# Incipiens

Zeitschrift für Erstpublikationen
aus der Philosophie und ihrer Geschichte

Ausgabe 2 1/2014



ISSN 2198-6843

Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, LMU München

# Incipiens

## Zeitschrift für Erstpublikationen aus der Philosophie und ihrer Geschichte

Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft
Ludwig Maximilians Universität München

Ausgabe 2
1/2014



**Issn:** 2198-6843

Veröffentlicht unter www.incipiens.de.

# Inhalt

# Logical Symbolicism vs. Dynamicist Connectionism: Is there a Difference in Computational Power?

Mario Günther

*I aim to resolve the philosophical controversy between logical symbolicism and dynamicist connectionism in cognitive science. A prominent philosopher, Timothy van Gelder, holds the view that the fundamental difference between symbolic systems and dynamical systems can be rendered explicit by looking at their computational powers. We argue – against van Gelder's view – for the claim that there is no difference* in principle *between the two types of systems concerning their representational and computational powers. In the argumentation we invoke Hannes Leitgeb's sequence of representation theorems that shows a profound equivalence between non-monotonic logics and dynamicist neural networks.*

*Ich ziele darauf ab, die philosophische Kontroverse zwischen logischen Symbolizisten und dynamischen Konnektionisten in der Kognitionswissenschaft zu beheben. Als einer der prominenten Philosophen des Dynamizismus vertritt Timothy van Gelder die Ansicht, dass der grundsätzliche Unterschied zwischen symbolischen Systemen und dynamischen Systemen explizit gemacht werden kann, indem man deren komputationelle Mächtigkeiten (oder äquivalent Leistungsfähigkeiten) kontrastiert. Ich argumentiere – entgegen van Gelders Ansicht – dafür, dass es keinen Unterschied* in principio *gibt zwischen den beiden Arten von Systemen hinsichtlich deren komputationellen Mächtigkeiten. In der Argumentation rekurriere ich auf Hannes Leitgebs Sequenz von Repräsentationstheoremen, die eine profunde Äquivalenz zwischen nicht-monotonen Logiken und dynamischen neuronalen Netzen beweist.*

## I Van Gelder's Claims

Apparently there is a philosophical cleavage in cognitive science. Some philosophers and/or cognitive scientists adhere to symbolicism, some adhere to connectionism. There have been fierce debates about the 'true nature' of cognition. We aim towards a resolution of this debate. The paper is supposed to be a first step. Philosopher Timothy van Gelder adds fuel to the fire concerning this ongoing debate. He holds both that only dynamicism can model cognition appropriately, and that dynamicism, while

compatible with connectionism, is not compatible with symbolicism. In the debate he defends his dynamical hypothesis as a substantive alternative to what he calls the 'computational' hypothesis; we will call it symbolic hypothesis for the sake of clarity (dynamical systems can also execute computations). Defending the dynamical hypothesis against the symbolic hypothesis means for van Gelder at first that there is a fundamental opposition between the two:

> Dynamical Hypothesis (DH): Cognitive agents instantiate dynamical systems.
>
> Symbolic Hypothesis (SH): Cognitive agents instantiate digital computers.

Of course, human beings are dynamical systems, and not digital computers.[1] However, van Gelder's claim is stronger: human cognition is best described by dynamical systems, and the description models *exactly the same* as the actual cognitive process.[2] Digital computers in van Gelder's sense are systems which operate by manipulating symbolic representations. He demands that this manipulation of representations exhibits a systematic interpretation such that the system's operations 'make sense'. This demand, however, is qualified by noting that "an interpretation in the current sense may not be enough to guarantee that the system has 'meaning' in some stronger sense, (and hence, perhaps, 'mind')".[3]

On the one hand van Gelder cites Newell and Simon's *physical symbol system hypothesis* (PSSH) as a famous presentation of the SH; on the other

1 Already one of the fathers of symbolicism, Alan M. Turing, was aware of that "[e]verything really moves continuously. But there are many kinds of machine which can profitably be thought of as being discrete state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them." Turing (1959): 439. The question is whether we can describe human cognition as one of Turing's discrete state machines.

2 In this paper, we do not consider the claim of a 1:1 mapping or 'identity instantiation' between reality and model (or description), but it seems metaphysically biased and concerning the complexity of human cognition (or the brain) to some extent odd. Abstractions are, after all, important in science, especially when the investigated entity is in reality very complex. For a survey about this issue see van Leeuwen (2005): 299-301.

3 Van Gelder (1998): 620, footnote 9.

hand he claims that connectionism can model cognition as the behaviour of dynamical systems. Moreover, he states that "one class of dynamical systems, recurrent neural networks [...] are *more* powerful – can compute a wider class of functions – than Turing Machines.[19]"[4] He adds in his footnote 19 that "dynamical systems can have 'super-Turing' capacities".[5]

We see that van Gelder tries to strengthen the dynamicist perspective on cognition against the symbolicist paradigm. To this end he proposes the DH as working hypothesis for cognitive scientists that, according to him, competes with and is fundamentally opposed to the symbolicist paradigm. In order to defend his hypothesis he roughly argues – based on his assumptions (i)-(v) – as follows:

> (i) There is a fundamental difference between symbolic systems and dynamical systems.
>
> (ii) If there is a fundamental difference between symbolic systems and dynamical systems, then there is a fundamental difference concerning their computational and representational powers.
>
> (iii) If symbolic and dynamical systems differ concerning the computational and representational powers, then dynamical systems have a greater computational power than symbolic systems.
>
> (iv) If dynamical systems have a greater computational power than symbolic systems, then symbolic systems are not sufficient to describe dynamical systems.
>
> (v) The (human) brain instantiates a dynamical system that describes (human) cognition.
>
> Therefore, (vi) symbolic systems are not sufficient to describe or model (human) cognition, whereas dynamical systems are sufficient.

