Chapter 7
Neural Network Models of Conditionals

Hannes Leitgeb

Abstract This chapter explains how artificial neural networks may be used as models for reasoning, conditionals, and conditional logic. It starts with the historical overlap between neural network research and logic, it discusses connectionism as a paradigm in cognitive science that opposes the traditional paradigm of symbolic computationalism, it mentions some recent accounts of how logic and neural networks may be combined, and it ends with a couple of open questions concerning the future of this area of research.

7.1 Introduction

Neural networks are abstract models of brain structures capable of adapting to new information. The learning abilities of artificial neural networks have given rise to successful computer implementations of various cognitive tasks, from the recognition of facial images to the prediction of currency movement. Under the heading 'deep learning', neural networks have become prominent again lately as major tools in the field of machine learning.

Logic deals with formal systems of reasoning; in particular, inductive logic studies formal systems of reasoning towards plausible but uncertain conclusions. As evidence accumulates, the degree to which it supports a hypothesis, as measured by the logic, should tend to indicate that the hypothesis is likely to be true.

Although sharing a joint focus on information and reasoning, until recently these two areas developed in opposition to each other: neural networks are quantitative dynamic systems, while logical reasoners must be symbolic systems; networks are described by mathematical equations, whereas logic is subject to normative statements about how we ought to reason; neural networks have been studied by

H. Leitgeb (✉)
Ludwig-Maximilians-University, Munich, Germany
e-mail: Hannes.Leitgeb@lmu.de

S. O. Hansson, V. F. Hendricks (eds.), *Introduction to Formal Philosophy*, Springer Undergraduate Texts in Philosophy, https://doi.org/10.1007/978-3-319-77434-3_7

scientists, whilst the classical "problem of induction" is regarded as belonging to philosophy. And so forth.

In recent years, however, this assessment has been changing: the emergence of logical formalisms for uncertain reasoning and the discovery that these formalisms apply to neural net processes on the representational level give rise to the expectation that the dynamics of artificial neural networks can be understood in terms of logically valid, and thus rational, rules of inference. As neural networks, commonsense reasoning, and maybe even scientific induction seem to conform to similar logical systems, a joint theoretical framework might be in the offing which might lead to new insights into the logical and cognitive basis of everyday reasoning, language, and science.

In this article we will focus on one outcome of these new developments: neural network semantics for conditionals. We will start with McCulloch's and Pitts' original interpretation of neural network components in terms of formulas of classical propositional logic, we will summarize the main features of connectionism which emerged as an alternative paradigm of cognitive science that was thought to be in opposition to logical takes on reasoning, and we will sketch how recent theories nevertheless attempt to describe states and processes in neural networks by means of logical terms. Finally, we will deal with one of these theories in more detail. Along the way we will also present very brief recaps of nonmonotonic reasoning and of the logical and philosophical literature on conditionals, as far as this serves the purpose of illuminating the neural networks models of conditionals that are the topic of this paper. We end with some tentative philosophical conclusions and with a list of interesting open questions. Clearly, this new field of research is relying heavily on the application of formal methods, mostly from logic and the mathematical theory of dynamical systems.[1]

7.2 Neural Networks as Models of Reasoning

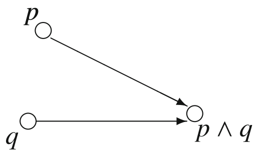
In their famous article "A Logical Calculus of the Ideas Immanent in Nervous Activity", McCulloch and Pitts [45] first introduced artificial neural networks as mathematical abstractions from neural circuits in the brain. A McCulloch-Pitts network consists of a set of nodes and a set of connections between these nodes. Each node can be in one of two possible states: it fires (1), or it does not (0). Each connection is of one of two possible kinds: along inhibitory connections, nodes receive inhibitory signals by which they get deactivated at the next point of time (on a discrete time scale). Via excitatory connections, signals are transferred from

[1] This article is a revised and extended version of: Leitgeb [41]. Some material contained in Leitgeb [37, 38], Ortner and Leitgeb [46], and in the popular and non-technical exposition of logic and neural networks in Leitgeb [39] was used, too.

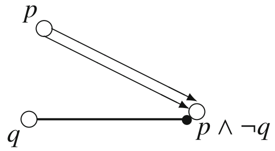
We are grateful for generous support received from the Alexander von Humboldt Foundation.

one node to another which have a stimulating effect on the target node: if the node does not get inhibited, and if the number of all incoming excitatory signals exceed or are identical to some fixed threshold value that is associated with the node, then the node fires at the next point of time. Although these appear to be quite simple devices, McCulloch and Pitts [45] effectively established that in principle every finite automaton can be realized by such a McCulloch-Pitts network (a formal result which was later made perfectly precise by the logician S. Kleene in his [31]). Furthermore, the state transitions which take place in such networks allow for a description in logical terms: if the activity of a node is considered as a truth value, then the node itself may be regarded as an entity which *has* a truth value, i.e., as a formula or proposition. If the "truth value" of a node does not depend on the "truth values" of other nodes (but, say, only on some given input), then it is indeed natural to regard such nodes as *atomic* formulas or propositions. Accordingly, if nodes are put together in a network, such that connections between nodes can cause the "truth values" of other nodes to be altered, then the latter nodes may be taken to correspond to *complex* formulas; the semantic dependency of the truth value of a complex formula on the truth values of its component formulas is thus represented by the network topology and the choice of thresholds.

As an example, consider two very elementary McCulloch-Pitts networks: In the first network, excitatory connections lead from nodes p and q to a third node. If this latter node has a threshold value of 2, then the node is going to fire if and only if both p and q were active at the previous point of time. So we can associate the formula $p \wedge q$ with this node:



In the second network, two excitatory lines lead from p to the output node, whereas q is connected to the latter by an inhibitory edge. If e.g. the output node has a threshold of 2, it will be activated at the next point of time if and only if p is set to 1 and q is set to 0 (and therefore does not have any inhibitory influence). Hence, the third node in the network corresponds to the formula $p \wedge \neg q$:



This way of associating nodes in networks with formulas in the language of classical propositional logic extends to more interesting networks with multiple layers of nodes and with more complex patterns of excitatory and inhibitory connections. E.g., it would be easy to extend the second network by a node that represents $\neg(p \wedge \neg q)$, i.e., a formula which is logically equivalent to the material conditional $p \supset q$. If our brains were, at least on some level, similar to neural networks of the McCulloch-Pitts kind, they could thus be understood as collections of simple logical units put together in order to calculate binary truth values from external or internal input. The calculation of the truth values of material conditionals would be a special case of this form of computational processing.

Of course, the McCulloch-Pitts networks are, in several respects, much too simple to be plausible models of actual neural networks in animal or human brains. In particular, they are not yet able to learn. The next decisive step in the development of artificial neural networks was to introduce variable weights that are attached to connections and which encode the degree of influence that nodes can exert on their target nodes via these connections. By sophisticated learning algorithms, these weights can be adjusted in order to map inputs to their intended outputs, e.g., facial images of persons to the names of these persons, or verbs to their correct past tenses. Despite some initial success in the 1950s and 1960s – mainly associated with F. Rosenblatt's *Perceptrons* which famously came under attack by M. Minsky's and S. Papert's monograph with the same title – it was only in the 1980s that artificial neural network models of cognition became serious contenders to the dominant symbolic computation paradigm in artificial intelligence. These new approaches to cognition are usually subsumed under the term 'connectionism'.[2] As we will explain below, the more recent neural network models do not only differ from the original McCulloch-Pitts networks in terms of complexity and learning abilities, they also differ in terms of the interpretation of their components: instead of assigning meaning – expressed by formulas – to *single* nodes, the modern approach emphasizes that it is rather *patterns* or *sets* of nodes which receive an interpretation.

How does 'cognition by neural networks' relate to the traditional 'cognition by symbolic computation' paradigm of cognitive science (exemplified by classical Artificial Intelligence)? According to the latter, (i) intelligent cognition demands structurally complex mental representations, such that (ii) cognitive processing is only sensitive to the form of these representations, (iii) cognitive processing conforms to rules, statable over the representations themselves and articulable in the format of a computer program, (iv) (standard) mental representations have syntactic structure with a compositional semantics, and (v) cognitive transitions conform to a computable cognitive-transition function (we adopt this characterization essentially from [30], with slight deviations). Intelligent cognition is supposed to be "systematic" and "productive" (see [21]), i.e., the representational capacities of intelligent agents are supposed to be necessarily closed under various representation-transforming and representation-generating operations (e.g., if an

[2] Rumelhart et al. [51] is still something like the "bible" of connectionism; Rojas [50] is a nice introduction to neural networks, and at ⟨http://plato.stanford.edu/entries/connectionism/⟩ the entry on connectionism in the *Stanford Encyclopedia of Philosophy* can be found – have a look at these for more background information.

agent is able to represent that aRb, it is also able to represent that bRa, etc.). This capacity is hypothesized to be due to the combinatorial properties of languages of mental symbols based on a recursive grammar. A cognitive agent that conforms to the symbolic computation paradigm has the belief that φ if and only if a corresponding sentence φ is stored in the agent's symbolic knowledge base. The rules that govern cognitive processes according to the symbolic computation paradigm are either represented within the cognitive agent as symbolic entities themselves, or they are hard-wired. Inference processes are taken to be internalizations of derivation steps within some logical system, and the alleged "systematicity" of inferences (see again [21]) is explained by the internal representation or hard-wiring of rules which are only sensitive to the syntactic form of sentential representations.

