there is such a stable state*; otherwise $Cl(s^*)$ is left undefined:

Definition 2.6. *(Closure Operator)*
For every finite inhibition net $\mathcal{I}$, let $Cl : S \to S$ be a partial mapping, s.t. for all $s^* \in S$:
if there is a state s' which is stable under the input s^*, s.t. for all $s \in S$ there is an $i \in \mathbb{N}$ with $F^i_{s^*}(s) = s'$, then let $Cl(s^*) := s'$.

The uniqueness of closure states follows immediately from the properties that we have postulated for s', i.e., Cl is well-defined. Closure operators for FHINs are total by theorems 2.3 and 2.5. But it is easy to see that this is not the case in general: there are non-hierarchical nets $\mathcal{I}$ and inputs s^*, s.t. there is no stable state under s^*. E.g., there is no closure for $\{n_1\}$ in $\mathcal{I}_2$, since the activity states of n_2 and n_3 oscillate from being both active to being both inactive and vice versa under the constant input $\{n_1\}$; all other states of $\mathcal{I}_2$ indeed have closure states. Note that there also exist non-hierarchical nets $\mathcal{I}$ and inputs s^*, s.t. there is more than one stable state under s^*. If the closure of an input state is defined, every initial state of the network converges to the closure state, and thus the closure state is only dependent on the fixed input state and not on some initial state of the network. Later we will exploit this input-determinedness of closure states in the way that we interpret input as premises of an inference and closures as their corresponding conclusions; input-determinedness entails that the conclusions only depend on the premises.

Where it is defined, Cl has the following obvious properties:

Remark 1. *(Properties of Cl)*
For every finite inhibition $\mathcal{I} = \langle N, E, I, bias\rangle$, for every state s which has a closure state:

1. *$s \subseteq Cl(s)$ (Inclusion).*
2. *$Cl(s) = Cl(Cl(s))$ (Idempotence).*

Due to the presence of inhibitory connections, the state transitions in inhibition nets are not generally monotonic, i.e. if $s_1 \subseteq s_2$ then it does not necessarily follow that also $Cl(s_1) \subseteq Cl(s_2)$. E.g. in the case of $\mathcal{I}_1$ we have $\{n_1\} \subseteq \{n_1, n_4\}$ but $Cl(\{n_1\}) = \{n_1, n_2, n_3\} \not\subseteq \{n_1, n_3, n_4\} = Cl(\{n_1, n_4\})$. As substitutes for monotonicity, the following two weakenings of monotonicity may be proved for those states of finite inhibition nets which have closure states, and thus in particular for *all* states of finite *hierarchical* inhibition nets (we have proved the latter but not the former in Leitgeb[6], pp.171f):

Lemma 2.7. *(Cumulativity)*
For every finite inhibition net $\mathcal{I} = \langle N, E, I, bias\rangle$, for all states s_1, s_2 which have closure states:

$$\text{if } s_1 \subseteq s_2 \subseteq Cl(s_1), \text{ then } Cl(s_1) = Cl(s_2).$$

Proof:

Let s_1, $s_2 \subseteq N$ have closure states. We show that $F_{s_2}(Cl(s_1)) = Cl(s_1)$; after showing that we are done since $F_{s_2}(Cl(s_2)) = Cl(s_2)$ by the definition of Cl, and thus it follows from the uniqueness of closure states that $Cl(s_1) = Cl(s_2)$.
By definition of F_{s^*} we have for all $n \in N \setminus \{bias\}$:
$F_{s_1}(Cl(s_1))(n) = 1$ *iff*
$s_1(n) = 1$ *or*
$\exists n_1 \in N\,(Cl(s_1)(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(Cl(s_1)(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$
iff, since $F_{s_1}(Cl(s_1)) = Cl(s_1)$, $Cl(s_1)(n) = 1$;
thus it follows for all $n \in N \setminus \{bias\}$:
$F_{s_2}(Cl(s_1))(n) = 1$ *iff, by def of F_{s^*},*
$s_2(n) = 1$ *or*
$\exists n_1 \in N\,(Cl(s_1)(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(Cl(s_1)(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$
iff, since $s_2 = s_1 \cup s_2$,
$s_1(n) = 1$ *or* $s_2(n) = 1$ *or*
$\exists n_1 \in N\,(Cl(s_1)(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(Cl(s_1)(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$
iff, because of what we have shown for $F_{s_1}(Cl(s_1))$ above,
$s_2(n) = 1$ *or* $Cl(s_1)(n) = 1$ *iff, because* $Cl(s_1) \supseteq s_2$,
$Cl(s_1)(n) = 1$.
So we have $F_{s_2}(Cl(s_1)) = Cl(s_1)$ as claimed above. ■

In the case of FHINs we can add (see Leitgeb[6], pp.172f):

Lemma 2.8. *(Loop for FHINs)*
For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$*, for all states* $s_0, \ldots, s_j$*:*

$$\text{if } s_1 \subseteq Cl(s_0), s_2 \subseteq Cl(s_1), \ldots, s_j \subseteq Cl(s_{j-1}), s_0 \subseteq Cl(s_j),$$
$$\text{then } Cl(s_r) = Cl(s_{r'}) \text{ for all } r, r' \in \{0, \ldots, j\}.$$

For the terminology of 'cumulativity', 'loop', etc., see Makinson [11] and KLM[5].
If there are no inhibitory connections, Cl *is of course monotonic:*

Lemma 2.9. *(Monotonicity for Nets without Inhibition)*
For every finite inhibition net $\mathcal{I} = \langle N, E, \varnothing, bias \rangle$ *(i.e. without inhibitory connections) the operator* Cl *is monotonic.*

Let us now consider two important kinds of sets of nodes within FHINs and their corresponding closure properties, which we will need later for the representation theorem for the system P. We devote the subsequent section to this topic.

3. Two Important Kinds of Sets of Nodes within FHINs

Later we are going to define the notions of a (i) preferential partially interpreted *antitone* inhibition network, and a (ii) preferential partially interpreted *odd* inhibition network. It will be shown that the class of preferential partially interpreted nets which are antitone is precisely the class of networks, which are disposed to draw inferences obeying the rules of the well-known nonmonotonic system P. The class of preferential partially interpreted networks which are odd will be proved to be a proper subclass of the latter class. The property of being antitone depends on the way the closure operator "behaves" in such networks; the property of being odd will be defined more directly by stating a constraint on the topology of networks, and thus is more informative than antitonicity. This is the main reason why oddness is interesting in itself, although the system P is not complete with respect to the class of preferential odd networks, but only sound, contrary to the class of preferential antitone networks, relative to which P is both sound *and* complete as we will prove later.

In order to be able to define in the subsequent section what odd, or antitone, preferential partially interpreted inhibition networks are, we have first to specify two auxiliary notions: the notion of a set of nodes being odd in a network, and the notion of a set of nodes being antitone in a network. We will furthermore prove some of their properties:

Definition 3.1. *(Odd)*
Let $\mathcal{I} = \langle N, E, I, bias \rangle$ *be an FHIN. Let* $\bar{N} \subseteq N$ *(later we will always assume that* $bias \in \bar{N}$*).*
$\bar{N}$ *is odd in* $\mathcal{I}$ *iff*
there is no path $n_0^1, \ldots, n_{k_1}^1, n_0^2, \ldots, n_{k_2}^2, \ldots, n_0^r, \ldots, n_{k_r}^r$ *in* $\mathcal{I}$ *with*

- $n_0^1 \in \bar{N}$, $n_0^1 \neq bias$, $n_{k_r}^r \in \bar{N}$,
- *for all* $i \in \{1, \ldots, r\}$, $j \in \{0, \ldots, k_i - 1\}$: $n_j^i \; E \; n_{j+1}^i$,

- *for all* $i \in \{1, \ldots, r-1\}$*: there is an* $m \in N$ *s.t.* $n^i_{k_i}$ I $\langle m, n^{i+1}_0 \rangle$,
- *and* $r-1$ *is even.*

See fig.3 for a "forbidden" path (given $\bar{N}$ is odd in $\mathcal{I}$, and given that the first and the last node in the path are members of $\bar{N}$):

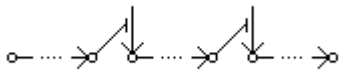

Fig.3: A "Forbidden" Path for Odd Subsets

Remark 2. *Def.3.1 implies that, if* $\bar{N}$ *is odd in* $\mathcal{I}$*, there are no* $n_0, n_1 \in \bar{N}$ *with* $n_0 \neq bias$*, s.t.* n_0 E n_1*; this says, roughly, that a node in a odd set* $\bar{N}$ *must not have any direct, excitatory influence on other nodes in* $\bar{N}$*. More generally, by def.3.1, a node in a odd set* $\bar{N}$ *must also not have any indirect excitatory influence on other nodes in* $\bar{N}$*, via an even number of inhibitions, s.t. the last inhibitory connection blocks a path originating from the bias node. Let us draw an analogy: suppose, for some reason, we want to exclude "directly positive" formulas* α*, i.e. formulas without negation (but just with, say, disjunction), from a propositional language; then we might also want to exclude the "indirectly positive" formulas of the form* $\neg\neg\alpha$*,* $\neg\neg\neg\neg\alpha, \ldots$*, where* α *is "directly positive", since the latter are logically equivalent to the former. In a similar way, direct and indirect excitatory connections between nodes of odd sets are excluded.*

Corollary 3.2. *Let* $\mathcal{I} = \langle N, E, I, bias \rangle$ *be an FHIN. Let* $\bar{N} \subseteq N$*, s.t.* $\bar{N}$ *is odd in* $\mathcal{I}$*.*

If there are nodes n_1*,* n_2*,* $n_3 \in \bar{N}$*, s.t.* n_2 *lies on a path leading from* n_1 *to* n_3*, then either* $n_1 = n_2$*, or* $n_2 = n_3$*.*

Proof:

If $n_1 \neq n_2$*, and* $n_2 \neq n_3$ *(*$n_1 \neq n_3$ *since* $\mathcal{I}$ *is an FHIN), it follows from* $\bar{N}$ *being odd in* $\mathcal{I}$ *that there is a path from* n_1 *to* n_2 *having an odd number of inhibitions, and that there is a path from* n_2 *to* n_3 *having an odd number of inhibitions. By concatenating those paths, there is a path from* n_1 *to* n_3 *having an even number of inhibitions, which contradicts* $\bar{N}$ *being odd in* $\mathcal{I}$*.* ■

For odd sets in FHINs one can show a distribution property of closure states:

Lemma 3.3. *(*$\bar{N}$*-Distribution I)*

For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$*, for every* $\bar{N} \subseteq N$*, s.t.* $bias \in \bar{N}$ *and* $\bar{N}$ *is odd in* $\mathcal{I}$*, it holds for all states* $s_1, s_2 \subseteq \bar{N}$*:*

$$\bar{N} \cap Cl(s_1) \cap Cl(s_2) \subseteq \bar{N} \cap Cl(s_1 \cap s_2).$$

Proof:

Let $n_1 \in \bar{N}$*: assume that* $Cl(s_1)(n_1) = 1$*,* $Cl(s_2)(n_1) = 1$*, and, for contradiction, suppose that* $Cl(s_1 \cap s_2)(n_1) = 0$*.*

1. *In this case* $n_1 \notin s_1 \cap s_2$, *and thus we may assume that* $n_1 \notin s_1$, *without loss of generality. But since* $Cl(s_1)(n_1) = 1$, *and since* $\mathcal{I}$ *is finite, it follows that there is a (not necessarily unique) E-path* $m^1_0, \ldots, m^1_{k_1}$ *of maximal length, s.t.* $Cl(s_1)(m^1_0) = 1$, $m^1_{k_1} = n_1$, *and there is no node* m *s.t.* $Cl(s_1)(m) = 1$, m I $\langle m^1_i, m^1_{i+1}\rangle$ *for some* $i \in \{0, \ldots, k_1 - 1\}$. *Since* $\bar{N}$ *is odd in* $\mathcal{I}$, m^1_0 *cannot be a member of* $\bar{N} \setminus \{bias\}$. *Thus, also* $m^1_0 \notin s_1 \setminus \{bias\}$. *Moreover, there is no node* m *s.t.* m E m^1_0, *and* $Cl(s_1)(m) = 1$, *because this would contradict the maximality of the path* $m^1_0, \ldots, m^1_{k_1}$. *Therefore, because of* $Cl(s_1)(m^1_0) = 1$, *it follows that* $m^1_0 = bias$.

