Computer Modeling in Philosophy

Robert Prentner* and Chris Fields

Using AI Methods to Evaluate a Minimal Model for Perception

Abstract: The relationship between philosophy and research on artificial intelligence (AI) has been difficult since its beginning, with mutual misunderstanding and sometimes even hostility. By contrast, we show how an approach informed by both philosophy and AI can be productive. After reviewing some popular frameworks for computation and learning, we apply the AI methodology of "build it and see" to tackle the philosophical and psychological problem of characterizing perception as distinct from sensation. Our model comprises a network of very simple, but interacting agents which have binary experiences of the "yes/no"-type and communicate their experiences with each other. When does such a network refer to a single agent instead of a distributed network of entities? We apply machine learning techniques to address the following related questions: i) how can the model explain stability of compound entities, and ii) how could the model implement a single task such as perceptual inference? We thereby find consistency with previous work on "interface" strategies from perception research.

While this reflects some necessary conditions for the ascription of agency, we suggest that it is not sufficient. Here, AI research, if it is intended to contribute to conceptual understanding, would benefit from issues previously raised by philosophy. We thus conclude the article with a discussion of action-selection, the role of embodiment, and consciousness to make this more explicit. We conjecture that a combination of AI research and philosophy allows general principles of mind and being to emerge from a "quasi-empirical" investigation.

Keywords: Computer Modeling; Agents; Networks; Perception; Consciousness; Embodiment; Interfaces; Attractors; Learning; Genetic Algorithms

1 Introduction

1.1 AI and Philosophy

The relationship between philosophy and research on artificial intelligence (AI) has been difficult since its beginning, with John Lucas, Hubert Dreyfus, John Searle, and Roger Penrose,[1] among many others, arguing that AI is impossible, pointless, or misguided. On the other side of the divide, philosophers such as Aaron Sloman, Jack Copeland, or Margaret Boden[2] are more sympathetic toward AI research as a means to advance the understanding of our minds.

At the same time, a major criticism targeted at philosophical research was mounted by AI researchers: without proper implementation, purely theoretical or conceptual research is futile. While the current hype

---

1 Lucas, "Minds, Machines, and Gödel"; Dreyfus, *What Computers Can't Do;* Searle, "Minds, Brains, and Programs"; Penrose, *The Emperor's New Mind.*
2 Sloman, *The Computer Revolution;* Copeland, *Artificial Intelligence: A Philosophical Introduction*; Boden, *The Creative Mind.*

---

*Corresponding author: Robert Prentner, University of California, Irvine, United States of America; E-mail: robert.prentner@uci.edu;
Chris Fields, independent scholar, Caunes Minervois, France

about AI with its focus on technical gadgets is deplorable, we take this critique seriously. In the first part of this contribution, we thus adopt AI's methodology[3] of "build it and see" to address a problem at the intersection of ontology and psychology: the problem of defining a minimal model in which "perception" and "agency" can be meaningfully discussed. Here we explicitly distinguish perception from mere sensation, following standard accounts in the field but also Gibson, the embodied cognition movement, and the theory of active inference in viewing perception as the acquisition of significant, actionable information from an uncertain, challenging environment.[4]

While the questions of philosophy may not be solved by AI, some of them can be translated into a language suitable to be explored using logic, computer modeling or other methods of AI research. Such questions are generically of the following type: provided any particular model, how can one study phenomenon x with the help of such methods? A pioneering example was given by Alan Turing. Given a machine which follows a finite set of instructions and can do nothing but read symbols from a tape, move the tape, write or erase symbols, and change its state, which kinds of computations can it (not) perform? Turing later embraced the question of whether such a machine could appear to think, that is, whether it could successfully play the "imitation game".[5] Historically, Turing's work led to the question how one could *actually build* a machine (classical or not) that could simulate a mind. To date, this has not been demonstrated. The methodology of testing theoretical ideas by implementation has, however, been applied to a wide variety of specific questions about the mind, from the requirements for logical reasoning[6] to the nature of creativity.[7] Two of the surprising insights to emerge from such work are that everyday human capabilities such as visual navigation or natural language understanding are far harder than playing chess or solving college-level math problems, and that experts in virtually any domain do not solve problems by following rules that they can report in language.[8]

Here we apply the AI methodology to a problem similar to Turing's. Our model comprises a network of very simple, but interacting agents which have binary experiences of the "yes/no" type and communicate their experiences with each other. How can a network of such simple agents act as a single, coherent, stable entity? How can communication within such a network lead to coherent collective information processing, e.g. the perception of external signals, by the network as a whole? To investigate these questions, we implement a computational model and apply standard tools from machine learning which could be regarded as techniques to infer and optimize model parameters.