4 Van Gelder (1998): 632.

5 Van Gelder (1998): 632, footnote 19.

> And so, (vii) dynamical systems are the appropriate tool to model human cognition, and thus cognitive scientists should prefer as working hypothesis the DH over the SH.

The plan is to investigate whether the mentioned claims (i), (ii), (iii), and (iv) are true, and thus the conclusion (vi) holds. In section II, we outline the symbolicist perspective van Gelder criticises. The picture of symbolicism he suggests by citing the PSSH is too strong. For an adequate view we replace the PSSH presented by van Gelder as central tenet of symbolicism by the psychological version of the Church-Turing thesis. In section III, we consider the claim that human beings can compute more (functions) than digital computers, which is incompatible with the Church-Turing thesis. However, it turns out that although theoretically possible, super-Turing machines are not physically plausible. We argue for the physical implausibility of super-Turing neural networks with which van Gelder underpins claim (iii).

In section IV, we introduce the idea behind the connectionist perspective on cognition, the neuron hypothesis and a dynamical alternative with respect to the alleged difference between symbolicism and connectionism. We end the section with the remark that logic can help to bridge the gap not only between algorithmic and implementation levels, but also to find structural similarities across different instantiations of cognition.

In section V, we present one result of such a logical search for similarities, namely McCulloch and Pitts representation theorem between a particular logical theory and a class of neural networks satisfying particular constraints. Their result raises the question whether there is some difference in principle between a symbolicist and connectionist analysis of cognition. In section VI, we address the controversy between symbolicists and connectionists about mental representations, in particular with respect to the 'symbolic-subsymbolic' distinction, in order to refine the question from the previous section. In section VII, we argue *via* representation theorems given by Hannes Leitgeb against van Gelder's claims (ii) and (iii). The representation theorems for non-monotonic logics in neural networks suggest that there is no substantive difference in the computational power of symbolic systems and that of dynamical neural networks.

## II The Symbolicist Perspective

As an opponent of the SH, van Gelder contests the claim that cognitive agents are (in some sense equivalent to) digital computers. In this section we outline the symbolicist perspective on cognition, which van Gelder characterises as the 'computationalist' perspective and inappropriate. We start with the Church-Turing thesis. The following serves to clarify what he means with 'digital computers', why the PSSH is a version of the SH, and why Turing machines play an important role in the symbolicist approach to cognition. Moreover, we try to draw a line between the necessary commitments of all members of the symbolicist approach and stronger commitments shared only by a few.

Human beings are physically realised cognizers. As such they are confronted with cognitive tasks or problems. One goal of Cognitive Science is to describe the cognitive processes involved in problem solving. The working hypothesis most cognitive scientists share – be they symbolicists, connectionists, or dynamicists – is that cognition can be considered as information processing, and thus as computation.

The symbolicist approach treats cognition as information processing in a symbolic system. This perspective claims that intelligent behaviour as we observe it in human beings can be seen as analogous to digital computers. But what does 'computation' mean here? This question leads us to the core of the symbolicist approach, a sophisticated conjecture of Church and Turing:

> Church-Turing Thesis (CTT): All computation, in the intuitive sense of a mechanical procedure for solving problems, is formally equivalent to computation by a Turing machine.

Almost all philosophers and cognitive scientists share the view that the (universal) Turing machine can be seen as the theoretical concept of computation. Van Gelder understands the SH in the following sense: cognitive agents instantiate realised Turing machines. The CTT cannot be formally proved. However, all attempts so far to explicate the intuitive notion of computability have turned out to define exactly the same class of problems. For instance, definitions *via* abstract machines (random access machines, cellular automata, genetic algorithms), formal systems (Church's λ-calculus, Post's rewriting systems), and particular classes of function (recursive functions) are all formally equivalent to the definition of Turing machine computability. These results provide support for the

Church-Turing thesis that all computation is equivalent to Turing computation.[6]

If we accept the Church-Turing thesis and we accept that the human mind cognises (i.e. executes computations), then we should also accept the psychological counterpart of the CTT:

> Psychological Version of the CTT: The human mind can only solve problems that are computable by a Turing machine.

Newell and Simon went one step further when they formulated their conjecture:

> Physical Symbol System Hypothesis (PSSH): "A physical symbol system has the necessary and sufficient means for general intelligent action."[7]

Presupposing that the human mind acts intelligently, the PSSH implies that the human mind is a symbolic system in some sense (for example a Turing machine) realised in the brain (similar to the implementation of a Turing machine in the physical appearance of digital computers). Moreover, the PSSH even implies that physical symbolic systems, i.e. realised machines, can be intelligent cognitive agents (similar to human beings).

One needs not necessarily embrace the PSSH in order to adhere to the symbolicist approach. In contrast, one needs to agree on the psychological version of the CTT, which states – in other words – that cognitive tasks can be described by computable functions (which express a 'Turing table', i.e. the programme or algorithm governing a Turing machine). In this form the PSSH implies the psychological version of the CTT, since the PSSH states that the intelligent performance while solving a cognitive task can be described by symbol manipulation, and functions fall within the domain of symbol manipulation. The converse implication does obviously not hold, and the psychological version of the CTT does not say anything about implementation issues.