2. $Cl(s_1 \cap s_2)(n_1) = 0$ *implies that there has to be a node* n_2, *s.t.* $Cl(s_1 \cap s_2)(n_2) = 1$, *and* n_2 I $\langle m^1_{i_1}, m^1_{i_1+1}\rangle$ *for some* $i_1 \in \{0, \ldots, k_1 - 1\}$. $n_2 \neq n_1$, *for* $\mathcal{I}$ *being hierarchical. From above we know that* $Cl(s_1)(n_2) = 0$, *therefore* $n_2 \notin s_1$, *and so we also have that* $n_2 \notin s_1 \cap s_2$, *and* $n_2 \neq bias$. *Again, it follows that there is a (not necessarily unique) E-path* $m^2_0, \ldots, m^2_{k_2}$ *of maximal length s.t.* $Cl(s_1 \cap s_2)(m^2_0) = 1$, $m^2_{k_2} = n_2$, *and there is no node* m *s.t.* $Cl(s_1 \cap s_2)(m) = 1$, m I $\langle m^2_i, m^2_{i+1}\rangle$ *for some* $i \in \{0, \ldots, k_2 - 1\}$. *Since* n_2 *is not necessarily a member of* $\bar{N}$, *we cannot simply infer again that* $m^2_0 = bias$. *But for the maximality of the selected path, we know at least that* $m^2_0 \in s_1 \cap s_2$, *and thus also* $s_1(m^2_0) = 1$.

   $Cl(s_1)(n_2) = 0$ *therefore implies that there has to be a node* n_3 *s.t.* $Cl(s_1)(n_3) = 1$, *and* n_3 I $\langle m^2_{i_2}, m^2_{i_2+1}\rangle$ *for some* $i_2 \in \{0, \ldots, k_2 - 1\}$. $n_3 \neq n_1$, $n_3 \neq n_2$, *since* $\mathcal{I}$ *is hierarchical. Suppose, for contradiction, that* $n_3 \in s_1$: *then* $n_3 \in \bar{N}$, $n_3 \neq bias$ *because* $Cl(s_1 \cap s_2)(n_3) = 0$, *and there is a path* $u_0, \ldots, u_k$ *in* $\mathcal{I}$ *with* $u_0 = n_3$ *(thus* $u_0 \in \bar{N}$*),* $u_1 = m^2_{i_2+1}, \ldots$, $u_{k_2 - i_2} = m^2_{k_2}$, $u_{k_2-i_2+1} = m^1_{i_1+1}, \ldots$, $u_k = m^1_{k_1}$ *(thus* $u_k \in \bar{N}$*), s.t.* u_0 I $\langle m^2_{i_2}, u_1\rangle$, $u_{k_2-i_2}$ I $\langle m^1_{i_1}, u_{k_2-i_2+1}\rangle$, *and between the rest of the nodes in the path, there are excitatory connections. But this contradicts* $\bar{N}$ *being odd in* $\mathcal{I}$. *Therefore,* $n_3 \notin s_1$.

3. *Now, we are in a similar situation, as we have been at the stage of proof item 1:* $n_3 \notin s_1$, $Cl(s_1)(n_3) = 1$, $Cl(s_1 \cap s_2)(n_3) = 0$. *There has to be a (not necessarily unique) E-path* $m^3_0, \ldots, m^3_{k_3}$ *of maximal length, s.t.* $Cl(s_1)(m^3_0) = 1$, $m^3_{k_3} = n_3$, *and there is no node* m *s.t.* $Cl(s_1)(m) = 1$, m I $\langle m^3_i, m^3_{i+1}\rangle$ *for some* $i \in \{0, \ldots, k_3 - 1\}$. *Since* $\bar{N}$ *is odd in* $\mathcal{I}$, m^3_0 *cannot be a member of* $\bar{N} \setminus \{bias\}$, *because otherwise we can find a path from* m^3_0 *to* n_1 *with an even number of inhibitions. Again, it follows that* $m^3_0 = bias$. *Extending the argument, analogously as above, it follows that there is an infinite sequence* $n_1, n_2, n_3, \ldots$ *of pairwise distinct nodes in* N, *contradicting the finiteness of* $\mathcal{I}$.

Therefore, $Cl(s_1 \cap s_2)(n_1) = 1$. ■

If Cl has the property that for every $\bar{N} \subseteq N$ s.t. $bias \in \bar{N}$, for all states $s_1, s_2 \subseteq \bar{N}$,

$$\bar{N} \cap Cl(s_1) \cap Cl(s_2) \subseteq \bar{N} \cap Cl(s_1 \cap s_2)$$

it will be called to satisfy *'($\bar{N}$-)distribution'* (corresponding to the distribution property of closure operators for sets of formulas; see Makinson[11], p.47).

Now we turn to antitonicity:

Definition 3.4. *(Antitone)*
Let $\mathcal{I} = \langle N, E, I, bias\rangle$ be an FHIN. Let $\bar{N} \subseteq N$ (later we will always assume that $bias \in \bar{N}$).
$\bar{N}$ is antitone in $\mathcal{I}$ iff
for all $n \in \bar{N}$ the mapping

$$\begin{array}{rccc} F_n: & \wp(\bar{N} \setminus \{n\}) & \rightarrow & \{0,1\} \\ & s & \mapsto & Cl(s)(n) \end{array}$$

is antitone, i.e., for all s_1, $s_2 \in \wp(\bar{N} \setminus \{n\})$: if $s_1 \subseteq s_2$ then $F_n(s_1) \geqslant F_n(s_2)$.
($\wp$ is the powerset operation.)

We can state some equivalent reformulations of def.3.4, and also one of its implications:

Corollary 3.5. *Let $\mathcal{I} = \langle N, E, I, bias\rangle$ be an FHIN. Let $\bar{N} \subseteq N$.*

1. *$\bar{N}$ is antitone in $\mathcal{I}$ iff*
   for all $n \in \bar{N}$, for all $s_2 \in \wp(\bar{N} \setminus \{n\})$:
   if $Cl(s_2)(n) = 1$ then for all s_1 s.t. $s_1 \subseteq s_2$, it holds that $Cl(s_1)(n) = 1$.
2. *If $\bar{N}$ is antitone in $\mathcal{I}$, for all $n \in \bar{N}$: if there is an $s \in \wp(\bar{N}\setminus\{n\})$ s.t. $Cl(s)(n) = 1$, then $Cl(\{bias\})(n) = 1$.*
3. *$\bar{N}$ is antitone in $\mathcal{I}$ iff*
   for all $X \subseteq \bar{N}$, for all s_1, $s_2 \in \wp(\bar{N} \setminus X)$:
   if $s_1 \subseteq s_2$ then $Cl(s_1) \cap X \supseteq Cl(s_2) \cap X$.
4. *In def.3.4, we could equivalently demand the mapping*

$$\begin{array}{rccc} F'_n: & \wp([\bar{N} \cap In(n)] \setminus \{n\}) & \rightarrow & \{0,1\} \\ & s & \mapsto & Cl(s)(n) \end{array}$$

   to be antitone, where $In(n)$ is the set of nodes, from which there are paths to n (i.e., which may have "causal influence" on n). This makes it easier to check whether a set $\bar{N}$ is antitone in $\mathcal{I}$.

Proof:

1. *straightforward;*
2. *this follows from claim 1, since $\{bias\}$ is a subset of every state;*
3. *"$\rightarrow$": assume that $\bar{N}$ is antitone in $\mathcal{I}$, let $X \subseteq \bar{N}$, s_1, $s_2 \in \wp(\bar{N} \setminus X)$, and $s_1 \subseteq s_2$. If $Cl(s_1) \cap X \not\supseteq Cl(s_2) \cap X$, then there is an $n \in Cl(s_2) \cap X$, s.t. $n \notin Cl(s_1) \cap X$. Since $n \in X$, it follows that $s_2 \in \wp(\bar{N} \setminus \{n\})$. But because*

$Cl(s_2)(n) = 1$, and since $\bar{N}$ is antitone in $\mathcal{I}$, we have that $Cl(s_1)(n) = 1$ (by claim 1 of this theorem), and therefore $n \in Cl(s_1) \cap X$, which is a contradiction.

"$\leftarrow$": assume the property stated in claim 3 on the right hand side of the equivalence sign, and suppose for contradiction that $\bar{N}$ is not antitone in $\mathcal{I}$. By 1, there is an $n \in \bar{N}$, an $s_2 \in \wp(\bar{N} \setminus \{n\})$, and an s_1 with $s_1 \subseteq s_2$, s.t. $Cl(s_2)(n) = 1$, and $Cl(s_1)(n) = 0$. Now we simply set $X := \{n\}$, and then we have: $\{n\} \subseteq \bar{N}$, $s_1, s_2 \in \wp(\bar{N} \setminus X)$, $s_1 \subseteq s_2$, but $Cl(s_1) \cap X = \emptyset \not\supseteq \{n\} = Cl(s_2) \cap X$, which is a contradiction.