Specifying stable patterns and solving a particular task (e.g. perception) are, we claim, necessary conditions for regarding such collective systems as agents in their own right. But they are not yet sufficient, a point we return to in our closing discussion.

In the second part of this contribution, we look at the way conceptual analysis could be beneficial to AI research. AI research, if it is intended to contribute to a more general understanding of intelligence, would benefit from explicitly considering issues previously raised by philosophy. For example, philosophers have long debated the nature of goals and intentionality in relation to mind and being, whether from a naturalist standpoint,[9] phenomenologically[10] or in an existentialist framework.[11] Some of this inspired later work on AI.[12] In this spirit, we discuss how systems of distributed agents could select appropriate actions, the role

3 Russell and Norvig, "Artificial Intelligence: A Modern Approach", 1.

4 Gibson, *The Ecological Approach to Visual Perception;* Knill and Richards, *Perception as Bayesian Inference*; Chemero, *Radical Embodied Cognitive Science*; Friston, "Life as we know it"; Hoffman et al., "Interface Theory of Perception".

5 Turing, *The Essential Turing*, 433. Note the particular specification given in the text of what is nowadays called a "Turing-machine". Turing's 1936 insights regarding the limits of computation did *not* concern machines in general but this specific architecture.

6 Newell and Simon, "The logic theory machine".

7 Hofstadter and Mitchell, "The copycat project: A model of mental fluidity and analogy-making".

8 For a general review, see Chater, *The Mind is Flat*.

9 Dretske, *Flow of Information;* Dennett, *Intentional Stance*.

10 Husserl, *Logical Investigations;* Merleau-Ponty, *Phenomenology of Perception*.

11 Heidegger, *Being and Time*; Sartre, *Being and Nothingness*.

12 Movements within AI that explicitly embrace such issues include artificial life, e.g. Steels and Brooks, *The Artificial Life Route*; embodied/enactive cognition, e.g. Froese and Ziemke, "Enactive artificial intelligence"; developmental robotics, e.g. Cangelosi and Schlesinger, *Developmental robotics*; and "biologically inspired" computing, e.g. Goertzel et al., "A world survey of artificial brain projects".

of embodiment, and how the external world would appear to them. In particular the role of learning is discussed. This would make the appearance of a single unified entity more cogent.

1.2 Existing Computational Frameworks and How They Are Used

Artificial neural networks (ANNs), also known as connectionist systems,[13] are layered networks consisting of many elementary constituents ("nodes") which are roughly modeled along physical neurons. Simple neural networks based on such nodes (also called "perceptrons") were first conceived by McCulloch and Pitts to study intelligent behavior.[14] In this model, each node receives one or more inputs and, after a weighted sum Σ has been computed, outputs 1 if Σ reaches a pre-defined threshold b. In subsequent elaborations, the output-characteristic of each neuron was taken to be sigmoidal, that is, it takes the value of the continuous function $\sigma(\Sigma, b) = (1 + \exp(-(\Sigma - b)))^{-1}$. But also functions other than sigmoidal are conceivable. For example, a popular choice nowadays is the "half-wave rectifier".[15]

The architecture of an ANN usually features three distinct sets of nodes, which are organized in layers. Nodes comprising the input layer receive environmental inputs and relay these inputs to nodes comprising one or more "hidden" layers (such an architecture is also known as "multilayer perceptron").
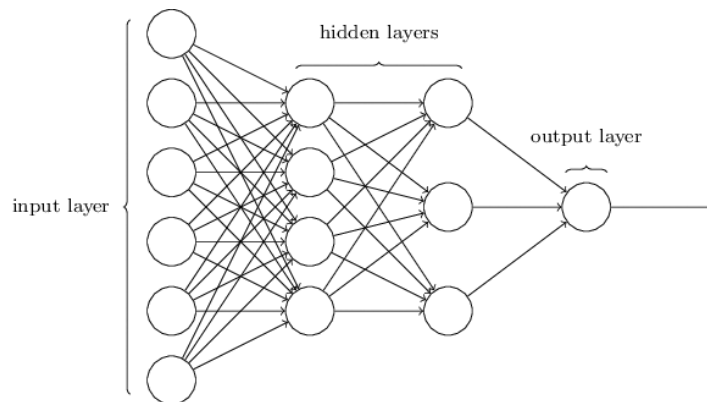

Figure 1. A simple ANN featuring two hidden layers of neurons; adapted from Nielsen, "Neural Networks and Deep Learning".