Cognition by artificial neural networks, on the other hand, belongs to the so-called dynamical systems paradigm of cognitive science which can be summarized by what van Gelder [59] calls the "dynamical hypothesis": "for every kind of cognitive performance exhibited by a natural cognitive agent, there is some quantitative [dynamical] system instantiated by the agent at the highest relevant level of causal organization [i.e., at the level of representations], so that performances of that kind are behaviors of that system" [59, p. 622]. A dynamical system may be regarded as a pair of a state space and a set of trajectories, such that each point of the space corresponds to a total cognitive state of the system, and every point of the space lies precisely on one trajectory. If a certain point corresponds to the system's total cognitive state at some time, the further evolution of the system follows the trajectory emanating at this point. Usually, such systems are either defined by differential equations, or by difference equations, defined over the points of the state space: in the first case one speaks of continuous dynamical systems with continuous time, while in the latter case one speaks of discrete dynamical systems with discrete time. In the discrete case, the set of trajectories may be replaced by a state-transition mapping, such that each trajectory is generated by the iterated application of the mapping. A *cognitive* dynamical system is a dynamical system with representations, i.e., where states and state transitions can be ascribed content or interpretation. The dynamic systems paradigm assumes that intelligent cognition takes place in the form of state-transitions in quantitative systems, i.e., systems in which a metric structure is associated with the points of the state space, and where the dynamics of the system is systematically related to the distances measured by the metric function. The distances between points may be regarded as a measure of their similarity *qua* total cognitive states. Moreover, the typical dynamical systems that are studied within the dynamical systems paradigm also have a vector space structure, and thus they "support a geometric perspective on system behaviour" [59, p. 619].

Connectionism is the most important movement within the dynamical systems paradigm: Artificial neural networks are the dynamical systems that the connectionists are interested in. Smolensky [54] characterizes connectionism by the following hypotheses: (i) "The connectionist dynamical system hypothesis: The state of the intuitive processor at any moment is precisely defined by a vector of numerical values (one for each unit). The dynamics of the intuitive processor are governed by a differential equation. The numerical parameters in this equation constitute the processor's program or knowledge. In learning systems, these parameters change according to another differential equation." (ii) "The subconceptual unit hypothesis: The entities in the intuitive processor with the semantics of conscious concepts of the task domain are complex patterns of activity over many units. Each unit participates in many such patterns." (iii) "The subconceptual level hypothesis: Complete, formal, and precise descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level." The subconceptual level is the level of analysis that is preferred by the connectionist paradigm, or, as Smolensky expresses it, by the *subsymbolic* paradigm; it lies "below" the conceptual level that is preferred by the symbolic computation paradigm, but "above" the neural level preferred by neuroscience.

Claim (i) proves connectionism to belong to the dynamical systems paradigm. The subconceptual unit hypothesis (ii) and the subconceptual level hypothesis (iii) highlight the main differences between the old McCulloch & Pitts approach presented above and modern day connectionism: by (ii), single nodes or single connections in a neural network are normally not supposed to carry any meaning at all; the representing units are distributed patterns of activation that involve a great number of nodes or even the network topology as a whole (see van Gelder [60] on "Distributed versus local representation"). In more metaphorical terms: there is not generally anything like a "grandmother cell", i.e., a single neuron that would correspond to a very complex formula which describes your grandmother and which would fire if and only if your grandmother were perceived. Rather, your grandmother's being perceived is represented by some complex pattern of activation which spreads throughout parts of the network at the time of perception. Furthermore, by (iii), if symbols can be attached to the activation patterns of nodes or to some other "global" aspects of neural networks at all, the transitions from one representing item – one pattern – to another will no longer be effected on the level of these representing items themselves but rather on the sub-symbolic level of nodes and edges. Therefore, for connectionists in the sense described, it seems impossible to translate the computations on the sub-symbolic level into sequences of rules on the symbolic level, let alone into logical rules which apply to complex symbolic expressions. Thus, McCulloch and Pitts' original *logical* approach to neural networks became something like the paradigmatic antagonist of the movement, and hence it had to be given up, or so it seemed. Instead of analyzing cognition in terms of localized representations of formulas – "hard constraints" – Smolensky [54, p. 18], suggests that connectionist cognition proceeds by means of "soft constraints": "Formalizing knowledge in soft constraints rather than hard rules has important consequences. Hard constraints have consequences singly; they are rules that can be applied separately and sequentially – the operation of each proceeding independently of whatever other rules may exist. But soft constraints have no implications singly; any one can be overridden by the others. It is only the entire set of soft constraints that has any implications. Inference must be a cooperative process [. . .] Furthermore, adding additional soft constraints can repeal conclusions that were formerly valid: Subsymbolic inference is fundamentally nonmonotonic." If human reasoning is as connectionists describe it, then McCulloch and Pitts' account of reasoning in terms of neural network implementations of truth functions in classical logic can hardly be adequate.

Even if this very last statement about McCulloch and Pitts' theory is true, this does not yet entail that the symbolic computation paradigm and the dynamical systems paradigm themselves have to be completely mutually exclusive, i.e.: significant aspects of the two paradigms could actually turn out to be compatible with each other. As Gärdenfors [23, p. 67f], suggests, the two paradigms might even be complementing each other: "they are best viewed as two different perspectives that can be adopted when describing the activities of various computational devices." Results on symbol manipulation in networks (see e.g. [12, 13, 55]), neural networks approaches to grammar representation (see e.g. [33, 56]), and hybrid systems that involve both neural network and symbolic components (see e.g. [10, 49]) indicate that there might be continuous paths of transition from the one paradigm to the other. In particular, the analysis of neural networks in terms of *logical laws and rules* has become a topic of research again in recent years, and on it we are going to focus now.

Here are some relevant references on logical accounts of neural network cognition (they can also be found in the bibliography – note that this is a *very* incomplete list though!):

- A.S. d'Avila Garcez, K. Broda, and D.M. Gabbay [15].
- A.S. d'Avila Garcez, K.B. Broda, and D.M. Gabbay [16].
- A.S. d'Avila Garcez, L.C. Lamb, and D.M. Gabbay [17].
- A.S. d'Avila Garcez et al. [18].
- S. Bader and P. Hitzler [3].
- C. Balkenius and P. Gärdenfors [4].
- R. Blutner [9].
- E.-A. Dietz, S. Hölldobler, and L. Palacios [19].
- P. Hitzler, S. Hölldobler, and A.K. Seda [26].
- S. Hölldobler [27].
- S. Hölldobler [29].
- S. Hölldobler and Y. Kalinke [28].
- H. Leitgeb [35].
- H. Leitgeb [37].
- H. Leitgeb [38].
- R. Ortner and H. Leitgeb [46].
- K. Stenning and M. van Lambalgen [57]

The main idea behind all of these theories is that if classical logic is replaced by a different logical calculus – in particular, by a system of nonmonotonic reasoning that is closer to the commonsense reasoning that our brains are usually involved in – then a logical description or characterization of neural network states and processes might be possible in a way, such that: (i) "The connectionist dynamical system hypothesis" is satisfied, maybe even in combination with (ii) "The subconceptual unit hypothesis", yet (iii) "The subconceptual level hypothesis" turns out to be false (for the precise statements of these theses see above). In other words: Logical descriptions of reasoning might become tractable again at the conceptual level, even when reasoning is realized in terms of the dynamics of an artificial neural network.[3]

Here is a brief, and very sketchy, guide to the literature as cited above:

A.S. d'Avila Garcez et al. [18], Bader and Hitzler [3], and Hölldobler [29] are very useful survey papers. Many authors and papers in this area of research can be found by checking the websites of the "NeSy" events in the workshop series on Neural-Symbolic Learning and Reasoning, which has been an ongoing endeavour since 2005.

The Hölldobler et al. group in Dresden has done pioneering work on how to generate neural networks from *logic programs*.[4] (See also Stenning and van Lambalgen, Chapter 7.) A logic program consists of rules which may look like this:

$$CanFly(Tweety) \Leftarrow Bird(Tweety), \neg Penguin(Tweety)$$

This is to be read as: if one has the information that Tweety is bird *but one lacks the information that Tweety is a penguin*, then one may infer that Tweety can fly. '$\Leftarrow$' here is much like the (non-material) if-then symbol in the sequent calculus of classical logic which connects the two sides of a sequent. Rational inferences that are based on such rules are *nonmonotonic*: given additional information, such as that Tweety is in fact a penguin, the inference would not longer be supported. Note that negation here is what is called *default negation*: e.g., $\neg Penguin(Tweety)$ expresses the *absence* of the positive information $Penguin(Tweety)$. What Hölldobler et al. managed to show was that it is possible to transform such logic programs into artificial neural networks, so that: the atomic formulas used in a given logic program correspond to the input nodes and to the output nodes in a feed-forward network; the rules in the logic program correspond to the nodes in the hidden layers; positive and negative information in the bodies of rule clauses correspond to excitatory and inhibitory connections, respectively; and additional feedback connections from the output nodes to the input nodes enable the network to converge on a model for the rules of the logic program, such that the model corresponds to a stable network state.

The group around d'Avila Garcez et al. has built on, and added to, this work, amongst others (i) by suggesting extraction methods that reverse the process just described by generating logic programs from (trained) neural networks, and (ii) by extending the results to logic programs that involve modal operators or that are based on intuitionistic logic. What the theories of these two groups have in common, too, is that they lie on the *consistency-based fixed point operator side*

[3] See Brewka et al. [11] for a very nice overview of nonmonotonic reasoning, Makinson [44] for a comprehensive logical treatment of the subject, and Schurz and Leitgeb [53] for a compendium of articles on cognitive aspects of nonmonotonic reasoning. Ginsberg [24] is an outdated collection of articles but it is still very useful if one wants to see what nonmonotonic reasoning derives from.