When van Gelder cites the PSSH in order to characterise the symbolicist approach, he takes one of the strongest claims ever raised by classical computationalists. As mentioned before, the adoption of the PSSH is not necessary in order to call oneself a symbolicist. It is necessary though to hold that – from an abstract perspective – a cognitive task can be described by an information-processing task:

6 Cf. Cooper (2003).

7 Newell/Simon (1976): 116.

> Given some input (e.g. a visual stimulus, a state of the world, a sensation of pain), produce an appropriate output (e.g. perform an action, draw a conclusion, utter a response).

Generally, cognitive tasks can be understood as functions from inputs to outputs, and (the psychological version of) the CTT states that the only realistic candidates for information processing tasks (performed by the human mind) can be expressed by *computable* functions. To be clear, a cognitive task can be described as the encoding of, for instance, a visual stimulus into input information, information processing (the respective function that is computed), and the decoding of output information, for instance, the performance of an action. Hereby, encoding and decoding are special forms of information processing (or symbol manipulation).

Not everyone accepts the psychological version of the CTT. In particular, some critics – including van Gelder – have argued that cognitive agents or systems can compute more than Turing machines (cf. claim (iii)). In the next section we consider one particular strand of argumentation for the claim that dynamical systems can have super-Turing capacities in contrast to symbolic systems. As it turns out, these super-Turing capacities seem to be only theoretical concepts.

## III Arguments against the Psychological Church-Turing Thesis

There are a bunch of arguments for the claim that some systems have super-Turing capacities. According to van Gelder at least some dynamical systems have such capacities. One line of argumentation for this claim that we consider here is motivated by Kurt Gödel's famous theorems. The foundation for the argumentation is that Gödel's incompleteness results somehow demonstrate that the human mind cannot have an algorithmic nature. Lucas, for example, claimed in his "Minds, machines and Gödel": "Gödel's Theorem seems to me to prove Mechanism is false, that is, that minds cannot be explained [which implies here that they cannot be described] as machines."[8] Lucas' claim obviously contradicts the PSSH and the psychological version of the CTT, according to which the processes in the human mind can be described as a symbolic system or Turing machine. Lucas provides the following argument for his claim: A digital computer

8 Lucas (1961): 112.

behaves according to a Turing table or programme; hence we can view such a digital computer as a formal system. Applying Gödel's theorem to this system, however, we obtain a true sentence which is unprovable in the system. Thus, the machine does not 'know' that the sentence is true while we can see that it is true. Therefore, we cannot be a machine.[9]

If the arguments based on Gödel's theorems are sound, then we must accept the theoretical possibility of 'super-Turing' computation, i.e. theoretical devices which are strictly more powerful with respect to computability than Turing machines. Examples of such powerful devices have been explored theoretically, for instance Zeno machines (accelerated Turing machines), which allow – in contrast to ordinary Turing machines – a countably infinite number of algorithmic steps to be performed in finite time, or analog neural networks, which allow computation over arbitrarily precise real values.[10] So it seems van Gelder is right in claiming that some class of neural networks can have super-Turing capacities. However, no plausible account of how such devices could be physically realised has been offered so far. Both Penrose's appeal to quantum properties of the brain and Siegelmann's arbitrarily precise neural networks fail to take into account the noise inherent in any real world analog system. In general, any physical system, including the brain, is susceptible to thermal noise. This simple fact defeats the possibility of the arbitrarily precise information transfer required for super-computation.[11]

Van Gelder follows H. T. Siegelmann, who repeatedly appeals to a result of Siegelmann and Sonntag, when they argue in later papers that analog neural networks do not require arbitrary precision, and thus are physically realisable.[12] In particular, Siegelmann and Sonntag's Lemma 4.1 shows that for every neural network which computes over real numbers, there exists a neural network which computes over truncated reals, i. e. reals precise only to a finite number of digits. However, the length of truncation required is a function of the length of the computation – longer computations require longer truncated strings. Consequently, if length of

9 Lucas' argument was revived by Penrose, who supplemented it with the claim that quantum properties of the brain allow it to solve problems that are uncomputable by a Turing machine. See Penrose (1994). Lucas' argument has been strongly criticised by logicians and philosophers (e.g. Benacerraf (1967), Pudlak (1999)), as has Penrose's (e.g. Feferman (1995)).

10 Cf. Syropoulos (2008).

11 Cf. Arora/Barak (2009).

12 Cf. Siegelmann/Sonntag (1994).

computation is allowed to grow arbitrarily, so must the length of the strings of digits over which the computation is performed in a truncated network. Therefore, one still must allow computation over arbitrarily precise reals if one is considering the computational properties of analog neural networks in general, i.e. over arbitrarily long computation times. Otherwise it is not clear whether such analog neural networks have super-Turing capacities.

In the current state of research it does not seem plausible that we can realise devices with super-Turing capacities, or that the brain is a realisation of a super-Turing device. In contrast, we already implement Turing machines in digital computers. This simple fact renders the CTT at least not physically implausible. In contrast, we have rendered physically implausible the existence of dynamical neural networks as specified in van Gelder's footnote 19. We slowly turn to the question whether his claim that the symbolic hypothesis is opposed to the dynamical hypothesis is substantive. But let us first introduce the ideas behind the connectionist perspective on cognition in order to be able to judge, whether there is a substantive difference between the two paradigms.

## IV The Connectionist Perspective

Connectionists try to figure out how computations are implemented in the brain. The neuron hypothesis of Ramón y Cajal has dominated neuroscience since the late 19th century. He was the first to observe and report the division of brain tissue into distinct cells: neurons. More importantly, he postulated a flow of information from axon to dendrite through the web of neural connections, which he denoted by drawing arrows on his illustrations of neural tissue. It suggests itself (also from a symbolicist perspective) to identify this flow of information from neuron to neuron as the *locus* of computation for solving cognitive tasks.