Each hidden element sends a signal, based on the input it receives, to neurons in the next layer and so on. Finally, output nodes receive signals from hidden elements and specify the output of the neural net. The individual weights are representing the connection strengths between nodes. Standardly, ANNs do not have connections withing layers. However, this feature is implemented in *recurrent* neural nets (RNNs), e.g. Grossberg's "adaptive resonance" networks.[16] RNNs are promising tools, even though they are more difficult "to train" than conventional ANNs. The weighting of connections between nodes is central for learning, which is usually implemented via adjusting weights, using, for example backpropagation techniques.[17] Employing hierarchically structured ANNs equipped with multiple hidden layers of processing neurons led to the development of deep learning methods. ANNs are now the best publicly-known tools from AI research due to their broad applicability and some astonishing success stories (e.g. google's *alpha go*[18]).

---

13 McClelland et al., *Parallel distributed processing*. Early connectionist models interpreted nodes roughly as concepts, along the lines of semantic network models. Later models interpreted nodes "sub-symbolically," more in line with ANNs; see Smolensky, "On the proper treatment of connectionism".
14 McCulloch and Pitts, "A logical calculus".
15 LeCun et al., "Deep learning".
16 Grossberg, "Nonlinear neural networks".
17 Rumelhart et al., "Learning representations by back-propagating errors". This technique has been widely applied and extended to multi-layer "deep learning" networks, c.f. Schmidhuber, "Deep learning in neural networks".
18 Silver et al., "Mastering the game of Go".

ANNs and their focus on optimization techniques strongly emphasizes learning as a primary hallmark of intelligent systems.

However, there exist also other computational frameworks for modeling. One is the framework of cellular automata, which was originally conceived of by John von Neumann and Stanislaw Ulam in order to study the process of self-replication.[19] Cellular automata came to fame when in 1970 a Scientific American article[20] presented John Conway's "game of life". In a cellular automaton, unit cells are placed along a k-dimensional lattice (grid) and are updated according to a (globally valid) rule that specifies the evolution of each cell as a (local) function of the states of its neighbors. While lattice topology and neighborhood could be of different sizes, it was observed that already very simple setups (1-2 dimensions, 3-5 neighbors) could lead to the formation of seemingly complex and meta-stable patterns. In the "game of life", for example, a simple rule defined on a 2-dimensional grid would lead to the formation of entities that seem to move along the grid, eating other entities and reproducing (Fig. 2), even though any "movement" was only apparent, and any "causal" power of these entities was just due to the "hidden" rule governing the underlying update scheme.
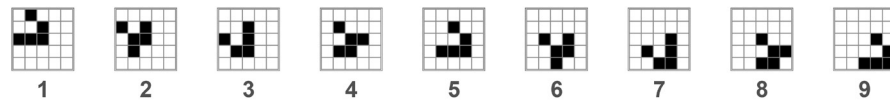

Figure 2. Time steps in Conway's game of life. A "gilder" is an emergent entity that seemingly "moves" across the grid. Created with Wolfram's Mathematica 11.

One philosophical lesson is that any ascription of causal powers to macroscopic "objects" arising from the cellular dynamics is precarious. Similarly, any claim regarding the "causal powers of the brain"[21] or other macroscopic physical objects might be equally precarious, e.g., when taken as referring to the capacity of some brain areas to "produce" mind. Any such causation might be distributed among many interconnected parts instead.

The perhaps most influential recent development in the theory of cellular automata came from Stephen Wolfram, who devised a numbering scheme of all "elementary" (1-dimensional) cellular automata and who classified cellular automata into different classes based on their long-term behavior.[22] These correspond to (i) homogenous structures, (ii) simple and regular structures, (iii) pseudo-random and chaotic structures, and (iv) complex structures with a mixture of order and randomness. Relatively simple cellular automata can be designed which are computationally universal and even reversible. The framework of cellular computing is often summarized as "simple + parallel + local"[23] and is appealing both in terms of computerization and in terms of application to biological and social systems.