[4] Brewka et al. [11] includes a very clear and accessible introduction to logic programming.

of nonmonotonic reasoning: explained in terms of the rule above, as long as it is *consistent* to assume that Tweety is not a penguin, one may infer that Tweety is bird; what one is ultimately supposed to believe given evidence is computed by generating a fixed point of an immediate-consequence operator that is determined by the logic program.[5]

However, there is also the more recent *preference-based nonmonotonic inference relation* side of nonmonotonic reasoning, which became prominent through the now classical articles by Kraus, Lehmann, and Magidor [32] and Lehmann and Magidor [34]. In these approaches, metalinguistic statements such as

$$Bird(Tweety) \mid\!\sim CanFly(Tweety)$$

are considered which are now interpreted as saying: *in the most preferred (most normal, most plausible) worlds* in which Tweety is a bird, Tweety is also able to fly. Here, '$\mid\!\sim$' is a binary metalinguistic predicate which is syntactically like the symbol '$\models$' for classical logical consequence. If one replaces such metalinguistic statements by conditionals in the object language, such as by

$$Bird(Tweety) \Rightarrow CanFly(Tweety),$$

and one makes the preference-based semantics for these conditionals precise, the resulting semantics ends up being very close indeed to standard semantics for conditional logic as developed by philosophical logicians since the 1960s and 1970s (about which more in the next section). This preference-based approach is characterized by having much nicer logical properties than its consistency-based fixed-point operator counterpart. For instance, in all of the preference-based calculi, the following two rules (now spelled out in terms of conditionals)

$$\frac{\varphi \Rightarrow \psi, \varphi \wedge \psi \Rightarrow \rho}{\varphi \Rightarrow \rho} \text{ (Cautious Cut)}$$

$$\frac{\varphi \Rightarrow \psi, \varphi \Rightarrow \rho}{\varphi \wedge \psi \Rightarrow \rho} \text{ (Cautious Monotonicity)}$$

are logically valid. The combination of these two rules is usually referred to by the term 'cumulativity' (see [32]). Cumulativity expresses that adding inferred formulas to the evidence neither increases nor decreases the inferential strength of the evidence. However, the rule

$$\frac{\varphi \Rightarrow \rho}{\varphi \wedge \psi \Rightarrow \rho} \text{ (Monotonicity)}$$

[5] Almost all of the classical approaches to nonmonotonic reasoning from the 1980s, such as default logic, inheritance networks, truth maintenance systems, circumscription, and autoepistemic logic belong to this class of nonmonotonic reasoning mechanisms.

is not logically valid anymore, which is why one cannot simply infer from

$$Bird(Tweety) \Rightarrow CanFly(Tweety)$$

that also

$$Bird(Tweety) \wedge Penguin(Tweety) \Rightarrow CanFly(Tweety)$$

holds. In contrast with the former approach to nonmonotonic reasoning, exceptions do not have to be stated explicitly anymore in the relevant rules or conditionals.[6]

It is nonmonotonic inference in this latter preferential sense that Balkenius and Gärdenfors represented in terms of state transitions within artificial neural networks, and which they studied by means of concrete experiments in computer simulations. Leitgeb's work builds on Balkenius and Gärdenfors' approach but adds soundness and completeness proofs for systems of nonmonotonic reasoning or conditional logic based on a corresponding neural network semantics. Blutner also starts from Balkenius and Gärdenfors but represents nonmonotonic inferences in so-called weight-annotated Poole systems by means of state-transitions in Hopfield networks, relating the results so obtained to Harmony or Optimality Theory in the sense of Smolensky and Legendre [56]. One point of difference between Blutner's and Leitgeb's theories – and one of agreement between Blutner's theory and the theories by Hölldobler et al. and d'Avila Garcez et al. – is that while Blutner represents atomic formulas in neural networks in terms of *nodes*, Leitgeb represents atomic formulas as *distributed patterns of activity*. Accordingly, generally, connections between nodes cannot be assigned any local symbolic interpretation anymore in Leitgeb's account. Since distributed representation was supposed to be one of the hallmarks of connectionism – in correspondence with (ii) "The subconceptual unit hypothesis" from above – we will concentrate on Leitgeb's theory in Sect. 7.4, where we will present the theory as a neural networks semantics for conditionals.

If any of these logical accounts of neural network cognition were to prove successful in the long run (logically, philosophically, and in applications), the gap between the dynamic systems paradigm and the symbolic computation paradigm in cognitive science would be bridged, or, at the very least, diminished. This would also constitute an important step in understanding what neural networks actually do; otherwise, we might be stuck with an ingenious technical machinery that maps an input to its desired output, but where the process that leads from the one to the other remains uninterpreted, unexplained, and unjustified. While it is certainly true that current implementations of machine learning do not themselves rely on the application of logical methods (see Wheeler [61]), logic might still play a role in the *rational reconstruction* and *assessment* of machine learning: in checking whether

[6] For more on the differences between the two sides of nonmonotonic reasoning, see Brewka et al. [11].

the “black box” conforms to norms of rationality and, perhaps, morality. Progress on the logic of neural networks might also lead to new insights in uncertain reasoning, induction, and even the philosophy of science – we will return to this in the final section of this article, which will include a list of open questions.[7]

7.3 A Brief Recap on Conditionals

Before I turn to a concrete example of a neural network semantics for conditionals, let me say a bit more about the conditionals that will be involved. Conditionals are sentences of an ‘if… then…’ form; so, the logical form of a conditional is an expression of the form

$$\text{If } \varphi, \text{ then } \psi$$

or, more formalized,

$$\varphi \Rightarrow \psi$$

where φ is called the ‘antecedent’ of the conditional and ψ its ‘consequent’; both the antecedent and the consequent of a conditional are sentences. (See also John Cantwell’s chapter 6 on conditionals in this handbook.)

Conditionals are crucial in everyday communication, especially when we want to convey information that goes beyond the currently present perceptual situation. Conditionals also play a major role in philosophical theories about dispositions, causality, laws, time, conditional norms, probability, belief, belief revision, and so forth. Finally, conditionals are closely related closely to quantifiers, such as ‘All φ are ψ’, ‘There are φ which are ψ’, ‘Most φ are ψ’, etc.[8] But note that in these latter cases, ‘φ’ and ‘ψ’ are place holders for *open formulas* – formulas with a free variable – rather than sentences.

Amongst conditionals in natural language, usually the following distinction is made[9]:

1. If Oswald had not killed Kennedy, then someone else would have.
2. If Oswald did not kill Kennedy, then someone else did.

[7] We should add that there are also results concerning the description of neural network states and processes by means of *classical* logic, over and above the traditional McCulloch and Pitts approach: see Pinkas [48] and Bechtel [5] for examples.

[8] See van Benthem [58] for a nice discussion of this relationship between conditionals and quantifiers; more can be found by consulting the theory of *generalized quantifiers* – see e.g. Peters and Westerstahl [47].

[9] The following famous example is due to Ernest Adams.

2 is accepted by almost everyone, whilst we do not seem to know whether 1 is true. This invites the following classification: a conditional such as 2 is called *indicative*, whereas a conditional like 1 is called *subjunctive*. In conversation, the antecedents of subjunctive conditionals are often assumed or presupposed to be false: in such cases, one speaks of these subjunctive conditionals as *counterfactuals*. Roughly, indicative conditionals represent the denoted act or state as an objective fact, while subjunctive conditionals represent a denoted act or state not as fact but as contingent or possible. Subjunctive and indicative conditionals may have precisely the same antecedents and consequents (as in the example above) while differing only in their conditional connectives, i.e., their 'if'-'then' occurrences having different meanings.

When logic developed into a serious philosophical and mathematical discipline in the late nineteenth and the early twentieth century, logicians quickly came up with two suggestions of how to formalize conditionals, whether indicative or subjunctive:

- $\varphi \supset \psi$: Formalization by means of material conditionals (material implications).
- $\varphi ⥽ \psi$: Formalization by means of strict conditionals (strict implications).

From an axiomatic point of view, the meaning of the former is given by any of the typical deductive systems for classical propositional logic. The logical systems for the latter were investigated intensively by C.I. Lewis, however it was only after the axiomatic systems of normal modal logic had been developed by S. Kripke that the analysis of $\varphi ⥽ \psi$ in terms of $\Box(\varphi \supset \psi)$ emerged as a standard (where $\Box$ is the necessity operator studied by modal logicians). On the semantic side, the meaning of $\supset$ is given by its well-known truth table, whereas the semantics of $⥽$ can be stated on the basis of the usual Kripkean possible worlds semantics of $\Box$.