4. *This follows from the fact that, for $s \in \wp(\bar{N} \setminus \{n\})$, $Cl(s)(n)$ does not depend on the values of s for nodes outside of $In(n)$.* ■

For antitone sets in FHINs one can also show a distribution property of closure states:

Lemma 3.6. *($\bar{N}$-Distribution II)*
For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$, for every $\bar{N} \subseteq N$: if $\bar{N}$ is antitone in $\mathcal{I}$, then for all states $s_1, s_2 \subseteq \bar{N}$:

$$\bar{N} \cap Cl(s_1) \cap Cl(s_2) \subseteq \bar{N} \cap Cl(s_1 \cap s_2).$$

Proof:

Suppose that $\bar{N} \subseteq N$ is antitone in $\mathcal{I}$.
Let $n \in \bar{N} \cap Cl(s_1) \cap Cl(s_2)$. We distinguish the following two cases:

1. *$n \in s_1$, $n \in s_2$. But then also $n \in s_1 \cap s_2 \subseteq Cl(s_1 \cap s_2)$, and thus $n \in \bar{N} \cap Cl(s_1 \cap s_2)$.*

2. *$n \notin s_1$, or $n \notin s_2$. Without restricting generality, assume that $n \notin s_2$. Now we have that $n \in \bar{N}$, $s_2 \in \wp(\bar{N} \setminus \{n\})$, $Cl(s_2)(n) = 1$, $s_1 \cap s_2 \subseteq s_2$, and $Cl(s_1)(n) = 1$. By claim 1 of corollary 3.5, it follows that $Cl(s_1 \cap s_2)(n) = 1$, and therefore $n \in \bar{N} \cap Cl(s_1 \cap s_2)$.* ■

Antitonicity does not only entail distribution, but it is also itself *entailed* by distribution:

Lemma 3.7. *($\bar{N}$-Distribution III)*
For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$, for every $\bar{N} \subseteq N$: if $\bar{N}$ is not antitone in $\mathcal{I}$, then not for all states $s_1, s_2 \subseteq \bar{N}$:

$$\bar{N} \cap Cl(s_1) \cap Cl(s_2) \subseteq \bar{N} \cap Cl(s_1 \cap s_2).$$

Proof:

Suppose that $\bar{N} \subseteq N$ is not antitone in $\mathcal{I}$.
Then there is an $n \in \bar{N}$, and there are s_1, $s_2 \in \wp(\bar{N} \setminus \{n\})$ with $s_1 \subseteq s_2$, s.t. $Cl(s_1)(n) \not\geq Cl(s_2)(n)$. Therefore, $Cl(s_1)(n) = 0$, $Cl(s_2)(n) = 1$. So we have that $n \in Cl(s_1 \cup \{n\})$, $n \in Cl(s_2)$, thus also $n \in Cl(s_1 \cup \{n\}) \cap Cl(s_2)$, but simultaneously $n \notin Cl([s_1 \cup \{n\}] \cap s_2) = Cl(s_1)$, which contradicts $\bar{N}$-distribution. ■

Lemmata 3.6 and 3.7 show that $\bar{N}$-distribution corresponds precisely to $\bar{N}$ being antitone in $\mathcal{I}$.

It follows:

Corollary 3.8. *For every FHIN $\mathcal{I} = \langle N, E, I, bias \rangle$, for every $\bar{N} \subseteq N$ s.t. $bias \in \bar{N}$: if $\bar{N}$ is odd in $\mathcal{I}$, then $\bar{N}$ is antitone in $\mathcal{I}$.*

Proof:

According to lemma 3.3, if $\bar{N}$ is odd in $\mathcal{I}$, $\mathcal{I}$ satisfies $\bar{N}$-distribution. But because of lemma 3.7, if $\mathcal{I}$ satisfies $\bar{N}$-distribution, $\bar{N}$ is antitone in $\mathcal{I}$. ■

Remark 3. *It is not difficult to show (by counterexamples) that the other direction, i.e., from being antitone to being odd, is not necessarily satisfied.*

4. Inhibition Nets as Cognizers

Inhibition nets will now be regarded as cognitive systems subserving cognitive agents which have beliefs and which draw inferences. In the following we will ascribe two kinds of beliefs to networks: (i) factual, occurrent, short-term beliefs, and (ii) normic, dispositional, long-term beliefs. Let us turn to the former first.

4.1. Ascribing Factual Beliefs to Inhibition Networks

The factual beliefs that we ascribe to inhibition nets are identified with occurrent, i.e., causally active, patterns of excitation. We may think of these beliefs as either being directly caused by the current state of the environment – in this case they are perceptual beliefs – or as being indirectly caused by such a perceptual belief via an intermediate inference process – in this case they are inferential beliefs. We use a propositional language $\mathcal{L}$ (the "factual" language) in order to ascribe such beliefs to inhibition nets. $\mathcal{L}$ shall consist of finitely many propositional variables and should be closed under the application of the standard logical connectives $(\neg, \wedge, \vee, \rightarrow, \leftrightarrow, \top, \bot)$ in the usual manner; we use small Greek letters with or without indices as metavariables ranging over the formulas of $\mathcal{L}$. E.g., we might use the propositional variables b and f in order to ascribe to a network the belief that the entity right in front of the network agent is a bird and is able to fly, i.e., the propositional formulas b and f are actually abbreviations for singular sentences of the form $Bird(a)$ and $CanFly(a)$. The corresponding singular beliefs are about current facts in the world, which usually change rather quickly.

Following the connectionist approaches to representation within artificial neural networks, we employ a distributed form of representation, i.e., it is not the nodes or the edges of an inhibition network that represent the contents of occurrent beliefs, but rather patterns of activation that are distributed over the whole ensemble of nodes. We define a net agent to believe φ in a state s iff a set $\mathfrak{I}(\varphi)$ of nodes (associated with φ) is active in s. $\mathfrak{I}$ (the "interpretation") assigns sets of nodes to formulas of $\mathcal{L}$; an inhibition net together with an interpretation mapping defines an interpreted inhibition net which might be thought of as being a part of an agent's cognitive

system. $\mathfrak{I}$ may be demanded to satisfy different kinds of constraints, and, as we will see below, the constraints in turn entail different kinds of logical properties satisfied by the inferences that are drawn by such networks. Let us first present the types of interpreted networks which we will focus on; we will discuss and motivate the corresponding formal constraints later on:

Definition 4.1. *(A Variety of Interpreted Inhibition Networks)*

1. *(Cumulative Interpreted Inhibition Networks)*

   A cumulative interpreted inhibition network $\mathfrak{N}$ *is a triple* $\langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$*, where*

   1. $\mathcal{I} = \langle N, E, I, bias \rangle$ *is a finite inhibition net, s.t.* $Cl(\{bias\}) \subsetneqq N$*,*
   2. $\mathcal{L}$ *is a language as characterized above,*
   3. $\mathfrak{I} : \mathcal{L} \to \wp(N)$ *satisfies:*
      1. $\mathfrak{I}(\top) = \{bias\}$ *(*$\top$ *is the logical verum),* $\mathfrak{I}(\bot) = N$ *(*$\bot$ *is the logical falsum),*
      2. *let* $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\, \mathfrak{I}(\varphi) = \{bias\}\}$*:*
         for all $\varphi, \psi \in \mathcal{L}$*: if* $\mathcal{TH}_{\mathfrak{I}} \vdash \varphi \to \psi$ *then* $\mathfrak{I}(\varphi) \supseteq \mathfrak{I}(\psi)$*,*
      3. *for all* $\varphi, \psi \in \mathcal{L}$*:* $\mathfrak{I}(\varphi \wedge \psi) = \mathfrak{I}(\varphi) \cup \mathfrak{I}(\psi)$*,*
      4. *for all* $\varphi \in \mathcal{L}$*:* $bias \in \mathfrak{I}(\varphi)$*,*
   4. *for all* $\varphi \in \mathcal{L}$*:* $Cl(\mathfrak{I}(\varphi))$ *is defined.*

2. *(Cumulative-Ordered Interpreted Inhibition Networks)*

   A cumulative-ordered interpreted inhibition network $\mathfrak{N}$ *is a triple* $\langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$ *defined analogously to 1, with the only difference that* $\mathcal{I} = \langle N, E, I, bias \rangle$ *is an FHIN.*

3. *(Preferential Partially Interpreted Antitone Inhibition Networks)*

   A preferential partially interpreted antitone inhibition network $\mathfrak{N}$ *is a quadruple* $\langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$ *defined analogously to 2, with the differences that*

   1. *there is a distinguished non-empty subset* $\bar{N}$ *of* N*, s.t.* $Cl(\{bias\}) \subsetneqq \bar{N}$*, and* $\bar{N}$ *is antitone in* $\mathcal{I}$*,*
   2. $\mathfrak{I} : \mathcal{L} \to \wp(\bar{N})$*, s.t.* $\mathfrak{I}$ *additionally satisfies:*
      1. *for all* $\varphi, \psi \in \mathcal{L}$*:* $\mathfrak{I}(\varphi \vee \psi) = \mathfrak{I}(\varphi) \cap \mathfrak{I}(\psi)$*,*
      2. *for all* $\varphi \in \mathcal{L}$*:* $\mathfrak{I}(\neg\varphi) = \bar{N} \setminus \mathfrak{I}(\varphi)$*.*

4. *(Preferential Partially Interpreted Odd Inhibition Networks)*

   Those are defined just as in 3 with the only exception that the term 'antitone' is replaced by 'odd'.

5. *(Simple Cumulative Interpreted Inhibition Networks)*

   A simple cumulative interpreted inhibition network $\mathfrak{N}$ *is a cumulative interpreted inhibition network* $\langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$*, where* $\mathcal{I}$ *has no inhibitory connections.*

6. *(Simple Preferential Partially Interpreted Inhibition Networks)*

   A simple preferential partially interpreted inhibition network $\mathfrak{N}$ is a preferential partially interpreted inhibition network $\langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$, where $\mathcal{I}$ has no inhibitory connections.

We have added the constraint that $Cl(\{bias\}) \subsetneqq N$ since otherwise for every parameter-setting s lemma 2.7 would entail that $Cl(s) = N$, since $\{bias\} \subseteq s \subseteq N = Cl(\{bias\})$. In this case the cognitive activity generated by an interpreted network would always trivially converge to a stable state identical to N. As we will see, for the class of net agents that we are going to define later, N is identical to the neural interpretation of the logical falsum, i.e. a net agent would always finally believe a contradiction, if we allowed for $Cl(\{bias\}) = N$.

Note that clause 1.4 above does not say that closure states have to exist for *all* states but just for the states which are represented in the object language, i.e., for all "cognitive" states of a network; usually, by far not every state of a network will be cognitive in this sense, any thus by far not every state of an interpreted network is demanded to have a closure state. In the case of cumulative-ordered/preferential/simple preferential interpreted inhibition networks clause 1.4 may be dropped because in each of the latter cases Cl is defined everywhere.

In the case of a (preferential/simple preferential) *partially* interpreted network, the nodes contained in $N \setminus \bar{N}$ may considered to be auxiliary "inter-neurons", without any representational function. The nodes contained in $\bar{N}$ might be called 'cognitive'.

Now we can state our informal presentation of the association of net states and belief states from above more precisely:

Definition 4.2. *(Ascription of Factual Beliefs)*
Let $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$ be an interpreted inhibition network of one of the types of def.4.1. Let s be a state in the state space S of $\mathcal{I}$:

$$Bel(\mathfrak{N}, s, \varphi) \text{ iff } \mathfrak{I}(\varphi) \subseteq s$$

(i.e.: $\mathfrak{N}$ believes in s the formula φ iff the pattern associated with φ is active in s).

We can also introduce the important concept of a *total* belief expressed by an *all-the-agent-believes* predicate (Levesque[9] has introduced the same notion as a sentential operator):

Definition 4.3. *(Ascription of Total Factual Beliefs)*
Let $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$ be an interpreted inhibition network of one of the types of def.4.1. Let s be a state in the state space S of $\mathcal{I}$:

$$AllBel(\mathfrak{N}, s, \varphi) \text{ iff } \mathfrak{I}(\varphi) = s$$

(i.e.: all that $\mathfrak{N}$ believes in s is the formula φ iff the pattern associated with φ is identical to the set of active nodes in s).