Random Boolean Networks (RBN) are a related computational model, originally proposed by Stuart Kaufman to study gene regulatory networks in the late 1960's.[24] RBNs consist of a finite number (sometimes in the range of only a few dozens) of nodes which could be in either of two states. The states of the individual nodes evolve according to an updating rule which is a function of the states of each node and its neighbors, similar as in the cellular-automata framework. The future state of the network is determined fully by its present state. The topology of the network is given by a graph which determines which nodes interact with each other or, equivalently, the neighborhood structure of the RBN. RBNs are good model systems to study emergent behavior and phase transitions that occur already within a reasonably sized network. Often the topology is randomized, and typical research questions pertain to the relation between statistical properties of the underlying graph (e.g., its degree-distribution) and the dynamics of the network or, more generally speaking, to a possible characterization of RBNs in terms of their information-processing capacities (e.g., mutual-information between nodes or network states).

19 Von Neumann and Burks, "Theory of self-reproducing automata".

20 Gardener, "Life", 120.

21 Searle, "The Causal Powers of the Brain".

22 Wolfram, *A New Kind of Science,* 231ff.

23 Sipper, "Simple + parallel + local = Cellular Computing".

24 Kauffman, *Origins of Order,* 465.

A particularly useful concept which is similar to ideas in physics is the one of "attractors" in "state space". The latter is defined as the space spanned by all possible states (configurations) of the network; the dynamics of the network, starting in some initial state, corresponds to a trajectory through state space. Eventually, in a finite network, such trajectories reach an attractor, referring to a set of states which seems to "pull" the states in their "vicinity" towards the set (formally, the set of states which eventually land in an attractor is called "the basin" of the attractor).[25]

Early on, Kauffman speculated whether biological systems are such that they exhibit order at "the edge of chaos", occupying regions in state space near the transition from ordered to chaotic behavior. [26] This is an example for how technical questions in computer modeling could shed light on fundamental questions, such as "What is life?".

Another model which is related to the cellular-automata framework is agent-based modeling (ABM). It was introduced to model the complex behavior of a system of interacting individual agents. Agents could make simple decisions based on the (limited) information available to them. The information available to an agent is strictly local, however, the patterns that emerge are global. A prominent example was devised by the Nobel laureate Thomas Schelling, who studied the effects of segregation in the United States using a very simple model.[27] Given a fictional city, individuals can choose whether to change their neighborhood, based on their preference regarding their (local) neighborhood structure. Segregation (a global phenomenon) could occur even if an agent's preference was just to live in a community where one is not in a minority. This suggested how a society can (arguably) exhibit racist tendencies even without (explicit) racist prejudices on the individual level and without a global mechanism that "concerts" individual actions. The agents' "decisions" resemble the update rules of a cellular automaton.[28] In many cases, the agents used in ABMs are mere sketches of individuals and are more similar to "dumb" molecular machines rather than to complex psychological entities. Despite this, ABMs are used as prolific tools in sociology,[29] and they are conceptually close to Bruno Latour's Actor-network theory,[30] when "actors" are defined as any entity which influences the state of other entities in the network. ABMs illustrate that much of what we call complex or cognitive (psychological) capacities could already be modeled in terms of structured societies of relatively simple agents.

2 Perceptual Networks

2.1 The Role of Ontology in Models

Underlying any model in science are assumptions about which things exist and how they relate to each other. Often, computational models are inspired by empirical processes. ANNs are modeled after the structure of physical neurons. RBN are modeled after ("first-order") chemical reactions. ABM are modeled after decision-making individuals. These models do not strive to be realistic representations and involve idealization. Real neurons are much more complicated than perceptrons and involve processing at many biological levels (pertaining to synaptic potentials, neuro-transmitters, architectural constraints etc.).