These formalizations of the 'if… then…' in classical logic proved to be enormously successful, especially in the formalization of mathematical theories and of fragments of empirical theories. However, there was still a problem: both $\supset$ and $⥽$ are *monotonic*, i.e., the rule $\frac{\varphi \Rightarrow \psi}{\varphi \wedge \rho \Rightarrow \psi}$ is logically valid if '$\Rightarrow$' is replaced by either of the two connectives. On the other hand, there seem to be many instances of indicative and subjunctive conditionals in natural language which are *nonmonotonic*, i.e., for which the rule $\frac{\varphi \Rightarrow \psi}{\varphi \wedge \rho \Rightarrow \psi}$ should not assumed to be valid. E.g., 'If it rains, I will give you an umbrella' does not seem to logically imply "If it rains and I am in prison, I will give you an umbrella", nor does 'If it rained, I would give you an umbrella' seem to logically imply "If it rained and I were in prison, I would give you an umbrella". Accordingly, add e.g. '…and Kennedy in fact survived all attacks on his life' to the antecedent of 'If Oswald did not kill Kennedy, then someone else did' and the resulting conditional does not seem acceptable anymore. Therefore, philosophical logicans started to investigate new logical systems in which monotonicity (or *strengthening of the antecedent*) would not turn out to be logically valid. Lewis [43] is the classic treatise on counterfactuals as nonmonotonic conditionals, in which subjunctive conditionals are evaluated based on similarity orderings of possible worlds (which are similar to the preference orderings of possible worlds used in nonmonotonic reasoning). Since the nonmonotonicity phenomenon had already been well known in probability theory – a conditional probability $P(Y|X)$ being high does not entail the conditional probability $P(Y|X \cap Z)$ being high – it is not surprising that some of the modern accounts of conditionals instead relied on a probabilistic semantics: indeed, Adams [1] famously developed a probabilistic theory of indicative conditionals that does not support monotonicity (see Adams [2] for a more general overview of probability logic).[10]

It is these axiomatic and semantic systems of conditionals which got rediscovered a bit later (note: philosophy was *first*!) by theoretical computer scientists who initiated the field of nonmonotonic reasoning. For assume you want to represent in a computer system what happens to your car when you turn the ignition key: well, you might say, the car starts, so 'if the ignition key is turned in my car, then the car starts' seems to describe the situation properly. But what if the gas tank is empty? You better improve your description by saying 'if the ignition key is turned in my car and the gas tank is not empty, then the car starts'. However, this could still be contradicted by a potato that is clogging the tail pipe, or by a failure of the battery, or by an extra-terrestrial blocking your engine, or. . . The possible exceptions to 'if the ignition key is turned in my car, then the car starts' are countless, heterogeneous, and unclear. Nevertheless, we seem to be able to communicate information with such simple conditionals, and, equally importantly, we are able to reason with them in a rational manner. In order to do so we make use of a little logical "artifice": we do not really understand 'if the ignition key is turned in my car, then the car starts' as expressing that it is not the case that the ignition key is turned and the car does not start – after all, what is negated here might indeed be the case in exceptional circumstances – but rather that *normally*, or with a *high probability*, given the ignition key is turned, the car starts. Instead of trying to enumerate the indefinite class of exceptions in the if-part of a material or strict conditional, we tacitly or explicitly qualify 'if the ignition key is turned in my car, then the car starts' as holding only in normal or likely circumstances, whatever these circumstances may look like. As a consequence, the logic of such normality claims again differs from the logic of material or strict conditionals: 'if Tweety is a bird, then [normally] Tweety is able to fly' is, presumably, true, but 'if Tweety is a penguin bird, then [normally] Tweety is able to fly' is not, and neither is 'if Tweety is a dead bird, then [normally] Tweety is able to fly' or 'if Tweety is a bird with his feet set in concrete, then [normally] Tweety is able to fly'. So computer scientists found themselves in need of describing reality in terms of nonmonotonic normality conditionals on the basis of which computers should be able to draw justified inferences about the everyday world while being unaffected by the omnipresence of exceptions. And this need eventually led to conclusions very similar to those drawn by philosophers who cared about the logic and semantics of conditionals in natural language.

In the next section we will suggest that so-called *interpreted dynamical systems* may be used to yield a semantics for nonmonotonic conditionals. The logical

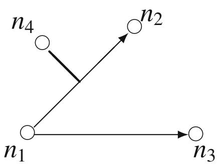
[10] For a textbook-like overview of the philosophical literature on indicative and subjunctive conditionals, see Bennett [7].

systems which turn out to be sound and complete with respect to such semantics are standard systems of conditional logic which have been studied both in philosophical logic and nonmonotonic reasoning. Interpreted artificial neural networks will be shown to be the paradigm case examples of such interpreted dynamical systems. Although the conditionals that are satisfied by such interpreted artificial neural networks are represented distributedly by these networks, the logical rules they obey are precisely the rules of systems which had been developed in order to make computers cope with the real world by means of symbolic computation, and which had been investigated even before by philosophers who intended to give a proper logical analysis of indicative and subjunctive conditionals. Since the dynamics of state changes in interpreted neural networks can be described correctly and completely by sets of conditionals that are closed under the rules of such logical systems, neural networks may be understood as nonmonotonic reasoners who, when they evolve under an input towards a state of "minimal energy", draw conclusions that follow from premises in all minimally abnormal cases.

7.4 From Dynamical Systems to Conditionals: Interpreted Dynamical Systems

Following Gärdenfors' proposal mentioned above, we will study cognitive dynamical systems from two complementary perspectives. On the one hand, cognitive dynamical systems such as neural networks can be described in terms of differential or difference equations, i.e., as *dynamical systems*. On the other hand, they exemplify cognitive states and processes that can be ascribed propositional contents which may in turn be expressed by sentences or formulas; so they are also *cognitive agents* or *reasoners*.

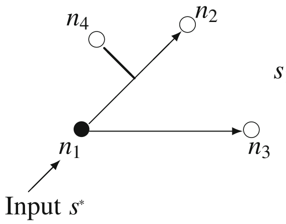
Here is an example. For the sake of simplicity, let us forget about the weights again that are attached to the edges of a typical neural network, and let us also assume that the activation functions that are defined for the nodes in such a network are as straight-forward and simple as in the case of the McCulloch-Pitts networks. Then we might, e.g., end up with a simple qualitative neural network that looks like this[11]:



[11] Such networks are called 'inhibition networks' in Leitgeb [35].

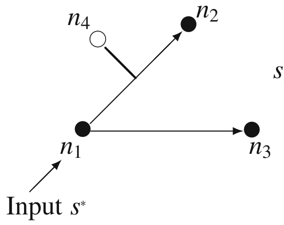
This is a network with four nodes. n_1 is connected both to n_2 and n_3 by excitatory connections. In contrast with traditional McCulloch-Pitts networks, there is also an inhibitory connection that leads from n_4 to the *excitatory connection* from n_1 to n_2. So, if n_4 is active, this is not going to directly inhibit the activation of some other node at the next point of time, but instead any activity by n_4 will have the effect that no excitatory stimulus will be able to pass the edge from n_1 to n_2 at the next point of time.

Now, say, node n_1 gets activated by some external stimulus, e.g., by some sensory signal coming from outside. We will assume that such inputs always remain constant for sufficiently long, hence, in the present example, one should think of n_1 as being activated from the outside until the computational process that we are interested in has delivered its final output. Formally, we can describe what is going on in the following way: the network is in an initial state s_0; e.g., the state in which no node fires. This state s_0 may be regarded as a mapping from the set of nodes into the set $\{0, 1\}$, such that each node is mapped to 0 or "inactivity". Furthermore, the network is fed an input s^* that makes n_0 fire but which activates no other node: it is useful to identify such an input with the network state that the input would generate just by means of external influences on the network. Thus, in our case, s^* will be the state in which the node n_0 is mapped to 1 and in which all other nodes are mapped to 0. The resulting dynamics of the network can be described by means of a state transition mapping F_{s^*} that is given relative to the (constant) incoming input – s^* – and which is applied to the initial state s_0 in order to determine the next state s_1 of the network. Since no node is active in s_0, the only nodes which will be active in s_1 will be those activated by the input itself, i.e., n_1. So we have:



$$s_1 = F_{s^*}(s_0)$$

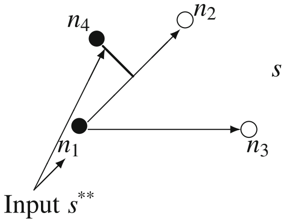
Accordingly, in order to determine the next state s_2 of the network, the state transition mapping F_{s^*} is applied again. The state transition will be such that the activity of n_1 in s_1 spreads to n_2 and n_3, which yields:



$$s_2 = F_{s^*}(s_1)$$

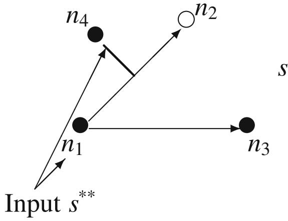
If the state transition mapping is applied again, then nothing is going to happen anymore (until the input to the network changes): hence, $s_3 = F_{s^*}(s_2) = s_2$. Connectionists regard such a *stable* or *equilibrium* state as a network's "answer" to the "question" posed by the input. So, s_2 – the state in which only n_1, n_2, n_3 fire – is the output that belongs to the input s^*. As we will also say, s_2 is an s^*-stable state.

What would happen if we applied a different input to the same initial state? Let s^{**} be the state in which both n_1 and n_4 fire, i.e., where the external input now causes these two nodes to become active. Then we have, by the same token as before:



$s_1' = F_{s^{**}}(s_0)$

But now the state transition will be such that the activity of n_4 in s_1 blocks the excitation of n_2 by n_1. In other words:



$s_2' = F_{s^{**}}(s_1')$

Once again, a stable state is reached after two computation cycles, and this time the output to the input state s^{**} is the state in which n_1, n_3, n_4 fire, i.e., s_2' is an s^{**}-stable state.

What we have said so far constitutes a typical description of (simplified) network processes in the language of the theory of dynamical systems. Out goal is now to complement this description by one according to which cognitive dynamical systems have beliefs, draw inferences, and so forth. So if x is a neural network, we want to say things like

- x believes that $\neg\varphi$
- x infers that $\varphi \vee \psi$ from φ

$\vdots$

where φ and ψ are *sentences*. Our task is thus to associate *states* of cognitive dynamical systems with *sentences* or *propositions*: the states of such dynamic systems ought to carry information that can be expressed linguistically.

Let us make this idea more precise. In order to do so, we first have to abstract from the overly simplified dynamical systems that were given by the qualitative neural networks sketched above. Indeed, we want to leave open at this point what our dynamical systems will be like – whether artificial neural networks or not – as long as they satisfy a few abstract requirements.