An alternative to the neuron hypothesis comes from the dynamical systems perspective, which asserts that the behaviour of a family of neurons cannot be reduced to signals between them. Instead, the dynamical perspective asserts that computations should be modeled in terms of a dynamical system seeking basins of attraction.[13] Neuroscientists such as

13 An attractor is the point, or set of points, a trajectory in a dynamical system tends towards over time. A basin of attraction is defined by a collection of points in state space that a trajectory can start out to eventually arrive at the attractor.

Freeman find support for the view in observed neural dynamics.[14] That a family of neurons is irreducible to signals between them corroborates van Gelder's view that the DH is a substantive alternative to the SH.

In the following section, we begin to argue that there is no logical incompatibility between neural networks and formal theories (logics) *per se*. Quite the contrary, neural networks were developed historically in very close relation to classical logic. It was even shown that the first artificial neural networks can be represented by classical logic. But first we want to clarify the role logic can play in the analysis of van Gelder's opposition claim.

Logic provides an abstract symbolic perspective on neural computation given the neuron hypothesis is true. As such, logic can never be the whole story of the implementation level, which by definition involves the physical instantiation of an algorithm. Nevertheless, logic can help to bridge the gap between the implementation and algorithmic levels by analysing structural similarities across different proposed instantiations. For example, if we subscribe to the neuron hypothesis, it seems obvious to look for logic gates in the wiring between neurons; if we subscribe to the alternative dynamical systems hypothesis, it seems obvious to look for logic gates corresponding to family patterns of neurons, or equivalent structure in the relations between basins of attraction in a dynamical system. Logical analysis can distinguish the commonalities across different implementation levels from their true disagreements as we will see in the next section. Representation theorems play a major role in such a logical search for similarities.

## V Logical Neurons and Representation Theorems

Classical logic does not only provide the groundwork for the abstract theory of computation, it also motivated models of neural behaviour, i.e. neural networks, both historically and today.[15] The classic work of McCulloch and Pitts proved the first representation theorem for a logic in an artificial neural network. In general, a representation theorem demonstrates that for every model of a theory, there exists an equivalent model within a distinguished subset. In McCulloch and Pitts' case, the 'theory' is just a

14 Freeman (1972); see also Freeman (2000).

15 Cf. the classic work of McCulloch/Pitts (1943), and a more recent work of Sandler/Tsitolovsky (2008).

time-stamped set of propositional formulas representing a logical derivation, and the distinguished subset in question is the set of neural networks satisfying a particular set of assumptions, namely, neural firing is 'all or none', the only delay is synaptic delay, and the wiring of the network does not change over time. McCulloch and Pitts show the opposite direction as well: the behaviour of any network of the specified type can be represented by a sequence of time-stamped propositional formulas. The propositions need to be time-stamped to represent the evolution of the network through time: the activations of neurons at time $t$ are interpreted as a logical consequence of the activations of neurons at time $t + 1$.

McCulloch and Pitts had shown how neurons could be interpreted as performing logical calculations, and thus, how their behaviour could be described and analysed by logical tools. Furthermore, their approach was modular, as they demonstrated how different patterns of neural wiring could be interpreted as logic gates: signal junctions which compute the truth value of the logical connectives (conjunction, disjunction, or negation) of incoming signals. The applications of this result are, however, limited by its idealising assumptions. As neurophysiology has enriched our understanding of neural behaviour, the hypothesis of synchronised computations cascading through a structurally unchanging network has become too distant from neural plausibility to resolve debates about implementation in the brain.

Nevertheless, logical methods continue to provide insight into the structure of neural computation. In the face of an increasingly complex theory of neurophysiology, two research projects present themselves. The first focuses on realistic models of individual neurons. Sandler and Tsitolovsky for example, begin with a detailed examination of the biological structure of the neuron, and then develop a model of its behaviour (using fuzzy logic).[16] A second project focuses on artificial neural networks designed to mimic brain dynamics as closely as possible. For example, Vogels and Abbott ran a number of simulations on large networks of integrate-and-fire neurons.[17] These artificial neurons include many realistic features such as a resting potential and a reset time after each action potential is generated. After randomly generating such networks, Vogels and Abbott investigated their behaviour to see if patterns of neurons exhibited the characteristics of logic gates. They successfully identified patterns of

16 Sandler/Tsitolovsky (2008).

17 Vogels/Abbott (2005).

activation corresponding to NOT, XOR, and other types of logic gate within their networks.

The idealising assumptions of these models continue to temper the conclusions which can be drawn from them. Nevertheless, there is a trend of increasing fit between mathematical models of neural behaviour and the richness of neurophysiology; and logic continues to guide our understanding of neurons as computational units. But from the standpoint of cognitive science, an describing and/or explanatory question remains: are these computational units the right primitives for analysing cognition? More generally, is there some in principle difference between an analysis offered in terms of neural networks and one offered in terms of logical rules as suggested by van Gelder? In order to make this question more precise we present in the following section a sketch of the debate between symbolicists and connectionists about mental representations.