25 While RBNs are conceptually quite simple, it took several decades until their mathematical understanding took off. For example, an intriguing question pertains to the relation of network-size N and attractor length. In 1969, Kauffman suggested that the mean number of attractors scales with the square-root of N. While it was later shown mathematically that the mean number of attractors grows faster than polynomial, subsequent work on asynchronously updated RBNs confirmed a power law distribution for critical networks as initially proposed by Kauffman; cf. Samuelsson and Troein, "Superpolynomial Growth in Kauffman-networks", 589; Greil and Drossel, "Dynamics of Critical Kauffman-networks", 180.

26 Kauffman, *Origins of Order,* 27.

27 Schelling, "Dynamic Models of Segregation", 149.

28 Note that the agents' decision-making could be described by the update rules of a cellular automaton if one regarded not the agents which move between houses but the houses, which do not move but are populated by agents, to be the elementary constituents of the model.

29 Macy and Willer, "From Factors to Actors". Hegselmann et al, *Simulation in the Social Sciences*.

30 Latour, *Reassembling the Social*.

For example, the dynamics of real (biological) neurons has a "refractory period", that is, a time after previous excitation during which a neuron cannot again be excited. Refractory periods are important biologically, since they prevent neurons from constant firing. Also, learning in biological systems is different to the current learning methods employed for ANNs. Biologically plausible mechanisms should inherently be local and could affect the topology of the network itself (other than merely adjusting weights in a pre-defined network). A prominent early proposal is Hebbian learning[31] – neurons that "fire together, wire together".

The other models listed above are highly idealized too. Chemical (enzymatic) reactions are not of a digital (on/off) type and could be of various orders (first, second or third). Individuals are not simple quasi-physical mechanisms but complex psychological agents. Correspondingly, AI frameworks are not necessarily used to theorize about the things they were originally modeled on. Google's *alphaGo*, for example, does not try to re-engineer or explain neural processes but to solve the problem of defeating human Go-players. It is also possible to use frameworks to study processes which at first sight seem more appropriate to be studied with different models. In principle, cellular automata could be used to study the dynamics of the brain even though the brain's constituents are, to a first approximation, much more similar to the elements of an ANN.[32] In general, then, the elementary entities of a computational model need not resemble its explanandum nor do they need to encode a (realistic) representation of it. However, the computational model should be able to reproduce (at least in an idealized way) the explanandum's formal (structural) properties. Put more philosophically, there is a certain looseness of fit between the ontology of a model and the thing studied. On a more radical reading, there may simply be no uniquely specifiable ground of being, other than the ontologies we use to inform our models. The model presented in this article is premised on the existence of agents which have experiences and communicate them to other agents in the network. There is no ontological structure other than this.

2.2 Computational Framework

The model which we propose in this paper, *perceptual networks*, combines many of the crucial features of the above modeling paradigms but gives them a more psychological interpretation. More concretely, perceptual networks are defined as networks of individual and excitable agents that (i) have experiences corresponding to their states, (ii) are located on a graph which encodes their interaction-topology, and (iii) whose individual actions affect the performance of the network. Such agents perceive certain messages and send messages, based on their experiences (states), to other agents in their environment.

Similar to ANNs, perceptual networks can learn from experience by either changing their updating rule or by adjusting their connectivity within the network. These two processes correspond to a perceptual learning process (i.e. how to represent incoming information) and an action-related learning process (i.e. where to send information to) respectively. The analysis of information processing capacities and the "information geometry"[33] embodied by a perceptual network is afforded by its basic graph structure. Perceptual networks share this feature with RBNs, and the idea that processes on the individual (local) level affect the global functioning of the network is consistent with the main thrust of ABMs and cellular automata.

We will demonstrate in the following sections how perceptual networks could address two different but related problems. First is the problem of how our simple model could give rise to stable structures. Given the more radical reading advanced above, this problem is an *ontological* one. Without the bedrock of some essential "core" which our model is about, how is it that the basic setup of our model (a network of agents which occasionally get excited) could give rise to a stable structure? Second is the problem of how these

---

31 Hebb, *The Organization of Behavior*.
32 Acedo, "A cellular automaton model for collective neural dynamics".
33 Amari, *Information Geometry*.

structures enable "perception" at the scale of the whole network, where perception is understood broadly in terms of a "representation" which results from the application of (internal) rules embodied in the network, under conditions of uncertainty (modeled as noise). Note that our account of perception already includes a sense of "action" (in the form of communication between agents (nodes) within the network). Further ascriptions of agency pertain to more elaborate decision-making capabilities, the body and consciousness, which will be further commented on in the final section of the paper.