Here is what we will presuppose: we are dealing with a discrete dynamical systems with a set S of states. On S, a partial order[12] $\leqslant$ is defined, which we will interpret as an ordering of the amount of information that is carried by the states in question; so, $s \leqslant s'$ will be read as: s' carries at least as much information as s does. We will also assume that $\leqslant$ is "well-behaved" in so far as for every two states s and s' there is a uniquely determined state $sup(s, s')$ (i) that carries at least as much information as s, (ii) that carries at least as much information as s', and (iii) which is the state with the least amount of information among all those states for which (i) and (ii) hold. Formally, such a state $sup(s, s')$ is the *supremum* of s and s' with respect to the partial order $\leqslant$. Finally, an internal next-state function is defined for the dynamical system, such that this next-state function is like the state transition mapping described above except that – for the moment – we will disregard possible inputs to the system. Hence, in the examples above, an application of the corresponding next-state mapping would lead to the transmission of the activity of n_1 to n_3 once n_1 gets activated, but it will never lead to any activation of n_1 itself since n_1 can only be activated by external input.

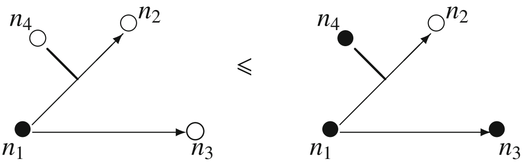
Summing up, we determine what is called an 'ordered discrete dynamical system' by Leitgeb [38]:

Definition 1 $\mathcal{S} = \langle S, ns, \leqslant \rangle$ is an ordered discrete dynamical system if and only if

1. S is a non-empty set (the set of states).
2. $ns : S \rightarrow S$ (the internal next-state function).
3. $\leqslant \subseteq S \times S$ is a partial order (the information ordering) on S, such that for all s, $s' \in S$ there is a supremum $sup(s, s') \in S$ with respect to $\leqslant$.

In the example networks above, we had $S = \{s \mid s : N \rightarrow \{0, 1\}\}$ with $N = \{n_1, n_2, n_3, n_4\}$ being the set of nodes. In order to define a suitable information ordering $\leqslant$ on S, we can, e.g., use the following idea: the more nodes are activated in a state, the more information the state carries. Thus we would have, e.g.:

[12] A partial order $\leqslant$ (on S) is a reflexive, antisymmetric, and transitive binary relation, i.e.: for all $s \in S$: $s \leqslant s$; for all $s, s' \in S$: if $s \leqslant s'$ and $s' \leqslant s$ then $s = s'$; for all $s_1, s_2, s_3 \in S$: if $s_1 \leqslant s_2$ and $s_2 \leqslant s_3$ then $s_1 \leqslant s_3$.



If $\leqslant$ is defined in this way, then $sup(s, s')$ turns out to be the union of the activation patterns that correspond to s and s'; in such a case one may also speak of $sup(s, s')$ as the "superposition of the states s and s'". The internal dynamics of the network is captured by the next-state mapping ns that is determined by the pattern of excitatory and inhibitory edges in the network.

Just as in the examples above, we are now ready to also consider an input, which is regarded to be represented by a state $s^* \in S$, and which is supposed to be held fixed for a sufficiently long duration of time. The state transition mapping F_{s^*} can then be defined by taking both the internal next-state mapping and the input s^* into account: the next state of the system is given by the superposition of s^* with the next internal state $ns(s)$, i.e.,

$$F_{s^*}(s) := sup(s^*, ns(s)).$$

The dynamics of our dynamical systems is thus determined by applying F_{s^*} iteratively to the initial state. Fixed points s_{stab} of F_{s^*}, i.e., where $F_{s^*}(s_{stab}) = s_{stab}$, are again regarded to be the "answers" the system gives to s^*; any such state s_{stab} is called s^*-stable (relative to the given ordered discrete dynamical system). Note that in general there may be *more than just one stable state* for the state transition mapping F_{s^*} that is determined by the input s^* (and by the given dynamical system), and there may also be *no stable state* at all for F_{s^*}: in the former case, there is more than just one "answer" to the input, in the latter case there is no "answer" at all. The different stable states may be reached by starting the computation in different initial states of the system.

Finally, we are ready to assign formulas to the states of ordered discrete dynamical system. These formulas are supposed to express the content of the information that is represented by these states. For this purpose, we fix a propositional language $\mathcal{L}$ which (i) includes finitely many propositional variables $p, q, r, \ldots$, and (ii) is closed under the application of the standard classical propositional connectives, i.e., $\neg$, $\wedge$, $\vee$, $\supset$, $\top$, $\bot$, where $\top$ is the *logical verum* (a tautology) and $\bot$ is the *logical falsum* (a contradiction). The formulas of $\mathcal{L}$ do not yet include any of the nonmonotonic conditional signs such as $\Rightarrow$ that we are interested in. The assignment of formulas to states is achieved by an interpretation mapping $\Im$. If φ is a formula in $\mathcal{L}$, then $\Im(\varphi)$ is the state that carries exactly the information that is expressed by φ, i.e., not less or more than what is expressed by φ. So we presuppose that for every formula in $\mathcal{L}$ there is a uniquely determined state the total information of which is expressed by that formula. If expressed in terms of belief, we can say that in the state $\Im(\varphi)$ *all the system believes is that* φ, i.e., the system only believes φ and all the propositions which are contained in φ from the viewpoint of the system (compare [42] on the modal logic of the 'all I know' operator). We will not demand that every state necessarily receives an interpretation but just that every formula in $\mathcal{L}$ will be the interpretation of some state. Furthermore, not just any assignment of states to formulas will do, but we will additionally assume certain postulates to be satisfied which will guarantee that $\Im$ is compatible with the information ordering that was imposed on the states of the system beforehand. An ordered discrete dynamical system together with such an interpretation mapping is called an 'interpreted ordered system' (cf. [38]). This is the definition stated in detail:

Definition 2 $\mathcal{S}_{\Im} = \langle S, ns, \leqslant, \Im \rangle$ is an interpreted ordered system if and only if

1. $\langle S, ns, \leqslant \rangle$ is an ordered discrete dynamical system.
2. $\Im : \mathcal{L} \to S$ (the interpretation mapping) is such that the following postulates are satisfied:
   (a) Let $\mathcal{TH}_{\Im} = \{\varphi \in \mathcal{L} \,|\text{for all } \psi \in \mathcal{L}\text{: } \Im(\varphi) \leqslant \Im(\psi)\}$:
   then it is assumed that for all $\varphi, \psi \in \mathcal{L}$: if $\mathcal{TH}_{\Im} \models \varphi \supset \psi$, then $\Im(\psi) \leqslant \Im(\varphi)$.
   (b) For all $\varphi, \psi \in \mathcal{L}$: $\Im(\varphi \wedge \psi) = sup(\Im(\varphi), \Im(\psi))$.
   (c) For every $\varphi \in \mathcal{L}$: there is an $\Im(\varphi)$-stable state.
   (d) There is an $\Im(\top)$-stable state s_{stab}, such that $\Im(\bot) \not\leqslant s_{stab}$.

$\mathcal{S}_{\Im}$ satisfies the uniqueness condition if and only if for every $\varphi \in \mathcal{L}$ there is precisely one $\Im(\varphi)$-stable state.

How can these postulates be justified? First of all, $\mathcal{TH}_{\Im}$ is the set of formulas that are the interpretations of states which carry less information than, or an equal amount of information as, *any* other state with an interpretation. Hence, if $\varphi \in \mathcal{TH}_{\Im}$, then the information expressed by φ is contained in every interpreted state of the system. If spelled out in terms of belief, we may say: φ is believed by the system in every state that has an interpretation. For the same reason, such a belief cannot be revised by the system – it is "built" into the interpreted ordered system independently of its current input or state, as long as the state it is in has an interpretation at all. In more traditional philosophical terms, we might say that every such formula is believed a priori by the system. So if a material conditional $\varphi \supset \psi$ follows logically from $\mathcal{TH}_{\Im}$, then – since (rational) belief is closed under logical deduction – $\varphi \supset \psi$ must also be (rationally) believed by the system in every interpreted state whatsoever; indeed we may think of such a conditional as a strict a priori conditional: it is a material conditional which is epistemically necessary in the sense of being entailed by $\mathcal{TH}_{\Im}$, hence, if $\Box$ expresses entailment by $\mathcal{TH}_{\Im}$, then for every conditional $\varphi \supset \psi$ that is derivable from $\mathcal{TH}_{\Im}$ it holds that $\Box(\varphi \supset \psi)$. But if this is so, then the system must regard the propositional information that is expressed by ψ to be included in the propositional information that is expressed by φ – from the viewpoint of the system, φ must express a stronger proposition than ψ. In this case, with respect to the information ordering of the system, the state that belongs to ψ should be "below" the state that is associated with φ, or at worst the two states should be equal in the information ordering. In other words, $\Im(\psi) \leqslant \Im(\varphi)$

ought to be the case. That is exactly what is expressed by postulate 2a. $\mathcal{TH}_{\Im}$ may be interpreted as the set of "hard laws" or "strict laws" represented by the interpreted system.

Postulate 2b is more easily explained and justified: the state that belongs to a conjunctive formula $\varphi \wedge \psi$ should be the supremum of the two states that are associated with the two conjuncts φ and ψ, just as the proposition expressed by a conjunctive sentence is the supremum of the propositions expressed by its two conjuncts in the partial order of logical entailment.

Postulate 2c makes sure that we are dealing with systems that have at least one "answer" – whether right or wrong – to every "question" posed to the system.

Postulate 2d only allows for interpreted ordered systems which do not end up believing a contradiction when they receive a trivial or empty information (i.e., $\top$) as an input.