By def.4.3, if $AllBel(\mathfrak{N}, s, \varphi)$ then also $Bel(\mathfrak{N}, s, \varphi)$, and $\mathfrak{N}$ does not believe a formula in s that is "stronger" than φ, i.e., which has a larger associated pattern of activation.

The postulates for $\mathfrak{I}$ which characterize cumulative interpreted networks may be motivated by (i) rationality/justification considerations, and by (ii) considering the properties of distributed representations in general (see Rumelhart[13], chapter 3). Since, if an interpreted inhibition network is aimed to be both rational and connectionist:

- $\mathfrak{I}(\top) = \{bias\}$, $\mathfrak{I}(\bot) = N$:

  $\mathfrak{N}$ should believe that $\top$ is true in every possible parameter-setting, and thus, by def.4.2, $\mathfrak{I}(\top)$ has to be a pattern which is active in every possible parameter-setting; we choose $\{bias\}$ as such a pattern. The postulate for $\bot$ is a kind of "normalization" constraint – if one liked to drop it, one might simply replace 'N' in the considerations below by '$\mathfrak{I}(\bot)$'.

- For all φ, $\psi \in \mathcal{L}$: if $\mathcal{TH}_{\mathfrak{I}} \vdash \varphi \rightarrow \psi$ then $\mathfrak{I}(\varphi) \supseteq \mathfrak{I}(\psi)$ (where $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\}$):

  $\mathcal{TH}_{\mathfrak{I}}$ is the set of formulas φ, s.t. φ is believed by the net in every possible parameter-setting (since $\mathfrak{I}(\varphi) = \{bias\} \subseteq s$ for arbitrary s). If $\mathcal{TH}_{\mathfrak{I}} \vdash \varphi \rightarrow \psi$ the net should also believe that $\varphi \rightarrow \psi$ is true in every parameter-setting. Now suppose the net is in the parameter-setting $\mathfrak{I}(\varphi)$, i.e., all and only the nodes within $\mathfrak{I}(\varphi)$ fire; in this case the net also believes that φ is true by def.4.2 again. But then, by detachment, the net agent should also believe that ψ is true in this case, which entails, according to the way in which we have associated net states with belief states, that $\mathfrak{I}(\varphi)$ must be a superset of $\mathfrak{I}(\psi)$.

  We forgo to postulate also the direction from the right to the left, i.e. [if $\mathfrak{I}(\varphi) \supseteq \mathfrak{I}(\psi)$ then also $\mathcal{TH}_{\mathfrak{I}} \vdash \varphi \rightarrow \psi$] since this will have some technical advantages concerning the proof of the representation theorem in the next section.

- For all φ, $\psi \in \mathcal{L}$: $\mathfrak{I}(\varphi \wedge \psi) = \mathfrak{I}(\varphi) \cup \mathfrak{I}(\psi)$:

  at first it might look strange that the interpretation of a conjunction should be identical to the *union* of the component interpretations, since we are rather used to define it by some *intersection* of the component values.But the postulate matches intuitively the interpretation of neurons as "elementary-feature detectors": suppose there are just two neurons n_1 and n_2; n_1 fires iff a red object has been detected, whereas n_2 fires iff a large object has been detected. If now a both red *and* large object has been detected, this will be the case if and only if both n_1 *and* n_2 fire, i.e. the set of firing neurons will be identical to the *union* of $\{bias, n_1\}$ and $\{bias, n_2\}$ and not to their intersection (but still this postulate is not as unproblematic as it may seem: compare the discussion in Leitgeb[6], p.177),

- For all $\varphi \in \mathcal{L}$: $bias \in \mathfrak{I}(\varphi)$:

  bias fires in every state anyway.

Match the first two items in this list with the following quotation of Rumelhart et al.[13], p.84, on distributed representations: "... the relation between a type and an instance can be implemented by the relationship between a set of units and a larger set that includes it. Notice that the more general the type, the *smaller* the set of units used to encode it. As the number of terms in an *intensional* description gets smaller, the corresponding *extensional* set gets larger." Furthermore, compare the third item to the following quotation taken again from Rumelhart et al.[13], p.94: "A distributed representation uses a unit for a set of items, and it implicitly encodes a particular item as the intersection of the sets that correspond to the active units"; on p.95, such distributed representations are explicitly referred to as "conjunctive".

The postulates for cumulative interpreted networks thus seem to be quite natural and have the consequence that the set of factual beliefs of such networks are closed both under conjunction and modus ponens. On the other hand, our postulates for interpretation mappings with respect to disjunction and negation – i.e., as far as preferential interpreted networks are concerned – are not motivated as clearly, or so it seems; the main reason why we have introduced them is for the sake of the soundness and completeness results for the system P below. The same holds for the introduction of antitonicity and oddness. However, the constraints which we have imposed on preferential networks should not be mistaken for implying such obviously counter-intuitive postulates like: if $\mathfrak{N}$ rationally believes in s the formula $\varphi \vee \psi$, then she rationally believes in s the formula φ, or she rationally believes in s the formula ψ; if $\mathfrak{N}$ does not rationally believe in s the formula φ, then she rationally believes in s the formula $\neg\varphi$. The latter are indeed not entailed.

4.2. Ascribing Normic Beliefs (Nonmonotonic Inference Dispositions) to Interpreted Inhibition Networks

Now let us turn to the second class of belief states which we are going to ascribe to inhibition networks: the class of normic beliefs. The contents of such beliefs are not expressed by singular sentences about the current state of the world, but rather by normic laws of the form "normal φs are ψs", i.e., their contents are *general* laws, which do not change in time. We use a "conditional" language $\mathcal{L}_{\Rightarrow}$ in order to ascribe normic beliefs inhibition networks, where the members of $\mathcal{L}_{\Rightarrow}$ are conditionals $\varphi \Rightarrow \psi$, for φ, $\psi \in \mathcal{L}$. E.g., we might use the conditional $b \Rightarrow f$ to ascribe the normic belief "normally, a bird can fly" to a network, i.e., the conditional $b \Rightarrow f$ is actually an abbreviation for a general sentence of the form $Bird(x) \Rightarrow CanFly(x)$, where x may be considered bound by $\Rightarrow$. Such general beliefs states are identified with long-term dispositional states of a network, i.e., they manifest in the way the node activities change given certain circumstances. In our case, these circumstances are constituted by the current external input s^* and the current activity state s of the network, and the way the activity state of the network changes under s^* plus s, is determined by the topology of the network. Thus, whether a network has a certain normic belief or not depends on what the topology of the network looks like. If we are given an *interpreted* network, every trajectory of activity states corresponds to a trajectory of factual *beliefs*; every such trajectory might be interpreted as a nonmonotonic inference

which is drawn by the interpreted network; every such trajectory is determined, on the one hand, by a fixed input state, and, on the other hand, by a dispositional normic belief state. More precisely, we are going to ascribe to an interpreted net the normic belief expressed by $\varphi \Rightarrow \psi$ if and only if the interpreted network is disposed to draw a nonmonotonic inference from the total premise belief that φ is true, to the conclusion belief that ψ is true. A conclusion is reached when a stable closure state is obtained; the external input corresponds to the activation pattern that is associated with φ, and the stable closure state contains the activation pattern that is associated with ψ (recall definition 4.2 of factual belief ascription). We may think of a closure state as a plausible hypothesis generated by the agent in light of the evidence given by the input. In our example from above we would thus ascribe the normic belief expressed by $b \Rightarrow f$ to an interpreted inhibition net iff the interpreted network is disposed to draw a nonmonotonic inference from the total belief that there is a bird to the belief that there is something which is able to fly.

Since we regard the topology of inhibition networks as fixed, we also have to consider normic beliefs as being fixed and unalterable, although a natural extension of our approach would be to view such beliefs as being the results of a learning procedure in an artificial neural network. In order to do so we should rather develop our network semantics for artificial neural networks with weights, real-valued activation functions, etc., which is another topic (but some results in this direction are sketched in Leitgeb[6], section 7).

If we state our conception of normic beliefs formally, we get:

Definition 4.4. *(Ascription of Normic Beliefs/Nonmonotonic Inference Dispositions)*
Let $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$ be an interpreted network of one of the types of def.4.1:
for $\varphi \Rightarrow \psi \in \mathcal{L}_{\Rightarrow}$ we say that

$$\mathfrak{N} \models \varphi \Rightarrow \psi$$

($\varphi \Rightarrow \psi$ is true in $\mathfrak{N}$) iff for all $s \in S$: if $AllBel(\mathfrak{N}, s, \varphi)$ then $Bel(\mathfrak{N}, Cl(s), \psi)$.

Instead of saying that $\varphi \Rightarrow \psi$ is true in $\mathfrak{N}$, we might equivalently say that $\mathfrak{N}$ has the (normic) belief that normal φs are ψs, or that $\mathfrak{N}$ is disposed to draw the nonmonotonic inference from the total belief that φ is true to the belief that ψ is true. The potential nonmonotonicity of such inferences is due to the potential nonmonotonicity of Cl for an arbitrary inhibition network.

Our definiens has the following equivalent (re-)formulations:

Remark 4. *$\mathfrak{N} \models \varphi \Rightarrow \psi$ iff $Bel(\mathfrak{N}, Cl(\mathfrak{I}(\varphi)), \psi)$ iff $\mathfrak{I}(\psi) \subseteq Cl(\mathfrak{I}(\varphi))$.*

A clause similar to the last one of remark 4 has been used by Gärdenfors[4], p.63, in order to introduce nonmonotonic inferences to neural networks. The only minor difference is that Gärdenfors does not interpret object languages (like $\mathcal{L}$) by patterns, but instead he talks about the patterns in the metalanguage without making use of an object language at all.

Using def.4.4 we are able to associate theories of conditionals with interpreted networks:

Definition 4.5. *(Conditional Theories Corresponding to Interpreted Inhibition Nets)*
Let $\mathcal{TH}_{\Rightarrow}(\mathfrak{N}) = \{\varphi \Rightarrow \psi \,|\mathfrak{N} \models \varphi \Rightarrow \psi\}$.
$\mathcal{TH}_{\Rightarrow}(\mathfrak{N})$ *is the conditional theory corresponding to* $\mathfrak{N}$.

$\mathcal{TH}_{\Rightarrow}(\mathfrak{N})$ is the total description of the set of normic beliefs of $\mathfrak{N}$, and also of the set of nonmonotonic inferences $\mathfrak{N}$ is disposed to draw; our calling $\mathcal{TH}_{\Rightarrow}(\mathfrak{N})$ a conditional *theory* will be justified in the next section.