Perceptual networks are a simplification of a framework previously proposed to study consciousness.[34] Given an external measurable space (a "world"), a "conscious agent" in this framework is a six-tuple consisting of two measurable spaces ("perception space" and "action space"), three kernels that connect the world and these spaces ("perception", "decision" and "action") and a time-counter which counts the number of kernel executions. This definition was refined and a network consisting entirely of "reduced conscious agents", a concept introduced to distinguish "internal" aspects of a conscious agent from its "extrinsic" aspects, was introduced.[35] It was shown theoretically that such a network could be usefully applied to problems in psychology and reproduce the architecture of many received models in cognitive science.

Perceptual networks feature node ("agents") that integrate information ("messages") and relay this information to neighboring nodes in the network. The mapping of sensory information could be represented, for each node j, by a rule R_j which, given a message m_j, together with its current state x_j, outputs its next state x'_j. The message m_j is determined by the topology of the network A (represented by its adjacency matrix in the following).

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \underline{A} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{pmatrix} \{\underline{R_j}\} \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix} \rightarrow \text{etc} \tag{1}$$

For simplicity, we look at deterministic rules, but the model should eventually be extended to model probabilistic inferences. We stress that such an update would represent a Markov process.[36]

The construction leaves open the possibility that the rules might change in the course of evolution. This is a prerequisite for any kind of learning in the network. Learning could affect the rules describing information integration, $\{R_j\}$, as well as network topology A.

2.3 Asymptotic States and Attractors

In total, the rules define a circular updating scheme from network states to network states. We assume a homogenous updating rule $R = \{R_j\}$ of the form:

$$m_j = (Ax)_j, \qquad x'_j = \begin{cases} 1 \text{ if } m_j \geq 1 \\ 0 \text{ else} \end{cases}, \tag{2}$$

and we illustrate the state evolution by a toy network comprising $N = 4$ agents, as shown in Table 1. The continuous and dotted lines represent two different adjacencies A and A' (left and middle column), giving rise to two different state evolutions, each starting in the initial state (1,0,1,0) (right column). Since the network is finite and closed, the state evolution will either land in a stationary state or, more generally, in an attracting set.

---

34 Hoffman and Prakash, "Objects of Consciousness", 6.

35 Fields et al. "CA-networks", 190.

36 More precisely, the update process corresponds to a homogenous discrete-time Markov chain, which would feature at least one stationary distribution (under the assumption that it is irreducible and non-periodic), cf. Chapter 3 in Gebali, *Computer Networks*.

Table 1. Example of a Perceptual Network with two different adjacencies (left column; continuous and dotted lines respectively). Initially, nodes 1 and 3 are excited, the rules (given in text) and the adjacencies of the network (middle column) determine the state evolutions (right column).

Network	Adjacencies	State Evolutions

Which attracting set a state is part of (or whether it is a transient state) is determined by the network topology and the rules governing its evolution. In our toy model, the same initial state (1,0,1,0) reaches two different attracting sets respectively (depicted by the arrows in Table 1). In the first case, the network realizes a flip-flop circuit which permanently oscillates between the states (1,0,1,0) and (0,1,0,1) Table 1, right column, left-hand side). In the second case, the network falls into an attracting set which cycles through its states with period 3 (Table 1, right column, right-hand side). In this case the initial state furthermore is transient. In both cases, an initial state of (0,0,0,0) would be stationary and could be interpreted as ground- or terminal state of the network.

The stationary states and attracting sets represent the stable entities which are associated with the network. This gives an explicit answer to the question how the structure underlying our model could give rise to stable patterns on the level of networks. We will now show how such patterns could also be conceived of as encoding perceptual inferences.

3 Perception

3.1 Basic Facts About Perception

In physics, the notion of "observation" remains controversial, with "observers" defined by different authors as anything from a moving coordinate system (a "Galilean observer") to a complex information processing system capable of (usually Bayesian) inference.[37] Most agree, however, that an observer is a physical system, and that any physical system can be considered to be an observer. Often "observation" is understood simply in terms of a symmetric interaction between entities which thus "observe" each other (e.g. photons "observe" an electron passing through a slit) or between entities which are in a definite state of motion relative to some other state (e.g. a lamp that "observes" the movement of my hand when writing this paper[38]). In perception science, by contrast, one usually makes a difference between a mere "sensation" and the "perceptual representation" of this sensation, such as the difference between the recording of an optical signal on the retina and some additional neuronal processing of this signal in the brain. This additional processing is modeled as inference, typically Bayesian inference.[39] Alternatively, proponents of "direct" perception theory argue for a direct relation between environmental input and behavioral output

---

37 For an overview and arguments favoring the latter, information-processing view, see Fields, "If physics is an information science, what is an observer?"