Finally, we are in the position to define what it means for a *nonmonotonic* conditional to be satisfied by an interpreted ordered system. Consider an arbitrary conditional $\varphi \Rightarrow \psi$ where φ and ψ are members of our language $\mathcal{L}$ from above, and where $\Rightarrow$ is a new nonmonotonic conditional sign. Then we say that a system satisfies $\varphi \Rightarrow \psi$ if, and only if, whenever the state that is associated with φ is fed into the system as an input, i.e., whenever the input represents a total belief in φ, the system will eventually end up believing ψ in its "answer states", i.e., the state that is associated with ψ is contained in all the states that are stable with respect to this input. If we collect all such conditionals $\varphi \Rightarrow \psi$ satisfied by the system, then we get what we call the 'conditional theory' corresponding to the system. In formal terms:

Definition 3 Let $\mathcal{S}_{\Im} = \langle S, ns, \leqslant, \Im \rangle$ be an interpreted ordered system:

1. $\mathcal{S}_{\Im} \models \varphi \Rightarrow \psi$ if and only if for every $\Im(\varphi)$-stable state s_{stab}: $\Im(\psi) \leqslant s_{stab}$.
2. $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\Im}) = \left\{\varphi \Rightarrow \psi \,\middle|\, \mathcal{S}_{\Im} \models \varphi \Rightarrow \psi\right\}$
   (the conditional theory corresponding to $\mathcal{S}_{\Im}$).

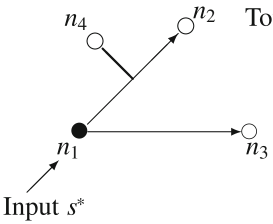
$\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\Im})$ may be interpreted as the set of "soft laws" or "normality laws" represented by the interpreted system. Leitgeb [40] gives an interpretation of the cognitive states that correspond to such conditionals in terms of so-called *conditional beliefs*, where conditional beliefs are to be distinguished conceptually from beliefs *in* conditionals.

Here is an example: consider again the simple qualitative network which we presented as a discrete ordered dynamical system above. In order to turn it into an *interpreted* ordered system, we have to equip it with an interpretation mapping $\Im$ that is defined on a propositional language $\mathcal{L}$. Let, e.g., $\mathcal{L}$ be determined by the set $\{b, f, w, p\}$ of propositional variables (for 'Tweety is a bird', 'Tweety is able to fly', 'Tweety has wings', 'Tweety is a penguin'). We choose the following interpretation mapping: let $\Im(b) = \{n_1\}$, $\Im(f) = \{n_1, n_2\}$, $\Im(w) = \{n_1, n_3\}$, $\Im(p) = \{n_1, n_4\}$, and $\Im(\neg\varphi) = 1 - \Im(\varphi)$, where the latter is to be understood in the way that whenever a node is active in $\Im(\varphi)$ then the same node is inactive in $\Im(\neg\varphi)$ and vice versa.[13]

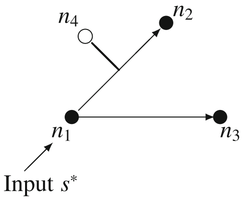
[13] So 1 here is actually the constant 1-function, i.e., the function that maps each node to the activation value 1.

One can show that there is one and only one interpretation that has these properties and which also satisfies the postulates in Definition 2. Note that we have assumed $\Im(\neg\varphi) = 1 - \Im(\varphi)$ just for convenience, as it becomes easier then to pin down an interpretation for our example. It is not *implied* at all by Definition 2 that the pattern of active nodes that is associated with a negation formula $\neg\varphi$ is actually identical to the complement of the pattern of active nodes that belongs to the formula φ; this is merely the way in which we have set up our example. One consequence of this choice of $\Im$ is that, e.g., the following material conditionals turn out to be members of $\mathcal{TH}_{\Im}$: $p \supset b$, $(p \wedge w) \supset b$, $\neg b \supset \neg p$, and so forth.

Reconsidering our example from above, the dynamics of the system which we studied back then now turns out to have the following symbolic counterparts:

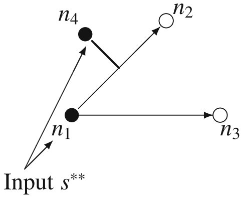


Total belief: b

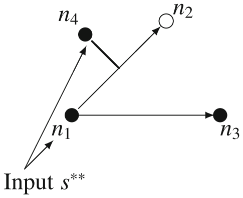


Believed: $b, f, w, (f \wedge w), (b \vee p), \ldots$
Not believed: $p, \neg b, \ldots$

Hence, $\mathcal{S}_{\Im} \models$
$b \Rightarrow f, b \Rightarrow w, b \Rightarrow (f \wedge w), (b \vee p) \Rightarrow f, \ldots$



Total belief: $b \wedge p$



Believed: $b, p, w, (b \wedge p \wedge w), \neg f, \ldots$
Not believed: $f, \neg b, (f \vee \neg b), \ldots$

Hence, $\mathcal{S}_{\Im} \models$
$(b \wedge p) \Rightarrow \neg f, p \Rightarrow b, p \Rightarrow \neg f, \ldots$

Obviously, there will be lots of “soft” if-then “laws” about birds and penguins which this interpreted ordered system will get wrong. After all, it would be very surprising indeed if a little network with just four nodes were able to represent all of the systematic relationships between birds and penguins and flying and having wings faithfully. But the example should suffice to give a clear picture of how the definitions above are to be applied.

So we find that in this case $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ contains, e.g., $b \Rightarrow f$, $b \Rightarrow w$, $b \Rightarrow (f \wedge w)$, $(b \vee p) \Rightarrow f$, $(b \wedge p) \Rightarrow \neg f$, $p \Rightarrow b$, $p \Rightarrow \neg f$ without containing, e.g., $b \Rightarrow p$, $(b \vee p) \Rightarrow p$, $(b \wedge p) \Rightarrow f$. In particular, we see that $b \Rightarrow f \in \mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ while $(b \wedge p) \Rightarrow f \notin \mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$.

What can be said in general terms about the conditional theories $\mathcal{TH}_{\Rightarrow}$ corresponding to interpreted dynamical systems? Here is the answer from the logical point of view:

Theorem 4 (Soundness of C)

Let $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$ be an interpreted ordered system:

Then $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ is sound with respect to the rules of the system C of nonmonotonic conditional logic (see [32] *for details on this system), i.e.:*

1. *For all $\varphi \in \mathcal{L}$: $\varphi \Rightarrow \varphi \in \mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ (Reflexivity)*
2. *$\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ is closed under the following rules: for φ, ψ, $\rho \in \mathcal{L}$,*
   $\frac{\mathcal{TH}_{\mathfrak{I}} \models \varphi \leftrightarrow \psi, \varphi \Rightarrow \rho}{\psi \Rightarrow \rho}$ *(Left Equivalence)*
   $\frac{\varphi \Rightarrow \psi, \mathcal{TH}_{\mathfrak{I}} \models \psi \supset \rho}{\varphi \supset \rho}$ *(Right Weakening)*
   $\frac{\varphi \Rightarrow \psi, \varphi \wedge \psi \Rightarrow \rho}{\varphi \Rightarrow \rho}$ *(Cautious Cut)*
3. *If $\mathcal{S}_{\mathfrak{I}}$ satisfies the uniqueness condition (remember Definition 2), then $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ is also closed under*
   $\frac{\varphi \Rightarrow \psi, \varphi \Rightarrow \rho}{\varphi \wedge \psi \Rightarrow \rho}$*(Cautious Monotonicity)*
4. *$\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$ is consistent, i.e., $\top \Rightarrow \bot \notin \mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}})$.*

So given the uniqueness assumption – an interpreted orderered system has a unique answer to each interpreted input – the class of conditionals it satisfies is closed under a well-known and important system of nonmonotonic conditional logic, namely the system C of *cumulative reasoning*, which is given by the rules listed above. Note that monotonicity, or strengthening of the antecedent, is *not* a valid rule for interpreted systems: as our example from above has shown, there may be formulas φ, ψ, ρ in $\mathcal{L}$, such that the conditional $\varphi \Rightarrow \psi$ is satisfied by a system but $\varphi \wedge \rho \Rightarrow \psi$ is not.

One can also show a corresponding completeness theorem for the system C with respect to this interpreted ordered systems semantics for $\Rightarrow$:

Theorem 5 (Completeness of C)

Let $\mathcal{TH}_{\Rightarrow}$ be a consistent theory of conditionals closed under the rules of C while extending a given classical theory $\mathcal{TH}$ as expressed by the Left Equivalence and the Right Weakening rules:

It follows that there is an interpreted ordered system $\mathcal{S}_{\mathfrak{I}} = \langle S, ns, \leqslant, \mathfrak{I} \rangle$, such that $\mathcal{TH}_{\Rightarrow}(\mathcal{S}_{\mathfrak{I}}) = \mathcal{TH}_{\Rightarrow}$, $\mathcal{TH}_{\mathfrak{I}} \supseteq \mathcal{TH}$, and $\mathcal{S}_{\mathfrak{I}}$ satisfies the uniqueness condition.

This means that whatever conditional theory you might be interested in, as long as it is closed under the rules of the system C, it is possible to find an interpreted ordered system which satisfies precisely the conditionals contained in that theory. These results can be found in Leitgeb [37].

It is also possible to extend these results into various directions. In particular, some interpreted ordered systems can be shown to have the property that each of their states s may be decomposed into a set of substates s_i which can be ordered in a way such that the dynamics for each substate s_i is determined by the dynamics for the substates $s_1, s_2, \ldots, s_{i-1}$ at the previous point of time. Such systems are called 'hierarchical' in Leitgeb [38]. We will not go into any details, but one can prove further soundness and completeness theorems for such *hierarchical* interpreted systems and the system CL = C + Loop of nonmonotonic conditional logic, where Loop is the following rule:

$$\frac{\varphi_0 \Rightarrow \varphi_1, \varphi_1 \Rightarrow \varphi_2, \ldots, \varphi_{j-1} \Rightarrow \varphi_j, \varphi_j \Rightarrow \varphi_0}{\varphi_0 \Rightarrow \varphi_j}\text{(Loop)}$$

Note that Loop is a weakened version of transitivity, whereas standard transitivity is *not* valid, just as the rule of cautious monotonicity above is a weakened version of monotonicity without standard monotonicity being valid anymore. (Consult [32] for more information on CL.)