Example 4.6. *Let* $\mathcal{I}_1$ *be as defined above,* $\mathcal{L}_1$ *is built from the propositional variables* b *("bird"),* f *("flyer"),* w *("wings"),* p *("penguin"), and* $\mathfrak{I}_1(b) = \{bias, n_1\}$, $\mathfrak{I}_1(f) = \{bias, n_1, n_2\}$, $\mathfrak{I}_1(w) = \{bias, n_1, n_3\}$, $\mathfrak{I}_1(p) = \{bias, n_1, n_4\}$, $\mathfrak{I}_1(\neg\varphi) := \{bias\} \cup N_1 \setminus \mathfrak{I}_1(\varphi)$, $\mathfrak{I}_1(\varphi \wedge \psi) := \mathfrak{I}_1(\varphi) \cup \mathfrak{I}_1(\psi)$ *for all* $\varphi, \psi \in \mathcal{L}_1$. *It is easy to see that* $\mathfrak{N}_1 = \langle \mathcal{I}_1, \mathcal{L}_1, \mathfrak{I}_1 \rangle$ *is an interpreted inhibition network. Since* $\mathcal{I}_1$ *is an FHIN,* $\mathfrak{N}_1$ *is cumulative-ordered (and even preferential with* $\bar{N} = N$*). The definitions of Bel, AllBel, and of satisfaction for conditionals entail that*
$\mathfrak{N}_1 \models \{b \Rightarrow f \wedge w, b \wedge f \Rightarrow w, b \wedge p \Rightarrow \neg f \wedge w, b \Rightarrow \neg p, b \vee p \Rightarrow f, \dots\}$,
$\mathfrak{N}_1 \not\models \{b \Rightarrow p, p \Rightarrow f, f \Rightarrow p, \top \Rightarrow b, \top \Rightarrow f \wedge w, p \Rightarrow \neg p, w \Rightarrow p, b \wedge p \Rightarrow f, \dots\}$.
Therefore, e.g., if all that $\mathfrak{N}_1$ *believes is that there is a bird, then she infers that there is something which is able to fly and which has got wings. But if all that* $\mathfrak{N}_1$ *believes is that there is a penguin bird, then she infers that there is something which is not able to fly but still has got wings.*

Example 4.7. *Let* $\mathcal{I}_2$ *be as defined above,* $\mathcal{L}_2$ *is built from the propositional variables* p *and* q*, and:* $n_1 \in \mathfrak{I}_2(\varphi)$ *iff* $p \wedge \neg q \not\models \varphi$ *or* $\neg p \wedge q \not\models \varphi$, $n_2 \in \mathfrak{I}_2(\varphi)$ *iff* $p \wedge q \not\models \varphi$ *or* $p \wedge \neg q \not\models \varphi$, $n_3 \in \mathfrak{I}_2(\varphi)$ *iff* $p \wedge q \not\models \varphi$ *or* $\neg p \wedge q \not\models \varphi$*, and always* $bias \in \mathfrak{I}_2(\varphi)$ *(for* $\varphi \in \mathcal{L}_2$*). It follows that, e.g.,* $\mathfrak{I}_2(p) = \{bias, n_1, n_3\}$, $\mathfrak{I}_2(q) = \{bias, n_1, n_2\}$, $\mathfrak{I}_2(\neg p) = \{bias, n_1, n_2, n_3\}$, $\mathfrak{I}_2(p \vee q) = \{bias\}$, $\mathfrak{I}_2(\neg(p \leftrightarrow q)) = \{bias, n_2, n_3\}$. *One can show that* $\mathfrak{N}_2 = \langle \mathcal{I}_2, \mathcal{L}_2, \mathfrak{I}_2 \rangle$ *is a cumulative interpreted inhibition network. The definitions of Bel, AllBel, and of satisfaction for conditionals entail that, e.g.,* $\mathfrak{N}_2 \models \top \Rightarrow \neg(p \leftrightarrow q)$*, but* $\mathfrak{N}_2 \not\models q \Rightarrow \neg(p \leftrightarrow q)$.

General beliefs are thus not represented in the network by patterns of activity but by the topology of the network. Such a way of coding is again a distributed kind of representation, since it is not a single node or a single connection which represents a normic belief but rather the whole network.

The role of interpreted inhibition networks within a possible architecture for "low-level" cognitive agents is treated extensively in Leitgeb[7].

5. The Representation Theorems

Interpreted inhibition nets may be shown to have nice logical properties. We can prove (i) soundness results: the sets $\mathcal{TH}_{\Rightarrow}(\mathfrak{N})$ of conditionals corresponding to interpreted inhibition networks of a certain class of nets are closed under the rules of well-known systems of nonmonotonic reasoning, (ii) completeness results: for every set $\mathcal{TH}_{\Rightarrow}$ of conditionals closed under the rules of such a well-known system of nonmonotonic reasoning, there is an interpreted inhibition net $\mathfrak{N}$ of a certain class of nets, s.t.,

$\mathcal{TH}_{\Rightarrow}(\mathfrak{N}) = \mathcal{TH}_{\Rightarrow}$. If we take soundness and completeness together we can formulate and prove corresponding representation theorems.

In order to state this more clearly, we need the notion of a conditional theory extending a deductive closed theory $\mathcal{TH}$ of factual formulas in $\mathcal{L}$, where the conditional theory conforms to the rules of the systems C, CL, P, CM, or M studied by KLM[5]. Note that where KLM refer to nonmonotonic inference *relations* $\mid\sim$ (see also Makinson[11]), we rather refer to *sets of conditionals* in the spirit of conditional logic:

Definition 5.1. *(Conditional Theories)*
Let $\mathcal{TH} \subseteq \mathcal{L}$ be a deductively closed theory:

1. *A conditional C-theory extending $\mathcal{TH}$ is a set $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ with the property that for all $\alpha \in \mathcal{L}$ it holds that $\alpha \Rightarrow \alpha \in \mathcal{TH}_{\Rightarrow}$ (Reflexivity),*

   and which is closed under the following rules:

   1. $\dfrac{\mathcal{TH} \vdash \alpha \leftrightarrow \beta, \alpha \Rightarrow \gamma}{\beta \Rightarrow \gamma}$ *(Left Equivalence)*
   2. $\dfrac{\gamma \Rightarrow \alpha, \mathcal{TH} \vdash \alpha \rightarrow \beta}{\gamma \Rightarrow \beta}$ *(Right Weakening)*
   3. $\dfrac{\alpha \wedge \beta \Rightarrow \gamma, \alpha \Rightarrow \beta}{\alpha \Rightarrow \gamma}$ *(Cautious Cut)*
   4. $\dfrac{\alpha \Rightarrow \beta, \alpha \Rightarrow \gamma}{\alpha \wedge \beta \Rightarrow \gamma}$ *(Cautious Monotonicity)*

   We refer to the axiom scheme and the rules above as the system C (see [5], pp.176-180). The rules are to be read in the following way:

   e.g., by Cut, if $\alpha \wedge \beta \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$ and $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$, then $\alpha \Rightarrow \gamma \in \mathcal{TH}_{\Rightarrow}$.

2. *A conditional C-theory $\mathcal{TH}_{\Rightarrow}$ is consistent iff $\top \Rightarrow \bot \notin \mathcal{TH}_{\Rightarrow}$.*

3. *A conditional CL-theory $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}$ is a conditional C-theory extending $\mathcal{TH}$, which is closed under the following rule:*

   $\dfrac{\alpha_0 \Rightarrow \alpha_1, \alpha_1 \Rightarrow \alpha_2, \ldots, \alpha_{j-1} \Rightarrow \alpha_j, \alpha_j \Rightarrow \alpha_0}{\alpha_r \Rightarrow \alpha_{r'}}$ *(Loop)*

   (r, r' are arbitrary members of $\{0, \ldots, j\}$)

   We refer to C+Loop as the system CL (see [5], pp.187).

4. *A conditional P-theory $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}$ is a conditional CL-theory extending $\mathcal{TH}$, which is closed under the following rule:*

   $\dfrac{\alpha \Rightarrow \gamma, \beta \Rightarrow \gamma}{\alpha \vee \beta \Rightarrow \gamma}$ *(Or)*

   We refer to CL+Or as the system P (see [5], pp.189-190).

5. *A conditional CM-theory $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}$ is a conditional C-theory extending $\mathcal{TH}$, which is closed under the following rule:*

$$\frac{\mathcal{TH} \vdash \alpha \rightarrow \beta, \beta \Rightarrow \gamma}{\alpha \Rightarrow \gamma} \textit{ (Monotonicity)}$$

We refer to C+Monotonicity as the system CM (see [5], pp.200-201).

6. *A conditional M-theory $\mathcal{TH}_{\Rightarrow}$ extending $\mathcal{TH}$ is a conditional C-theory extending $\mathcal{TH}$, which is closed under the following rule:*

$$\frac{\alpha \Rightarrow \beta}{\neg\beta \Rightarrow \neg\alpha} \textit{ (Contraposition)}$$

We refer to C+Contraposition as the system M (see [5], p.202).

(In each case for arbitrary α, β, γ, α_0, α_1,..., $\alpha_j \in \mathcal{L}$).

Now we can show the following representation results:

Theorem 5.2. *(Representation)*
Let $\mathcal{TH} \subseteq \mathcal{L}$ be a theory:

1. *$\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional C-theory extending $\mathcal{TH}$ iff there is a cumulative interpreted inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$.*

2. *$\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional CL-theory extending $\mathcal{TH}$ iff there is a cumulative-ordered interpreted inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$.*

3. *$\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional P-theory extending $\mathcal{TH}$ iff there is a preferential partially interpreted antitone inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$.*

4. *$\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional CM-theory extending $\mathcal{TH}$ iff there is a simple cumulative interpreted inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$.*

5. *$\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional M-theory extending $\mathcal{TH}$ iff there is a simple preferential partially interpreted inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$.*

Claim 3 of theorem 5.2 together with corollary 3.8 implies the following soundness theorem for P with respect to preferential partially interpreted *odd* networks:

Corollary 5.3. *If there is a preferential partially interpreted odd inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$, s.t. $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, and $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$, then $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ is a consistent conditional P-theory extending $\mathcal{TH}$.*

We will not prove every single claim contained in theorem 5.2, but we will rather concentrate on the more central ones, i.e., on completeness for C and for P. The proofs of soundness for both systems will only be sketched, since they are very similar to the soundness proof for CL in Leitgeb[6] (recall that claim 2 has been proved in Leitgeb[6]). Claims 4 and 5 follow easily from the other claims and are relatively trivial.

5.1. Proving Representation for C

Soundness, i.e., the right-to-left direction of 1 in theorem 5.2, is proved in nearly the same way as it has been proved for CL in [6] – use remark 1, lemma 2.7, item 1 of definition 4.1, definitions 4.2, 4.3, 4.4, and remark 4.

So let us turn to completeness:

Lemma 5.4. *(Completeness for C)*

Let $\mathcal{TH} \subseteq \mathcal{L}$ be a theory:

for every consistent conditional C-theory $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ extending $\mathcal{TH}$ there is a cumulative interpreted inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$, s.t.

- $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\, \mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, *and*
- $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$, *i.e. for every $\alpha \Rightarrow \beta \in \mathcal{L}_{\Rightarrow}$:*

$$\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow} \text{ iff } \mathfrak{N} \models \alpha \Rightarrow \beta.$$

Proof:

First we construct a network analogously as in the proof of lemma 5.6 of Leitgeb[6] (pp.186f), i.e.:

by theorem 3.25 by KLM[5], which is proved on pp.184-185, for every $\mathcal{TH}_{\Rightarrow}$ as above there is a finite cumulative model $\mathfrak{M}_c = \langle \bar{S}, l, \prec \rangle$ (based on the set of worlds satisfying $\mathcal{TH}$), s.t. $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$ iff $\mathfrak{M}_c \models \alpha \Rightarrow \beta$, i.e. all states minimal with respect to $\prec$, which make α true, also make β true. We use $\mathfrak{M}_c$ to construct the intended input-determined interpreted network $\mathfrak{N}$. We take '$\bar{s}$' with or without index to range over states in the sense of preferential models, and 's', as usual, to range over net states.