## VI The Controversy of Symbolicists and Connectionists about Representation

In an influential paper, Fodor and Pylyshyn argued that (i) mental representations exhibit systematicity; (ii) representations in neural networks do not exhibit systematicity (in contrast to symbolic representations); therefore, (iii) the appropriate formalism for modeling cognition is not connectionist (but symbolic). Systematicity here is just the claim that changes in the meaning of a representation correspond systematically to changes in its internal structure (e.g. from our ability to represent 'Bonnie loves Clyde', it follows that we are also able to represent 'Clyde loves Bonnie').[18] Fodor and Pylyshyn claim that the only case in which representations in a neural network do exhibit systematicity is when the network is a 'mere' implementation of a symbolic system.[19]

It is important to notice what is at stake here: if cognitive tasks manipulate representations, then the appropriate analysis of a cognitive task must respect the properties of those representations. According to Fodor and Pylyshyn, the claim that descriptions and/or explanations in cognitive science must be in terms of symbolic systems does not, however, restrict attention to the symbolic level. Paradigmatic examples of the symbolic

18 Fodor/Pylyshyn (1988).

19 They do, however, not indicate how such implementational networks avoid their general critique, cf. Chalmers (1990).

approach in cognitive science such as Chomsky investigate the role of particular algorithms for solving information processing tasks (such as extracting syntactic structure from a string of words).[20] Nevertheless, the claim is that somewhere between abstract task specification and physical implementation, descriptive and explanatory power breaks down, and neural networks (except 'mere' implementations of symbolic systems) fall on the implementation side of this barrier.

The response from connectionist modelers was vehement and univocal: Fodor and Pylyshyn had simply misunderstood the representational properties of neural networks. Responses elucidated how representations in neural networks are 'distributed' or 'subsymbolic'. Smolensky, van Gelder, Clark, and many others all emphasised the importance of acknowledging the distinctive properties of distributed representations in understanding the difference between neural networks and symbolic systems.[21] Yet it is difficult to put one's finger on just what the essential feature of a distributed representation is which makes it qualitatively different from a symbolic representation. Since the late 1990's, the supposed distinction has largely been ignored as hybrid models have risen to prominence, such as the ACT-R architecture of Anderson and Lebiere.[22] Such hybrid models combine neural networks (for learning) and symbolic manipulation (for high level problem solving). Although pragmatically satisfying, and strongly questioning van Gelder's incompatibility claim, the hybrid approach avoids rather than resolves questions about the essential difference between symbolic and distributed representations.

However, we are now able to pose the above question more precise: Is there some in principle difference between subsymbolic computation by neural networks over distributed representations and symbolic computation by Turing (or equivalent) machines? The representation theorem of McCulloch and Pitts discussed above suggests differently, namely that logical theories and neural networks are essentially the same, i.e. their

20 Chomsky (1957).

21 Smolensky (1987): 137-63; Smolensky (1988): 1-74; van Gelder (1990): 355-364; van Gelder (1991): 355-381; Clark (1993). In his paper "The dynamical hypothesis in cognitive science" van Gelder even holds an anti-representationalism, stating that "[u]nlike digital computers, dynamical systems are not inherently representational. A small but influential contingent of dynamicists have found the notion of representation to be dispensable or even a hindrance for their particular purposes." van Gelder (1998): 626.

22 Anderson/Lebiere (1998).

computational and representational properties respectively are logically equivalent. Can this result be extended to a larger class of neural networks including recurrent neural networks? The trick, as it turns out, is to treat neural computation as non-monotonic.

## VII Representation Theorems for Non-Monotonic Logics in Neural Networks

We can easily see that some particular non-monotonic theories may be represented by neural networks.[23] Consider, for instance, a system for reasoning about birds: two input nodes – one for $Bird(x)$ and one for $Penguin(x)$ – and an output node – for $Fly(x)$ – are all we need to model this system with a simple neural network. As long as there is an excitatory connection from $Bird(x)$ to $Fly(x)$ and at least as strong an inhibitory connection from $Penguin(x)$ to $Fly(x)$, this network will produce the same conclusions from the same premises as our non-monotonic example theory in the appendix. But this is just a simple specific case; a representation theorem for non-monotonic logics (NMLs) in neural networks would show us that for every non-monotonic theory, there is some neural network which computes the same conclusions. Such a theorem would demonstrate that NMLs and neural networks can compute the same functions, have the same computational capacity, exhibit the same computational power, or – simply put – are computationally equivalent.

Representation theorems for the computational equivalence of NMLs and neural networks have already been given. For instance, Balkenius and Gärdenfors consider the inferential relationship between a fixed input to a neural network and its so called 'resonant state', i.e. the stable activation state it reaches given the fixed input.[24] By partitioning the state space of these networks into 'schemata', i.e. informational components closed under conjunction, disjunction, and complementation, they demonstrate that the relation between input schemata and the corresponding resonant state or output schemata satisfies the axioms of a NML.

Balkenius and Gärdenfors' result sheds some light on the formal relationship between neural networks and symbolic systems.[25] However, their

23 A short glimpse on non-monotonic logics is given in the appendix. For an introduction to non-monotonic logics we recommend Makinson (2005).

24 Balkenius/Gärdenfors (1991): 34 f.

25 The result serves also a practical purpose. In practical applications, an algo-

research attention in "Nonmonotonic Inferences in Neural Networks" was directed to very simple networks in which the atomic schemata, treated as propositions in a NM inference relation, correspond to single nodes.[26] If we want to resolve the connectionist (dynamicist) vs. symbolic systems debate, then the result of Balkenius and Gärdenfors needs to be supplemented in two ways: (i) by extension to more realistic neural networks; (ii) by extension to the case of actually distributed representations. We do not consider the required supplement (i) here.[27] Instead, we directly jump to an examination of supplement (ii).