In Leitgeb [36, 37], further soundness and completeness theorems can be found for more restricted classes of interpreted dynamical systems and even stronger logical systems for nonmonotonic conditionals. E.g., the important system P of so-called *preferential reasoning*, where P results from adding the rule

$$\frac{\varphi \Rightarrow \rho, \psi \Rightarrow \rho}{(\varphi \vee \psi) \Rightarrow \rho}\text{(Or)}$$

to the system CL, is sound and complete with respect to another particular class of interpreted dynamical systems. P coincides with Adams' [1] logical system for indicative conditionals as well as with the "flat" fragment of Lewis' [43] logic for subjunctive conditionals. ('Flat' means: iterations of subjunctive conditionals and other compositional constructions on their basis are excluded.) Moreover, various semantical systems for nonmonotonic reasoning have been found to "converge" on system P as their logical calculus.

As one can show, if artificial neural networks with weights are extended by an information ordering as well as an interpretation mapping along the lines explained above, then they turn out to be special cases of interpreted ordered systems. Furthermore, if the underlying artificial neural network consists of layers of nodes, such that the layers are arranged hierarchically, and all connections between nodes reach from one layer to the next one, then the interpreted ordered system is indeed a hierarchical one.

In more formal detail: $\langle U, W, A, O, NET, ex\rangle$ is an artificial neural network if and only if

1. U is a finite and nonempty set of nodes.
2. $W : U \times U \to \mathbb{R}$ assigns a weight to each edge between nodes.
3. A maps each node $u \in U$ to an activation mapping $A_u : \mathbb{R}^3 \to \mathbb{R}$ such that the activation state $a_u(t+1)$ of u at time $t+1$ depends on the previous activation state $a_u(t)$ of u, the current net input $net_u(t+1)$ of u, and the external input $ex(u)$ fed into u, i.e. $a_u(t+1) = A_u(a_u(t), net_u(t+1), ex(u))$.
4. O maps each node $u \in U$ to an output mapping $O_u : \mathbb{R} \to \mathbb{R}$ such that the output state $o_u(t+1)$ of u at time $t+1$ is solely dependent on the activation state $a_u(t+1)$ of u, i.e. $o_u(t+1) = O_u(a_u(t+1))$.
5. NET maps every node $u \in U$ to a net input (or propagation) mapping $NET_u : (\mathbb{R} \times \mathbb{R})^U \to \mathbb{R}$ such that the net input $net_u(t+1)$ of u at time $t+1$ depends on the weights of the edges leading from nodes u' to u, and on the previous output states of the nodes u', i.e. $net_u(t+1) = NET_u(\lambda u'.\big\langle W(u', u), o_{u'}(t)\big\rangle)$.[14]
6. $ex : U \to \mathbb{R}$ is the external input function.

We can view such networks as ordered dynamical systems when we define:

1. $S = \{s \mid s : U \to \mathbb{R}\}$.
2. $ns : S \to S$ with $ns(s)(u) := A_u(s(u), NET_u(\lambda u'.\big\langle W(u', u), O_{u'}(s(u'))\big\rangle), 0)$
   (so, in the definition of the internal next-state function, $ex(u)$ is set to 0).
3. $\leqslant \subseteq S \times S$ with $s \leqslant s'$ if and only if for all $u \in U$: $s(u) \leqslant s'(u)$.
   (Thus, $sup(s, s')$ is simply $max(s, s')$.)

$\langle S, ns, \leqslant\rangle$ is an ordered discrete dynamical system, such that $F_{s^*}(s) = sup(s^*, ns(s)) = max(s^*, ns(s))$ which entails that $F_{s^*}(s)(u) = max(s^*(u), ns(s)(u)) = max(s^*(u), A_u(s(u), NET_u(\lambda u'.\big\langle W(u', u), O_{u'}(s(u'))\big\rangle), 0))$, which corresponds to the assumption that the external input to a network interacts with the current activation state of the network by taking the maximum of both. Given this assumption, the dynamics of artificial neural networks and the dynamics of the corresponding ordered dynamical systems coincide. If the network is layered, then the corresponding ordered system is hierarchical. Stable states are regarded as the relevant "answer" states just as it is the case in the standard treatment of neural networks. If such networks are equipped with a corresponding interpretation

[14] $\lambda u'.\big\langle W(u', u), o_{u'}(t)\big\rangle$ is the function that maps u' to the pair $\big\langle W(u', u), o_{u'}(t)\big\rangle$.

mapping $\Im$ as defined above, they satisfy conditional theories which are closed under the rules of well-established systems of logic for nonmonotonic conditionals.

Furthermore, on the level of representation or interpretation we have:

- In interpreted ordered systems, *propositional formulas* are represented as total states s of the system; in particular, in interpreted neural networks, propositional formulas are represented as patterns of activity distributed over the nodes of the network.
- In interpreted ordered systems, *nonmonotonic conditionals* are represented through the overall dynamics of the system; in particular, in interpreted neural networks, nonmonotonic conditionals are represented by means of the network topology and the manner in which weights are distributed over the connections of the network. It is not single edges that correspond to conditionals, but the conditional theory that belongs to an interpreted network is a set of soft constraints that is represented by the network as a whole.

Thus, in contrast with the old McCulloch-Pitts idea, the representation of formulas in interpreted dynamical systems is distributed, as suggested by connectionists. At the same time, the set of conditionals satisfied by an interpreted dynamical system is closed under the rules of systems of nonmonotonic conditional logic that were introduced, and which have been studied intensively, by researchers in the tradition of the symbolic computation paradigm of cognitive science. Subsymbolic inference may be fundamentally nonmonotonic, as claimed by Smolensky (reconsider Sect. 7.2), but that does not mean that it could not be formalized in logical terms – it only means that the formalization has to be given in terms of systems of nonmonotonic reasoning or conditional logic.

The dynamical systems paradigm and the symbolic computation paradigm may thus be taken to yield complementary perspectives on the one and the same cognitive system. The precise meaning of this 'complementarity' is given by soundness and completeness theorems. Although these results only apply to highly idealized imitations of actual structures in the brain, the possibility of having such correspondences between symbolic and dynamic descriptions at all should be of great interest to philosophers of mind. Moreover, since nonmonotonic conditionals have been shown to have interpretations in terms of preference or similarity orderings of possible worlds or in terms of conditional probability measures (recall Sect. 7.3), the nonmonotonic conditionals that are satisfied by interpreted dynamical systems may be taken to represent aspects of some of these semantic structures. In this way, neural networks – artificial ones, but maybe even biological ones – may be understood as representing orderings of possible worlds or conditional probability measures, accordingly. This might pave the way for new interpretations and explanations of cognition done by neural networks, which should be relevant to cognitive scientists. Finally, as follows from the results above, conditionals in natural language, normality conditionals used by computer scientists, and the conditionals by which one may describe the dynamics of neural networks all seem to converge on more or less the same logic. This constitutes tentative evidence for two conclusions: first, the correspondence with normality conditionals in computer science indicates why conditionals in natural language might have the logical properties that they have – because we, much as the computer systems in artificial intelligence, need to be able to cope with exceptions. Secondly, the neural network semantics above suggests how we, natural language speakers, are capable of determining whether or not a conditional is acceptable to us – by feeding the information that is conveyed by its antecedent into a neural network that is run offline and the stable states of which are checked for whether they contain the information conveyed by the consequent of the conditional. Both of these tentative conclusions should be of obvious relevance to philosophers of language who are interested in conditionals.

7.5 Some Open Questions

Here is an (incomplete) to-do-list in this area of research:

Extending soundness/completeness results: How can all of the logical systems discussed by Kraus et al. [32], Lehmann and Magidor [34], and beyond be characterized in terms of connectionistically plausible and elegant constraints on interpreted dynamical systems? So far, there only seem to be partial answers to this question, sometimes relying on very restricted classes of dynamic systems. Which logical systems do we get if we drop the uniqueness assumption (see Definition 2)? How can full-fledged systems of conditional logic for subjunctive conditionals, for which nestings of conditionals and the application of propositional connectives to conditionals are well-defined, be represented by means of interpreted dynamical systems? The results achieved up to this point seem more suitable for indicative conditionals for which the meaningfulness of nesting and the application of propositional connectives are less plausible.

Characterizing learning in neural networks by logical rules: As we have seen, state transitions in a fixed (possibly, trained) neural network can be described by means of conditionals. However, it is as yet unknown how learning processes in networks – by which the weights in a network change under the influence of a learning algorithm and training data – can be represented by logical rules. Learning schemes such as Hebbian learning or backpropagation (by which the weights of connections between co-active nodes are increased) might translate into particular systems of inductive logic in which inferences can be drawn from factual training data and conditionals to learned conditionals. In order to facilitate this study, computer implementations of interpreted networks and their learning algorithms will be crucial.

Applying the theory to open problems in uncertain reasoning: The results achieved by the previous tasks are expected to feed back on open problems in uncertain reasoning. E.g.: The standard theory of belief revision was created as a theory for the "one-shot" revision of beliefs by a single piece of evidence.[15] It is well-known that belief revision and (preferential) nonmonotonic reasoning are more or less intertranslatable. Attempts of extending the theory of belief revision to iterated occurrences of evidence led to a multitude of suggestions lacking clear philosophical interpretation. By means of the results achieved in this area, it might be possible to understand evidence-induced changes of networks as iterated belief revisions. We hypothesise that different schemes of iterated revision correspond to, and can be understood as, different learning algorithms for neural networks.