Let $N = \{bias\} \cup \bar{S}$. Let $E = \{\langle bias, \bar{s} \rangle \,|\, \bar{s}$ is not minimal according to $\prec\} \cup \{\langle \bar{s}, \bar{s}' \rangle \,|\, \bar{s} \prec \bar{s}'\}$. For every $\bar{s} \in \bar{S}$ let $L_{\bar{s}} = \{\bar{s}' \in \bar{S} \,|\, \bar{s}' \prec \bar{s}\}$; say, $L_{\bar{s}} = \{\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}\}$. Now we define $I_{\bar{s}} = \{\langle bias, \langle \bar{s}_1, \bar{s} \rangle \rangle, \langle \bar{s}_1, \langle \bar{s}_2, \bar{s} \rangle \rangle, \ldots, \langle \bar{s}_{r_{\bar{s}}-1}, \langle \bar{s}_{r_{\bar{s}}}, \bar{s} \rangle \rangle, \langle \bar{s}_{r_{\bar{s}}}, \langle bias, \bar{s} \rangle \rangle\}$. If $\bar{s}$ is minimal in $\prec$, then let $I_{\bar{s}} = \emptyset$. Let $I = \bigcup_{\bar{s} \in \bar{S}} I_{\bar{s}}$. Obviously, $I \subseteq N \times E$. Since $\mathcal{I}$ is finite (because $\bar{S}$ is), $\mathcal{I}$ is a finite inhibition net.

We define for $\varphi \in \mathcal{L}$: $\mathfrak{I}(\varphi) = \{bias\} \cup \{\bar{s} \,|\, \bar{s}$ does not make φ true$\}$. It is easy to see that $\mathfrak{I}$ is an interpretation mapping as demanded by 1 of definition 4.1. In order to show that $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I} \rangle$ is a cumulative interpreted inhibition network, it remains to prove that $Cl(\{bias\}) \subsetneqq N$ and that there is a closure for all representational states. In fact we will now show more: there is a closure $Cl(\mathfrak{I}(\alpha))$ for all $\alpha \in \mathcal{L}$, and $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ iff $\bar{s}$ is a minimal α-state according to $\mathfrak{M}_c$. From that it follows

that $\mathfrak{M}_c \models \alpha \Rightarrow \beta$ *iff* $\mathfrak{N} \models \alpha \Rightarrow \beta$*, which entails the intended completeness claim:* $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$ *iff* $\mathfrak{N} \models \alpha \Rightarrow \beta$.

The proof method we apply is not identical to the one used in the proof of lemma 5.6 of [6], but it might also be applied there. Let $\alpha \in \mathcal{L}$*, let* s *be a parameter-setting of* $\mathcal{I}$.

For all $\bar{s} \in \mathfrak{I}(\alpha)$ *we have that* $F_{\mathfrak{I}(\alpha)}(s)(\bar{s}) = 1$ *and also that* $\bar{s}$ *does not make* α *true in* $\mathfrak{M}_c$ *and is thus no minimal* α*-state according to* $\mathfrak{M}_c$*. The same holds for* $F^i_{\mathfrak{I}(\alpha)}(s)$ *where* $i \geqslant 1$.

So we can concentrate on the case where $\bar{s} \notin \mathfrak{I}(\alpha)$*:*

by the def. of F_{s^*} *we know that* $F_{\mathfrak{I}(\alpha)}(s)(\bar{s}) = 0$ *iff*

$\mathfrak{I}(\alpha)(n) = 0$*, and* $\neg\exists n_1 \in N\,(s(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(s(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$ *iff, since* $\bar{s} \notin \mathfrak{I}(\alpha)$ *by assumption,*

$\neg\exists n_1 \in N\,(s(n_1) = 1, n_1\ E\ n, \neg\exists n_2 \in N\,(s(n_2) = 1, n_2\ I\ \langle n_1, n\rangle))$.

Now we distinguish between two subcases:

1. $\bar{s}$ *is a minimal* α*-state according to* $\mathfrak{M}_c$*:*

   thus all states in $\mathfrak{M}_c$ *which are below* $\bar{s}$ *(if there are any) are no* α*-states. By the def. of* $\mathfrak{I}$ *it follows that* $\mathfrak{I}(\alpha) \supseteq L_{\bar{s}} = \left\{\bar{s}' \in \bar{S} \,|\, \bar{s}' \prec \bar{s}\right\}$*, and so for all* $\bar{s}' \in L_{\bar{s}}$*,* $i \geqslant 1$ *we have* $F^i_{\mathfrak{I}(\alpha)}(s)(\bar{s}') = 1$ *from the above. But then every excitatory connection to* $\bar{s}$ *is inhibited in* $F^i_{\mathfrak{I}(\alpha)}(s)$ *for all* $i \geqslant 2$ *by the def. of* I*. Since* $\bar{s} \notin \mathfrak{I}(\alpha)$ *this implies:* $F^i_{\mathfrak{I}(\alpha)}(s)(\bar{s}) = 0$ *for all* $i \geqslant 2$.

2. $\bar{s}$ *is no minimal* α*-state according to* $\mathfrak{M}_c$*:*

   in this case there is a $\bar{s}_j \in L_{\bar{s}}$ *s.t.* $\bar{s}_j$ *is a minimal* α*-state according to* $\mathfrak{M}_c$ *(this is by the smoothness of* $\mathfrak{M}_c$*). But we have just shown that* $F^i_{\mathfrak{I}(\alpha)}(s)(\bar{s}_j) = 0$ *for all* $i \geqslant 2$*. So there is an uninhibited excitatory connection to* $\bar{s}$ *for all* $i \geqslant 2$ *by the def. of* E*, and we have:* $F^i_{\mathfrak{I}(\alpha)}(s)(\bar{s}) = 1$ *for all* $i \geqslant 2$.

Summing up we have proved that for all $\bar{s} \in \bar{S}$*:* $F^i_{\mathfrak{I}(\alpha)}(s)(\bar{s}) = 0$ *(for all* $i \geqslant 2$*) iff* $\bar{s}$ *is a minimal* α*-state according to* $\mathfrak{M}_c$*. But since this holds for all net parameter-settings* $s \in S$ *we have: there is a closure* $Cl(\mathfrak{I}(\alpha))$ *for all* $\alpha \in \mathcal{L}$*, and* $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ *iff* $\bar{s}$ *is a minimal* α*-state according to* $\mathfrak{M}_c$*. So we are done.* ■

5.2. Proving Representation for P

Soundness for P follows from soundness for C and the following lemma:

Lemma 5.5. *Let* $\mathfrak{N} = \left\langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \right\rangle$ *be a preferential partially interpreted antitone (or odd) inhibition network; let* α*,* β*,* $\gamma \in \mathcal{L}$*: then* $\mathfrak{N}$ *satisfies*

Or: if $\mathfrak{N} \models \alpha \Rightarrow \gamma$*,* $\mathfrak{N} \models \beta \Rightarrow \gamma$*, then* $\mathfrak{N} \models \alpha \vee \beta \Rightarrow \gamma$.

Proof:

By assumptions and remark 4, $\mathfrak{I}(\gamma) \subseteq Cl(\mathfrak{I}(\alpha))$ *and* $\mathfrak{I}(\gamma) \subseteq Cl(\mathfrak{I}(\beta))$*; thus* $\mathfrak{I}(\gamma) \subseteq Cl(\mathfrak{I}(\alpha)) \cap Cl(\mathfrak{I}(\beta))$*, and since*
$Cl(\mathfrak{I}(\alpha)) \cap Cl(\mathfrak{I}(\beta)) \subseteq Cl(\mathfrak{I}(\alpha) \cap \mathfrak{I}(\beta))$ *by lemma 3.6 (or 3.3), and* $Cl(\mathfrak{I}(\alpha) \cap \mathfrak{I}(\beta)) = Cl(\mathfrak{I}(\alpha \vee \beta))$ *by 3 of definition 4.1, we are finished again by remark 4.* ■

Finally, we prove completeness for P:

Lemma 5.6. *(Completeness for P)*
Let $\mathcal{TH} \subseteq \mathcal{L}$ be a theory:
for every consistent conditional P-theory $\mathcal{TH}_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ extending $\mathcal{TH}$ there is a preferential partially interpreted antitone inhibition network $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$, s.t.

- $\mathcal{TH}_{\mathfrak{I}} = \{\varphi \in \mathcal{L} \,|\mathfrak{I}(\varphi) = \{bias\}\} \supseteq \mathcal{TH}$, *and*
- $\mathcal{TH}_{\Rightarrow} = \mathcal{TH}_{\Rightarrow}(\mathfrak{N})$, *i.e. for every $\alpha \Rightarrow \beta \in \mathcal{L}_{\Rightarrow}$:*

$$\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow} \text{ iff } \mathfrak{N} \models \alpha \Rightarrow \beta.$$

Proof:

By theorem 5.18 of KLM[5], which is proved on pp.193-196, for every $\mathcal{TH}_{\Rightarrow}$ as above there is a finite preferential model $\mathfrak{M}_p = \langle \bar{S}, l, \prec \rangle$ (based on the set of worlds satisfying $\mathcal{TH}$), s.t. $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$ iff $\mathfrak{M}_p \models \alpha \Rightarrow \beta$, i.e. all states minimal with respect to $\prec$, which make α true, also make β true. We use $\mathfrak{M}_p$ again to construct the intended partially interpreted network $\mathfrak{N}$.

For every $\bar{s} \in \bar{S}$ let $L_{\bar{s}} = \{\bar{s}' \in \bar{S} \,|\bar{s}' \prec \bar{s}\}$; say, $L_{\bar{s}} = \{\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}\}$. Furthermore, let $\bigwedge L_{\bar{s}}$ be a conjunction subnetwork for the nodes $\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}$; i.e., there is a node $\bar{s}_1 \wedge \ldots \wedge \bar{s}_{r_{\bar{s}}}$, the activity of which matches the Boolean conjunction of the activity states of $\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}$ (that one can construct such a network by means of an inhibition net is proved in Leitgeb[6], theorem 3.8, on p.173). If $\bar{s}$ is minimal in $\prec$, then let $\bigwedge L_{\bar{s}}$ be the empty subnet. Now we define:

let $N = \{bias\}$ joined with the set of nodes of the conjunction subnetwork $\bigwedge L_{\bar{s}}$ for each $\bar{s} \in \bar{S}$.

Let $E = \{\langle bias, \bar{s} \rangle \,|\bar{s}$ is not minimal according to $\prec\}$, joined with the set of excitatory connections within the conjunction subnetworks $\bigwedge L_{\bar{s}}$ for each $\bar{s} \in \bar{S}$.