Leitgeb identifies himself with the tradition originating from Balkenius and Gärdenfors, yet aims to establish a broader class of results. The theorems discussed so far address the relationship between a particular NML and a particular type of neural network. In contrast, Leitgeb proves a sequence of representation theorems for each non-monotonic system introduced by Kraus et al. in distinguished classes of neural networks.[28] Leitgeb's results involve inhibition nets with different constraints on internal structure, where an inhibition net is a spreading activation neural network with binary (i.e. firing or non-firing) nodes and both, excitatory and inhibitory connections. Leitgeb even extends his results by proving representation theorems for the same logics into interpreted dynamical systems.[29]

A dynamical system, at the most abstract level, is a set of states with a transition function defined over them. An interpretation function $I$ of a dynamical system maps formulas from a propositional language to regions of its state space. Leitgeb obtains closure under logical connectives *via* the same strategy Balkenius and Gärdenfors employ, namely by assuming an ordering $\leq$ over informational states. If $S_I$ is an interpreted dynamical

rithm for constructing a neural network from a set of non-monotonic inference rules has computational value, because it can efficiently find the fixed point which maximises satisfaction of the inference rules. Unfortunately, this computational efficiency can only be achieved on a case by case basis.

26 Cf. Balkenius/Gärdenfors (1991): 35-37.

27 Cf. for a treatment of supplement (i) Stenning/van Lambalgen (2008). They focus on neural networks which plausibly represent actual structure in the brain. There they consider networks made of nodes with sigmoid activation functions which model the behaviour of actual neurons more realistically.

28 Cf. Leitgeb (2001); cf. also Leitgeb (2003); cf. Kraus/Lehmann/Magidor (1990).

29 Cf. Leitgeb (2005).

system, then $S_I \models \phi \Rightarrow \psi$ iff $s$ is the resonant state of $S_I$ on fixed input $I(\phi)$ and $I(\psi) \leq s$. We call the set of all such conditionals $TS_I$. Then two of Leitgeb's theorems follow:

> Theorem 1. If $S_I$ is an interpreted dynamical system, then the theory $TS_I$ is closed under the rules of Kraus et al.'s non-monotonic system $C$.

> Theorem 2. If $T_\Rightarrow$ is a consistent theory closed under the rules of the non-monotonic system $C$, then there exists an interpreted dynamical system $S_I$ such that $TS_I \equiv T_\Rightarrow$.

Unlike Balkenius and Gärdenfors, Leitgeb takes pains to ensure that his representation theorems subsume the case of distributed representation. In particular, the interpretation function $I$ may map a propositional formula to a set of nodes, i.e. distributing its representation throughout the network. Hence, from a philosophical perspective, Leitgeb's results are highly interesting for the debate between symbolic and dynamicist (including some connectionist) approaches. He has shown that any dynamical system performing calculations over distributed representations may be interpreted as a symbolic system performing non-monotonic inference. His result appears to show that there is no substantive difference in the computational and/or representational power of symbolic systems and that of dynamical neural networks. It seems that Smolensky's subsymbolic hypothesis, "The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual-level description", is refuted, since Leitgeb has proven that every subsymbolic dynamical system admits a complete, formal, and precise symbolic description, namely a non-monotonic description.[30]

Leitgeb's results allow us to switch back and forth from a symbolic description of cognition to a subsymbolic one and justify the use of a hybrid system combining symbolic and subsymbolic components (cf. ACT-R). That means we can use the virtues of both descriptions and avoid simultaneously their shortcomings. A symbolic system, for instance, has the advantage that we can read off the assumptions on which its inferences have been based, and how the conclusions have been derived. A dynamicist artificial neural network (ANN), in contrast, gives us only information about the trajectory of states which led to its stable or resonant state (i.e. the 'conclusions').

30 Smolensky (1988): 6-7.

In case there is a principled difference between symbolic and subsymbolic systems, the key to articulating it may be embedded somewhere in Leitgeb's assumptions. The step of interest seems to be the ordering over informational states of the respective network; it is an open question whether the states of actual neural networks – to which connectionists attribute representational properties – satisfy such an ordering. So there is work yet to be done in providing a full resolution to the symbolic systems vs. connectionism debate (recall also required supplement (i)).

Even if there is no substantive difference between the representational capacities of symbolic systems and those of neural networks, there may be other differences between their computational powers, in particular concerning their computational efficiency. For instance, a NM system in the style of Kraus et al. (i.e. one with symbolic knowledge base) is able to integrate a new defeasible conditional to its knowledge base without further complications. In contrast, the *whole* topology of the corresponding ANN may, in this case, be required to readjust. On the other hand, if the ANN's topology satisfies all inferences that can be drawn in the corresponding NM system, the network is perhaps significantly quicker. These potential differences seem to be worth to pursue in order to determine – if at all – in which respects symbolic and subsymbolic systems differ.

## VIII Conclusion

Van Gelder's claims (ii) and (iii), namely if there is a fundamental difference between symbolic systems and dynamical systems, then there is a fundamental difference concerning their computational and representational powers and if symbolic and dynamical systems differ concerning the computational and representational powers, then dynamical systems have a greater computational power than symbolic systems, are refuted *via* Leitgeb's sequence of representation theorems, since they show that there is no substantive difference in the computational and representational power of symbolic systems and that of dynamical systems. However, it remains unclear whether claim (i) holds, namely that there is a fundamental difference of symbolic systems and dynamical systems. Indeed, modus tollens speaks against claim (i), but it seems intuitively more plausible that the consequence of van Gelder's claim (ii) is problematic. If there is any difference between the two kinds of systems, Leitgeb's results and the

corresponding assumptions should help us articulating it. Our suggestion is to take a closer look at Leitgeb's ordering over informational states of the respective network. Moreover, even if there is no fundamental difference, there may be less principled differences in the computational efficiency of the corresponding systems. A future investigation of these potential differences in computational efficiency would clarify in which respects symbolic and dynamical-connectionist systems differ.