Applying the theory in philosophy of science: In philosophy of science, it was realized early on that new empirical evidence can have the effect that previous hypotheses must be withdrawn, as a scientists might learn that what she had regarded likely is actually not. As Flach [20] argues, the same logics that govern valid commonsense inferences can be interpreted as logics for scientific induction, i.e., for data constituting incomplete und uncertain evidence for empirical hypotheses. Schurz [52] demonstrates that scientific laws are subject to normality or ceteris paribus restrictions that obey the logic of nonmonotonic reasoning. At the same time, the study of neural networks is expected to transform our philosophical understanding of science: Churchland [14] presents networks as models of scientific theories and regards prototype representations in networks as a system's explanatory understanding of its inputs. Bechtel [6] explains scientific model building in terms of the satisfaction of soft constraints represented in networks. Bird [8] observes: "The time is ripe for a reassessment of Kuhn's earlier work in the light of connectionist and neural-net research". Is it possible to throw some new light on these insights from the philosophy of science on the basis of new findings on logical accounts of neural networks and learning?

References

1. Adams, E. (1975). *The logic of conditionals: An application of probability to deductive logic* (Synthese library, Vol. 86). Dordrecht: Reidel.
2. Adams, E. (1998). *A primer of probability logic* (CSLI lecture notes). Stanford: Center for the study of language and information.
3. Bader, S., & Hitzler, P. (2005). Dimensions of neural-symbolic integration – a structured survey. In S. N. Artemov, H. Barringer, A. S. d'Avila Garcez, L. C. Lamb, & J. Woods (Eds.), *We will show them! Essays in honour of Dov Gabbay* (Federation for computational logic, pp. 167–194). London: College Publications, Int.
4. Balkenius, C., & Gärdenfors, P. (1991). Nonmonotonic inferences in neural networks. In J. A. Allen, R. Fikes, & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning* (pp. 32–9). San Mateo: Morgan Kaufmann.
5. Bechtel, W. (1994). Natural deduction in connectionist systems. *Synthese, 101*(3), 433–463.
6. Bechtel, W. (1996). What should a connectionist philosophy of science look like? In R. M. McCauley (Ed.), *The Churchlands and their critics* (pp. 121–44). Massachusetts: Basil Blackwell.

[15] Gärdenfors [22] is the classic reference, and Hansson [25] is a nice textbook on belief revision.

7. Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.
8. Bird, A. (2002). What is in a paradigm? *Richmond Journal of Philosophy, 1*(ii), 11–20.
9. Blutner, R. (2004). Nonmonotonic inferences and neural networks. *Synthese, 142*, 143–74.
10. Boutsinas, B., & Vrahatis, M. N. (2001). Artificial nonmonotonic neural networks. *Artificial Intelligence, 132*, 1–38.
11. Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning. An overview* (CSLI lecture notes, Vol. 73). Stanford: CSLI Publications.
12. Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science, 2*(1 & 2), 53–62.
13. Chen, C. -H., & Honavar, V. (1999). A neural-network architecture for syntax analysis. *IEEE Transactions on Neural Networks, 10*(1), 94–114.
14. Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. London: MIT Press.
15. d'Avila Garcez, A. S., Broda, K., & Gabbay D. M. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence, 125*, 153–205.
16. d'Avila Garcez, A. S., Broda, K. B., & Gabbay D. M. (2002). *Neural-symbolic learning systems*. London: Springer.
17. d'Avila Garcez, A. S., Lamb, L. C., & Gabbay, D. M. (2009). *Neural-symbolic cognitive reasoning*. Berlin: Springer.
18. d'Avila Garcez, A. S., Besold, T. R., de Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K. -U., Lamb, L. C., Miikkulainen, R., & Silver, D. L. (2015). *Neural-symbolic learning and reasoning: Contributions and challenges*. Paper presented at the 2015 AAAI spring symposium series, Stanford.
19. Dietz, E. -A., Hölldobler, S., Palacios, L. (2015). *A connectionist network for skeptical abduction*. Paper presented at NeSy 2015, Buenos Aires.
20. Flach, P. A. (2000). Logical characterizations of inductive learning. In D. M. Gabbay & R. Kruse (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 4, pp. 155–96). Dordrecht: Kluwer.
21. Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.
22. Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, MA: The MIT Press.
23. Gärdenfors, P. (1994). How logic emerges from the dynamics of information. In J. Van Eijck & A. Visser (Eds.), *Logic and information flow* (pp. 49–77). Cambridge: The MIT Press.
24. Ginsberg, M. L. (Ed.). (1987). *Readings in nonmonotonic reasoning* (pp. 1–23). Los Altos: Morgan Kaufmann.
25. Hansson, S. O. (1999). *A textbook of belief dynamics*. Dordrecht: Kluwer.
26. Hitzler, P., Hölldobler, S., & Seda, A. K. (2004). Logic programs and connectionist networks. *Journal of Applied Logic, 2*(3), 245–272.
27. Hölldobler, S. (1993). *Automated inferencing and connectionist models* (Post-doctoral thesis).
28. Hölldobler, S., & Kalinke, Y. (1994). Towards a massively parallel computational model for logic programming. In *Proceedings ECAI94 Workshop on Combining Symbolic and Connectionist Processing* (pp. 68–77). ECCAI.
29. Hölldobler, S. (2009). Cognitive science, computational logic and connectionism. In M. Adriani, et al. (Eds.), *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 1–6).
30. Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology*. Cambridge: The MIT Press.
31. Kleene, S. C. (1956). Representation of events in nerve nets and finite automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 3–42). Princeton: Princeton University Press.
32. Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence, 44*, 167–207.
33. Legendre, G., Miyata, Y., & Smolensky, P. (1994). *Principles for an integrated connectionist/symbolic theory of higher cognition*. Hillsdale: L. Erlbaum.

34. Lehmann, D., & Magidor, M. (1992). What does a conditional knowledge base entail? *Artificial Intelligence, 55*, 1–60.
35. Leitgeb, H. (2001). Nonmonotonic reasoning by inhibition nets. *Artificial Intelligence, 128*(1–2), 161–201.
36. Leitgeb, H. (2003). Nonmonotonic reasoning by inhibition nets II. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 11*(suppl., issue 2), 105–35.
37. Leitgeb, H. (2004). *Inference on the low level. An investigation into deduction, nonmonotonic reasoning, and the philosophy of cognition* (Applied logic series). Dordrecht: Kluwer/Springer.
38. Leitgeb, H. (2005). Interpreted dynamical systems and qualitative laws: From inhibition networks to evolutionary systems. *Synthese, 146*, 189–202.
39. Leitgeb, H. (2005). Reseaux de neurones capables de raisonner. *Dossier Pour la Science* (Special issue of the French edition of the *Scientific American*) October/December, 97–101.
40. Leitgeb, H. (2007). Beliefs in conditionals vs. conditional beliefs. *Topoi, 26*(1), 115–32.
41. Leitgeb, H. (2007) Neural network models of conditionals: An introduction. In X. Arrazola, J. M. Larrazabal, et al. (Eds.), *LogKCA-07, Proceedings of the First ILCLI International Workshop on Logic and Philosophy of Knowledge, Communication and Action* (pp. 191–223). Bilbao: University of the Basque Country Press.
42. Levesque, H. (1990). All I know: A study in autoepistemic logic. *Artificial Intelligence, 42*, 263–309.
43. Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
44. Makinson, D. (1994). 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* (Vol. 3, pp. 35–110.). Oxford: Clarendon Press.
45. McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–33; Reprinted in: W. S. McCulloch, *Embodiments of mind*. Cambridge, MA: The MIT Press (1965).
46. Ortner, R., & H. Leitgeb (2011). Mechanizing induction. In D. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic. Volume 10: Inductive logic* (pp. 719–772). Oxford: Elsevier.
47. Peters, S., & Westerstahl, D. (2006). *Quantifiers in language and logic*. Oxford: Oxford University Press.
48. Pinkas, G. (1991). Symmetric neural networks and logic satisfiability. *Neural Computation, 3*, 282–291.
49. Pinkas, G. (1995). Reasoning, nonmonotonicity and learning in connectionist networks that capture propositional knowledge. *Artificial Intelligence, 77*, 203–247.
50. Rojas, R. (1996). *Neural networks – a systematic introduction*. Berlin: Springer.
51. Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing* (Vols. 1 and 2). Cambridge: The MIT Press.
52. Schurz, G. (2002). Ceteris paribus laws: Classification and deconstruction. *Erkenntnis, 57*(3), 351–72.
53. Schurz, G., & Leitgeb, H. (Eds.). (2005). Special volume on "Non-monotonic and uncertain reasoning in the focus of paradigms of cognition. *Synthese, 146*, 1–2.
54. Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1–23.
55. Smolensky, P. (1990). Tensor-product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence, 46*, 159–216.
56. Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. Cambridge, MA: The MIT Press.
57. Stenning, K., & Lambalgen, M. van (2008). Human reasoning and cognitive science. Cambridge: The MIT Press.
58. van Benthem, J. (1984). Foundations of conditional logic. *Journal of Philosophical Logic, 13*(3), 303–49.
59. Van Gelder, T. J. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21*, 615–65.

60. Van Gelder, T. J. (1999). Distributed versus local representation. In R. Wilson & F. Keil (Eds.), *The MIT encyclopedia of cognitive sciences* (pp. 236–38). Cambridge: The MIT Press.
61. Wheeler, G. (2017). Machine epistemology and big data. In L. McIntyre & A. Rosenberg (Eds.), *The Routledge companion to philosophy of social science* (pp. 321–329). New York: Routledge.