Moreover, for non-minimal $\bar{s} \in \bar{S}$, and $L_{\bar{s}} = \{\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}\}$,
let $I_{\bar{s}} = \{\langle \bar{s}_1 \wedge \ldots \wedge \bar{s}_{r_{\bar{s}}}, \langle bias, \bar{s} \rangle \rangle\}$. If $\bar{s}$ is minimal in $\prec$, then let $I_{\bar{s}} = \emptyset$. Let $I = \bigcup_{\bar{s} \in \bar{S}} I_{\bar{s}}$ joined with the inhibitory connections within each conjunction subnetwork. Obviously, $I \subseteq N \times E$.

Let $\mathcal{I} = \langle N, E, I, bias \rangle$. Since $\mathcal{I}$ does not contain any cycles, and since $\mathcal{I}$ is finite (because $\bar{S}$ is), $\mathcal{I}$ is an FHIN.

Let $\bar{N} = \{bias\} \cup \bar{S}$. As a consequence of our def. of $\mathcal{I}$, our def. of $\bar{N}$, we see that $\bar{N}$ is antitone in $\mathcal{I}$. This follows because for all $\bar{s} \in \bar{N}$, for all $s_2 \in \wp(\bar{N} \setminus \{\bar{s}\})$: if $Cl(s_2)(\bar{s}) = 1$ then $\bar{s}$ has to be excited by the bias node via $\langle bias, \bar{s} \rangle$; but this is only possible if $Cl(s_2)(\bar{s}_1 \wedge \ldots \wedge \bar{s}_{r_{\bar{s}}}) = 0$ (for $L_{\bar{s}} = \{\bar{s}_1, \ldots, \bar{s}_{r_{\bar{s}}}\}$). Thus there has to be a node within $L_{\bar{s}}$, which is not active in $Cl(s_2)$. Now, for all s_1 s.t. $s_1 \subseteq s_2$, it also holds that there is a node within $L_{\bar{s}}$ which is not active in $Cl(s_1)$, because if every node in $L_{\bar{s}}$ were active in $Cl(s_1)$, there would be a node $\bar{s}^* \in L_{\bar{s}}$, s.t. $\bar{s}^* \notin Cl(s_2)$, but $\bar{s}^* \in Cl(s_1)$. $\bar{s}^*$ cannot be a member of layer 0, because $s_1 \subseteq s_2$. Thus $\bar{s}^*$ has to be excited in $Cl(s_1)$ by the bias node via $\langle bias, \bar{s}^* \rangle$; but this is only possible if $Cl(s_1)(\bar{s}^*_1 \wedge \ldots \wedge \bar{s}^*_{r_{\bar{s}^*}}) = 0$ (where $L_{\bar{s}^*} = \{\bar{s}^*_1, \ldots, \bar{s}^*_{r_{\bar{s}^*}}\}$). Therefore, there has to be a node $\bar{s}^*_i$ within $L_{\bar{s}^*}$ which is not active in $Cl(s_1)$. But $\bar{s}^*_i$ is also a member of

$L_{\bar{s}}$, which contradicts our assumption that every node in $L_{\bar{s}}$ is active in $Cl(s_1)$. So we have that there is also a node within $L_{\bar{s}}$ which is not active in $Cl(s_1)$. Thus, $Cl(s_1)(\bar{s}_1 \wedge \ldots \wedge \bar{s}_{r_{\bar{s}}}) = 0$, therefore $Cl(s_1)(\bar{s}) = 1$. So, $\bar{N}$ is antitone in $\mathcal{I}$.

Now we define for $\varphi \in \mathcal{L}$: $\mathfrak{I}(\varphi) = \{bias\} \cup \{\bar{s} \,|\, \bar{s}$ does not make φ true$\}$. $\mathfrak{I}$ is easy to be shown an interpretation mapping satisfying the postulates of 3 in definition4.1, and thus $\mathfrak{N} = \langle \mathcal{I}, \mathcal{L}, \mathfrak{I}, \bar{N} \rangle$ is a preferential partially interpreted and antitone inhibition network.

Now we show that $\mathfrak{M}_p \models \alpha \Rightarrow \beta$ iff $\mathfrak{N} \models \alpha \Rightarrow \beta$, which again entails: $\alpha \Rightarrow \beta \in \mathcal{TH}_{\Rightarrow}$ iff $\mathfrak{N} \models \alpha \Rightarrow \beta$.

Let $\alpha \in \mathcal{L}$. We will prove by induction that $\bar{s} \in \bar{S}$ does not fire in the net state $Cl(\mathfrak{I}(\alpha))$ iff $\bar{s}$ is a minimal α-state according to $\mathfrak{M}_p$. Let $\langle N_0, \ldots, N_k \rangle$ be the canonical partition of $\mathcal{I}$ as it has been defined on p.167 of Leitgeb[6] (there we have shown that every FHIN can be decomposed into layers, s.t., all connections starting from one layer only lead to nodes of layers with higher index):

- *Induction basis:*

  let $\bar{s} \in N_0$ ($\bar{s} \neq bias$ since $\bar{s} \in \bar{S}$);

  $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ iff $\bar{s} \notin \mathfrak{I}(\alpha)$ iff $\bar{s}$ is an α-state. Moreover, every state $\bar{s}$ in N_0 is minimal according to $\prec$ by def. of E and I.

- *Induction step:*

  assume that for every $\bar{s} \in N_0 \cup \ldots \cup N_i$: $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ iff $\bar{s}$ is a minimal α-state. Now consider an arbitrary $\bar{s} \in N_{i+1}$:

  $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ iff $\bar{s} \notin \mathfrak{I}(\alpha)$ and $\neg\exists m \in N_j$ with $j < i$ s.t.
  $(Cl(\mathfrak{I}(\alpha))(m) = 1, m\ E\ \bar{s}, \neg\exists m' \in N_u$ with $u < i\ (Cl(\mathfrak{I}(\alpha))(m') = 1, m'\ I\ \langle m, \bar{s} \rangle))$.
  But this is the case if and only if $\bar{s}$ is a minimal α-state, for the following reasons:

  first, $\bar{s} \notin \mathfrak{I}(\alpha)$ iff $\bar{s}$ is an α-state; at second, by def. of E and I, $\neg\exists m \in N_j$ s.t.

  $j < i$ and $Cl(\mathfrak{I}(\alpha))(m) = 1, m\ E\ \bar{s}, \neg\exists m' \in N_u$ with $u < i$ s.t.

  $(Cl(\mathfrak{I}(\alpha))(m') = 1, m'\ I\ \langle m, \bar{s} \rangle)$ iff

  $\forall \bar{s}' \in L_{\bar{s}}$ it holds that $Cl(\mathfrak{I}(\alpha))(\bar{s}') = 1$ iff

  (by induction hypothesis) $\forall \bar{s}' \in L_{\bar{s}}$ it holds that $\bar{s}'$ is no minimal α-state. But $\bar{s}$ is an α-state and $\forall \bar{s}' \in L_{\bar{s}}$ ($\bar{s}'$ is no minimal α-state) iff $\bar{s}$ is a minimal α-state (by the Smoothness Condition). Therefore, we have that $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ iff $\bar{s}$ is a minimal α-state.

We know that $\mathfrak{M}_p \models \alpha \Rightarrow \beta$ iff all minimal α-states are β-states. But the latter is the case, if and only if for all $\bar{s} \in \bar{S}$: if $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 0$ then $\bar{s} \notin \mathfrak{I}(\beta)$, or equivalently, for all $\bar{s} \in \bar{S}$: if $\bar{s} \in \mathfrak{I}(\beta)$ then $Cl(\mathfrak{I}(\alpha))(\bar{s}) = 1$. So, $\mathfrak{M}_p \models \alpha \Rightarrow \beta$ iff $\mathfrak{N} \models \alpha \Rightarrow \beta$. ■

Remark 5. *Note that the class of networks, which are constructed in the course of the proof of lemma, is actually a proper subclass of the class of all antitone networks (as may be shown by constructing counterexamples).*

We have seen that the systems of KLM are not just sound and complete with respect to the normal states semantics employed by KLM, and, in the case of P, to a probability semantics (see Pearl[12]) or a possibilistic semantics (see Dubois and Prade[3]), but also with respect to the network semantics that we have developed above for interpreted inhibition networks and which is based on the concept of distributed representation in networks. Interpreted inhibition networks may be viewed as cognizers which draw nonmonotonic inferences in correspondence to correct systems of nonmonotonic reasoning. Moreover, for every conditional theory conforming to the rules of one of the systems above, there is an interpreted inhibition networks which reasons precisely according to the given theory.

One could also study further semantical notions like validity, or logical implication, for which corresponding soundness and completeness results can be shown, but we omit this for the sake of brevity (but see Leitgeb[6] where we have done this for the system CL, and Leitgeb[7] where we have also done this for the other systems).

6. A Preliminary Comparison: Inhibition Nets and Logic Programs

Now we turn to the connections and/or differences between inhibition nets and logic programs. We will use the terminology of Lifschitz[10]: apart from the usual definitions of (basic, normal, hierarchical) programs, closure under a program, the consequence operator Cn, the reduct Π^X of program Π relative to sets X of literals, answer sets, etc., we use the immediate consequence operator T_Π, s.t., for a basic program Π and a consistent X, $T_\Pi(X) = \{Head \,|\, Head \leftarrow Body \in \Pi, Body \subseteq X\}$. It is easy to see that $Cn(\Pi)$ is the least fixed point of T_Π, and $Cn(\Pi)$ may be computed "bottom up" by $Cn(\Pi) = \bigcup_{n \geqslant 0} T_\Pi^n(\varnothing)$. For the historical origin of these definitions see [10], section 6.

We also need the notion of a *direct* consequence operator S_Π for programs with negation as failure, and we introduce a new consequence operator Cn^+:

Definition 6.1. *Let Π be a finite normal program.*

For every set X of literals, let $S_\Pi(X) = T_{\Pi^X}(X)$, i.e.,
$S_\Pi(X) = \{Head \,|\, Head \leftarrow Pos \cup not(Neg) \in \Pi, Pos \subseteq X, Neg \cap X = \varnothing\}$.

For every finite normal program Π, let $Cn^+(\Pi)$ be defined iff there is a fixed point X under S_Π, s.t. for all sets Y of literals there is an $i \in \mathbb{N}$ with $S_\Pi^i(Y) = X$; in the latter case, let $Cn^+(\Pi) = X$.

For basic Π we have $S_\Pi(X) = T_\Pi(X)$, i.e., S_Π is an extension of T_Π to the case of negation as failure. Moreover, for normal programs in general, X is closed under Π iff $S_\Pi(X) \subseteq X$, and X is supported by Π iff $X \subseteq S_\Pi(X)$; therefore, the fixed points of S_Π are precisely the sets which are closed under Π and supported by Π. Since answer sets are closed under Π, it also holds for all answer sets X that $S_\Pi(X) \subseteq X$. If Π is normal and hierarchical, then $Cn^+(\Pi)$ is defined and $Cn^+(\Pi) = Cn(\Pi)$. In the general case, Cn^+ is defined iff there is a fixed point under S_Π which is the limit of "bottom up" S_Π-iteration independently of what the initial set Y looks like, i.e., where the result only depends on Π.