Since claims (ii) and (iii) are refuted and thereby claim (i) rendered questionable, van Gelder's conclusion (vi) loses its force, namely that cognition described by dynamical systems cannot be described by symbolic systems. Indeed, every subsymbolic dynamical system admits a symbolic, non-monotonic description. Moreover, since the two paradigms seem to be equivalent with respect to computational and representational powers, there is *prima facie* no need for cognitive scientists to prefer the dynamicist approach to cognition over the symbolicist, or *vice versa*. Quite the contrary, Leitgeb's results suggest that they should – and, more importantly, can – be seen as *fruitfully complementing* each other. Hence we are not unjustified if we reject van Gelder's corollary (vii) and treat both, the symbolic systems perspective and the dynamical systems perspective, as substantive hypotheses within the computational paradigm of cognitive science. And wouldn't it be desirable for the future of cognitive science, if physical and biological analyses of real-life dynamical systems are translated into symbolic systems for implementation purposes, and the resulting constructions are analysed in terms of dynamical systems in order to compare the constructions with data of real-life cognisers?

## Appendix: Non-Monotonic Logics

Classical deductive reasoning does not exhaust logical inference. In a complex and changing world, cognitive agents must draw conclusions about their circumstances on the basis of incomplete evidence. Crucially, this evidence is defeasible, which means that conclusions drawn from it may be defeated by later evidence. Let us first see what monotonicity means for functions.

A funtion $f$ is said to be *monotonic* if $n \leq m$ implies $f(n) \leq f(m)$; as the input grows, the output grows as well. Reasoning in classical logic is monotonic, because adding new premises always allows you to generate more conclusions. Let $T$ and $T'$ represent consistent sets of sentences

and let $F(T)$ denote the deductive closure of $T$ (i.e. the set of all sentences which follow from T by some specified (classical) inference rules). Then, for all classical logics, $T \subseteq T'$ implies $F(T) \subseteq F(T')$.

Typically, a non-monotonic logic supplements an underlying classical logic with a new, non-monotonic connective and a set of inference rules which govern it. The rules describe a logic of *defeasible* inference, inferences which may be defeated by additional information. For instance, from the fact that *this is a bird*, we can usually conclude *this can fly*. This inference, however, can be defeated, if we learn that *this is a penguin*. Symbolically, we want our system to ensure that $Bird(x) \Rightarrow Fly(x)$, but $Bird(x) \wedge Penguin(x) \not\Rightarrow Fly(x)$. The example illustrates why such a system is non-monotonic, since $\{Bird(x)\} \subset \{Bird(x), Penguin(x)\}$ *yet* $F(\{Bird(x)\}) \not\subset (\{Bird(x), Penguin(x)\})$.

The basic idea is easy to see. If we allow ourselves default assumptions (that is formally speaking a set of conditional assertions in a knowledge base) about the state of the world, we can easily reason about how it changes. Without the basic assumption that features of the world not mentioned by the incoming evidence do not change, we would waste all our computational resources checking irrelevant facts about the world whenever we received new information. This consideration inspired John McCarthy's assertion that, not only do "humans use [...] 'non-monotonic' reasoning," but also "it is required for intelligent behavior".[31]

Kraus et al. provide a unified approach to a hierarchy of non-monotonic logics of varying strengths.[32] They use a preference ('plausibility') ordering over worlds as a model for non-monotonic inference. They achieved that increasingly strict constraints on this semantic ordering correspond to increasingly powerful sets of syntactic rules, and used this insight to define the systems $\mathbf{C} \subseteq \mathbf{CL} \subseteq \mathbf{P} \subseteq \mathbf{M}$, where **C** ('cumulative reasoning') is the weakest non-monotonic system they consider and **M** ('monotonic') is equivalent to standard propositional logic. Intermediary systems are characterised semantically by added constraints on the plausibility ordering over worlds and syntactically by the addition of stronger inference rules. For example, models for **C** are sets of worlds ordered by a relation which is asymmetric and well-founded. **C**, for instance, is strengthened to the system **CL** by adding the inference rule Loop:

31 McCarthy (1980): 28.
32 Kraus et al. (1990).

$$\frac{\phi_0 \Rightarrow \phi_1, \phi_1 \Rightarrow \phi_2, \ldots, \phi_{k-1} \Rightarrow \phi_k, \phi_k \Rightarrow \phi_0}{\phi_0 \Rightarrow \phi_k} \; .$$

Semantically, models for CL add the constraint that $\prec$ be transitive, i.e. form a strict partial order.

*About the author:*

*Mario Günther received his BA in Philosophy (major) and Cognitive Science (minor) from* Albert-Ludwigs-Universität Freiburg im Breisgau. *Currently he is enrolled in the MA program 'Logic and Philosophy of Science' offered by the* Munich Center for Mathematical Philosophy *at* Ludwig-Maximilians-Universität München.

*Über den Autor:*

*Mario Günther hat einen B. A. in Philosophie (Hauptfach) und Kogntionswissenschaft (Nebenfach) an der A*lbert-Ludwigs-Universität Freiburg im Breisgau *erlangt und ist im Masterstudiengang ‚Logic and Philosophy of Science' am* Munich Center for Mathematical Philosophy *an der* Ludwig-Maximilians-Universität München *immatrikuliert.*

## References

Anderson, John R.; Lebiere, Christian: *The Atomic Components of Thought*, Lawrence Erlbaum, Mahwah (NJ) 1998.