In [6], section 6, we have shown how to associate finite hierarchical inhibition nets with finite normal hierarchical logic programs, and vice versa. We will generalize these results now. If we are given an arbitrary finite inhibition net, we can construct a "counterpart program", s.t. excitatory connections are simulated by rules with positive bodies and inhibitory connections are replaced by negation as failure. Since inhibition nets always compute on input states, whereas logic programs do not have inputs, the input states have to be transformed into bodyless rules:

Definition 6.2. *Let $\mathcal{I} = \langle N, E, I, bias\rangle$ be a finite inhibition net:*
the program $\Pi(\mathcal{I})$ associated with $\mathcal{I}$ is defined in the following way:

1. *take N as the set P of propositional variables (but if there is no edge from bias to some other node or edge, simply drop bias);*
2. *for each $n \in N$ add all rules of the form $n \leftarrow n', not\ n_1, \ldots, not\ n_j$, where*
   - *$n'\ E\ n$,*
   - *for all i with $1 \leqslant i \leqslant j$: $n_i\ I\ \langle n', n\rangle$,*
   - *for all $n'' \in N$: if $n''\ I\ \langle n', n\rangle$ then $\exists i$ with $1 \leqslant i \leqslant j$ s.t. $n'' = n_i$.*
3. *do not add any further rules.*

Let $s^* \in S$; the program $\Pi(\mathcal{I}, s^*)$ associated with the net $\mathcal{I}$ and the input s^* is defined as follows:

1. *take N as the set P of propositional variables (but, again, if there is no edge from bias to some other node or edge drop bias);*
2. *add all rules contained in $\Pi(\mathcal{I})$;*
3. *add all bodyless rules with head n iff $s^*(n) = 1$;*
4. *do not add any further rules.*

In [6], p.173, we have shown that every Boolean mapping may be computed by a finite hierarchical inhibition net. In particular, one can always construct a conjunction node $n_1 \wedge \ldots \wedge n_i$ of nodes $n_1, \ldots, n_i$ (if $i = 1$ then we regard n_1 as a "conjunction" node), s.t. the conjunction node fires in a stable state if and only if each of $n_1, \ldots, n_i$ fires; for the construction one has to add a subnet of auxiliary inter-nodes. Let us assume that the signal propagation through the auxiliary nodes takes just one step of time. The next definition indicates how one may simulate any given finite normal logic program by means of an inhibition net, s.t. conjunction nodes replace the positive parts of rule bodies, and where inhibitory lines replace negation as failure. The input associated with a program is essentially defined as the class of heads of its bodyless rules:

Definition 6.3. *Let Π be a finite normal program (allowing for negation as failure) based on a set P of propositional variables:*
the inhibition net $\mathcal{I}(\Pi) = \langle N_\Pi, E_\Pi, I_\Pi, bias_\Pi\rangle$ associated with Π is given as follows:

1. $N_\Pi = P \cup \{bias_\Pi\} \cup$ *the set of auxiliary nodes needed for the construction of conjunction nodes;* $bias_\Pi$ *is some object not contained in* P,
2. *for all* $n, n_1, \ldots, n_{i+j} \in N$: $(n_1 \wedge \ldots \wedge n_i)\ E_\Pi\ n$ *iff there is a rule* $n \leftarrow n_1, \ldots, n_i,$ *not* $n_{i+1}, \ldots,$ *not* n_{i+j} *in* Π,
3. *for all* $n, n', n_1, \ldots, n_i, n_{i+2}, \ldots, n_{i+j} \in N$: $n'\ I\ \langle (n_1 \wedge \ldots \wedge n_i), n\rangle$ *iff there is a rule* $n \leftarrow n_1, n_2, \ldots, n_i,$ *not* $n',$ *not* $n_{i+2}, \ldots,$ *not* n_{i+j} *in* Π.

The input $s^*(\Pi)$ *associated with* Π *is defined as* $\{bias_\Pi\}$ *joined with the set of all propositional variables* n *s.t.* n *is contained in* Π *as a bodyless rule.*

The following theorem expresses that the two definitions above are, in a sense, sound and compatible:

Theorem 6.4. 1. *Let* $\mathcal{I} = \langle N, E, I, bias\rangle$ *be an FHIN,* $s^* \in S$, Π *be a finite normal hierarchical program:*

   1. $\Pi(\mathcal{I})$ *and* $\Pi(\mathcal{I}, s^*)$ *are finite normal hierarchical programs.*
   2. $\mathcal{I}(\Pi)$ *is an FHIN,* $s^*(\Pi)$ *is a state of* $\mathcal{I}(\Pi)$.
   3. $Cn(\Pi(\mathcal{I}, s^*)) = Cl_\mathcal{I}(s^*)$.
   4. $Cl_{\mathcal{I}(\Pi)}(s^*(\Pi)) \setminus [\{bias_\Pi\} \cup$ *the set of auxiliary nodes*$] = Cn(\Pi)$.

2. *Let* $\mathcal{I} = \langle N, E, I, bias\rangle$ *be a finite inhibition net,* $s^*, s \in S$, Π *be a finite normal program:*

   1. $\Pi(\mathcal{I})$ *and* $\Pi(\mathcal{I}, s^*)$ *are finite normal programs.*
   2. $\mathcal{I}(\Pi)$ *is a finite inhibition net,* $s^*(\Pi)$ *is a state of* $\mathcal{I}(\Pi)$.
   3. $S_{\Pi(\mathcal{I}, s^*)}(s) = F^\mathcal{I}_{s^*}(s)$.
   4. $F^{\mathcal{I}(\Pi)}_{s^*(\Pi)}(s) \setminus [\{bias_\Pi\} \cup$ *the set of auxiliary nodes*$] = S_\Pi(s)$.
   5. *If* $Cl_\mathcal{I}(s^*)$ *is defined, then:*
      $Cn^+(\Pi(\mathcal{I}, s^*)) = Cl_\mathcal{I}(s^*)$.
   6. *If* $Cn^+(\Pi)$ *is defined, then:*
      $Cl_{\mathcal{I}(\Pi)}(s^*(\Pi)) \setminus [\{bias_\Pi\} \cup$ *the set of auxiliary nodes*$] = Cn^+(\Pi)$.

3. *Let* $\mathcal{I} = \langle N, E, I, bias\rangle$ *be a finite inhibition net without inhibitory connections,* Π *be a finite basic normal program: results analogous to the ones in 1 hold.*

The last theorem shows that there is a close connection between inhibition nets and logic programs. Definitions and results involving the one family of mechanisms (e.g. concerning antitonicity and preferentiality) may be translated into definitions results for the other. The well-known translations between logic programming on the one hand and default logic, autoepistemic logic, circumscription, and truth maintencance systems on the other are thus also applicable to inhibition networks, although we do not have space to elaborate on this more specifically (but see Leitgeb[6], section 6).

But there is an essential difference between interpreted nets and logic programs on the interpretative level, i.e. on the level where some kind of meaning is assigned to entities like nodes, or to the processes acting upon these entities. The main idea used in logic programming is (i) to assign meaning to the very entities (nodes $\approx$ propositional variables) which are used as the constituents of the local rules governing the nonmonotonic inference process, and (ii) to assign meaning to the local rules themselves in some way. E.g., we might implement a logic program using the propositional variables b, p, f and the single rule $f \leftarrow b, not\ p$. The entities having representational function are the propositional variables b, p, f standing for birds, penguins, and flyers respectively, and the local rule $f \leftarrow b, not\ p$ by which birds are believed to be flyers as long as they are not believed to be penguins. The propositional variables which are subject to this local rule are thus also interpreted. On the other hand this is not true for interpreted inhibition nets (compare our two examples at the end of section 3): while the entities which represent are the patterns of activity, the entities subject to local activation rules are the nodes in a network. In particular, there are no abnormality nodes but abnormality is represented in the network "implicitly". Edges are not interpreted at all in the case of inhibition nets, and generally it will even be impossible to read any content into a single connection. Finally, note that in logic programs propositional variables are used as nodes, whereas in interpreted networks propositional variables are interpreted as sets of nodes. We also do not make use of "negative nodes" – corresponding to negative literals – because negation is just defined on the level of patterns and not below. In this way, full-fledged KLM-cumulative/loop-cumulative/preferential/cumulative monotonic/monotonic inference relations $\mid\!\sim$ may be associated with inhibition nets and thus also with logic programs, where the relation holds between formulas of arbitrary propositional complexity and not just between formulas involving (conjunctions of/disjunctions of) literals. Our results from above complement Dix's (see e.g. Brewka et al.[2], section 7.2) study of a classification of logic programming semantics in terms of inference relations with "strong principles", but with the difference that we do not refer to the level nodes but to the level of interpreted activation patterns. By this transition we are not just able to prove soundness but even completeness theorems in the sense of KLM[5]. The same results may be shown for logic programs if the network semantics that we have introduced above for inhibition nets is adopted for logic programs.

References

[1] C. Balkenius, P. Gärdenfors, "Nonmonotonic inferences in neural networks," in: J.A. Allen, R. Fikes, E. Sandewall (eds.), *Principles of Knowledge Representation and Reasoning*, San Mateo: Morgan Kaufmann, 1991, 32–39.

[2] G. Brewka, J. Dix, K. Konolige, *Nonmonotonic Reasoning. An Overview*, Stanford: CSLI Lecture Notes 73, 1997.

[3] D. Dubois, H. Prade, "Conditional objects, possibility theory and default rules," in: G. Crocco et al. (eds.), *Conditionals: From Philosophy to Computer Science*, Oxford: Oxford University Press, 301-336.

[4] P. Gärdenfors, "How Logic Emerges from the Dynamics of Information," in: J. Van Eijck, A. Visser (eds.), *Logic and Information Flow*, Cambridge: The MIT Press, 1994, 49–77.

[5] S. Kraus, D. Lehmann, M. Magidor, "Nonmonotonic reasoning, preferential models and cumulative logics," *Artificial Intelligence* 44 (1990), 167–207.

[6] H. Leitgeb, "Nonmonotonic reasoning by inhibition nets," *Artificial Intelligence* 128[1-2] (2001), 161-201.

[7] H. Leitgeb, *Inference on the Low Level*, Salzburg: Doctoral Dissertation, University of Salzburg, Austria, 2001.

[8] H. Leitgeb, "New Results on the Network Semantics for Cumulative Systems of Nonmonotonic Reasoning", *Informatikberichte* 287-8/2001, KI-2001 Workshop Uncertainty in Artificial Intelligence (2001), 37-52.

[9] H. Levesque, "All I know: a study in autoepistemic logic," *Artificial Intelligence* 42 (1990), 263–309.

[10] V. Lifschitz, Foundations of Logic Programming, in: G. Brewka (Ed.), *Principles of Knowledge Representation*, CSLI, Stanford, 1996, pp. 69–127.

[11] D. Makinson, "General Patterns in Nonmonotonic Reasoning," in: D.M. Gabbay, C.J. Hogger, J.A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming 3*, Oxford: Clarendon Press, 1994, 35–110.

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, San Mateo: Morgan Kaufmann, 1988.

[13] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel Distributed Processing*, Vol 1 and 2, Cambridge: The MIT Press, 1986.