Arora, Sanjeev; Barak, Boaz: *Computational Complexity: A Modern Approach*, Cambridge University Press, Cambridge/New York 2009.

Balkenius, Christian; Gärdenfors, Peter: "Nonmonotonic Inference in Neural Networks", in: J.A. Allen; R. Fikes; E. Sandewall (eds.), *Knowledge, Representation and Reasoning: Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo (CA) 1991, 32-39.

Benacerraf, Paul: "God, the Devil, and Gödel", in: *The Monist*, 51 (1967), 9-32.

Chalmers, David: "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation", in: *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Cambridge (MA) 1990, 340-347.

Chomsky, Noam: *Syntactic Structures*, Mouton de Gruyter, The Hague 1957.

Clark, Andy, *Associative Engines*, Bradford Books, Cambridge (MA) 1993.

Cooper, Barry S.: *Computability Theory*, CRC Mathematics Series, Chapman & Hall/CRC, London 2003.

Feferman, Solomon: "Penroses Gödelian Argument", in: *Psyche 2* (1995), online publication, available at http://www.calculemus.org/MathUniversalis/NS/10/04feferman.html.

Fodor, Jerry A.; Pylyshyn Zenon W.: "Connectionism and Cognitive Architecture: A Critical Analysis", in: *Cognition*, 28 (1988), 3-71.

Freeman, Walter J.: "Waves, Pulses and the Theory of Neural Masses", *Progress in Theoretical Biology*, 2 (1972), 87-165.

Freeman, Walter J.: *How Brains Make Up Their Minds*, Columbia University Press, New York (NY) 2000.

Kraus, Sarit; Lehmann, Daniel J.; Magidor, Menachem: "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", in: *Artificial Intelligence*, 44, 1-2 (1990), 167-207.

Leitgeb, Hannes: “Nonmonotonic Reasoning by Inhibition Nets”, in: *Artificial Intelligence*, 128, 1-2 (2001), 161-201.

Leitgeb, Hannes: “Nonmonotonic Reasoning by Inhibition Nets II”, in: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11, 2 (2003), 105-135.

Leitgeb, Hannes: “Interpreted Dynamical Systems and Qualitative Laws: From Neural Networks to Evolutionary Systems”, in: *Synthese* 146 (2005), 189-202.

Lucas, John R.: “Minds, Machines and Gödel”, Philosophy, 36, 137 (1961), 112-127.

Makinson, David: “Bridges from Classical to Nonmonotonic Logic”, in: I. Mackie (ed.), *Texts in Computing*, volume 5, King’s College London, London 2005.

McCarthy, John: “Circumscription – A form of Non-Monotonic Reasoning”, in: *Artificial Intelligence*, 13, 1-2 (1980), 27-39.

McCulloch, Warren S.; Pitts, Walter H.: “A Logical Calculus Immanent in Nervous Activity”, Bulletin of Mathematical Biophysics, 5 (1943), 115-133.

Newell, Allen; Simon, Herbert: “Computer Science as Empirical Inquiry: Symbols and Search”, in: *Communications of the ACM*, 19, 3 (1976), 113-126.

Penrose, Roger: *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press, New York 1994.

Pudlàk, Pavel: “A Note on Applicability of the Incompleteness Theorem to Human Mind”, in: *Annals of Pure and Applied Logic*, 96, 1-3 (1999), 335-342.

Siegelmann, Hava T.; Sonntag, Eduardo D.: “Analog Computation via Neural Networks”, in: *Theoretical Computer Science*, 131 (1994), 331-360.

Smolensky, Paul: “The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn”, in: *Southern Journal of Philosophy*, 26, Supplement (1987), 137-63.

Smolensky, Paul: “On the Proper Treatment of Connectionism”, in: *Behavioral and Brain Sciences*, 11 (1988), 1-74.

Stenning, Keith; van Lambalgen, Michiel: *Human Reasoning and Cognitive Science*, MIT Press, Cambridge (MA) 2008.

Syropoulos, Apostolos: *Hypercomputation: Computing Beyond the Church-Turing Barrier*, Monographs in Computer Science, Springer, Berlin 2008.

Turing, Alan Mathison: "Computing Machinery and Intelligence", in: *Mind*, 49 (1950), 433-460.

Van Gelder, Timothy: "Compositionality: A Connectionist Variation on a Classical Theme", in: *Cognitive Science*, 14 (1990), 355-364.

Van Gelder, Timothy: "Classical Questions, Radical Answers: Connectionism and the Structure of Mental Representations", in: T. Horgan, J. Tienson (eds), *Connectionism and the Philosophy of Mind*, Kluwer Academic Publishers, Dordrecht 1991, 355-381.

Van Gelder, Timothy: "The Dynamical Hypothesis in Cognitive Science", in: *Behavioral and Brain Sciences*, 21, 5 (1998), 615-665.

Van Leeuwen, Marco: "Questions for the Dynamicist: The Use of Dynamical Systems Theory in the Philosophy of Cognition", in: *Minds and Machines*, 15 (2005), 271-333.

Vogels, Tim P.; Abbott, Larry F.: "Signal Propagation and Logic Gating in Networks of Integrate-and-Fire Neurons", in: *The Journal of Neuroscience*, 25, 46 (2005), 10,786-10,795.