38 Rovelli, "Relational Quantum Mechanics", 1641.

39 Perception as Bayesian inference is the standard view in the field, see e.g. Knill and Richards, *Perception as Bayesian Inference*. The notion of "active inference" embeds this standard view within a larger framework describing a behaving organism within a responsive world, see Friston, "Life as we know it". We return to this in Sect. 5.

of the organism, thus dropping the concepts of information processing, inference, and representation from perceptual theory.[40] In any case, proponents of representationalism or direct-perception theory have to account for the difference between a perception and a mere physical recording as can be illustrated, for example, by the possibility to *mis*represent, or in direct-perception terms, behave maladaptively in response to an environmental signal. Without taking up the ontological baggage of either representationalism or direct-perception theory, we shall liberally use the notion of "representation" in the sense of a collectively stable state generated by the dynamics after perturbation in a (probabilistic) dynamical system. This does not imply that there is no sense in speaking of "information processing" that is going on in the system, nor does it mean that perception is fundamentally "dis-embodied" and purely abstract. We furthermore adopt the following set of assumptions:

1. Perception is an ill-posed problem. The perceptual system constructs a coherent percept based on the stimuli it receives. The mapping from stimuli to percepts is, however, not unique. For any set of stimuli, there are many percepts which are consistent with this set. The perceptual system has to choose the "right" percept out of many possible ones. Perception is thus akin to a form of inference which is abductive rather than deductive.

    Perception is furthermore a constructive process, it creates "objects that are not really there". Visual illusions are particularly suited to illustrate this. In neon-color spreading, a subject perceives an illusionary glowing disk even though no corresponding shape is explicitly part of the image the subject perceives (Figure 3). The visual system is "filling in" the glowing disk based on the stimuli it receives and the experimental setting.
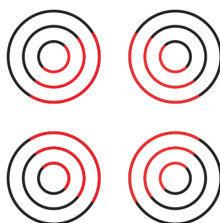

Figure 3. Neon-color spreading. A glowing disk is experientially filled-in by the visual system.

2. Perception is probabilistic. In the course of the last century two approaches in psychology have competed with each other.[41] In a logic-based account, reasoning cognition is modeled along the execution of deterministic rules. A second approach models cognitive processes probabilistically. To each possible outcome in a psychological experiment, one could associate a probability. For example, in word recognition to each word one can assign a probability on the basis of presented letters on a piece of paper. Given a flawless text, this task is rather trivial. However, when letters are missing, this might quickly become difficult. Does the word sh*p refer to a vessel or a place where one could buy things? In psycho-physical experiments devised by Richard Warren, it was demonstrated that subjects report different percepts based on the same partially-masked phoneme ("$eel") they hear in different contexts (sentences):[42]

What subjects actually heard	What subjects believed to have heard
The $eel is on the shoe	The heel is on the shoe
The $eel is on the car	The wheel is on the car
The $eel is on the table	The meal is on the table
The $eel is on the orange	The peel is on the orange

---

40 Gibson, *The Ecological Approach to Visual Perception*; Chemero, *Radical Embodied Cognitive Science*.
41 Mumford, "Pattern Theory", 2.
42 Warren, "Restoration of missing speech".

Note that the solutions that subjects could come up with are mostly ambiguous. Different meanings are in principle consistent with a given (gapped or distorted) sequence of letters. Nevertheless, subjects are able to approximately "choose" the "right" percept. The probabilistic approach is inherently related to the idea that psychology presents us with ill-posed problems. The probabilistic approach has been successfully applied in linguistics since the 1950's. Since the 1980's it came to dominate AI and perception research. It is also worthy to note that the word-recognition task implies a high degree of contextuality. It is the context – and not knowledge about the "inherent" meaning of a word – that lets us get the message approximately right.

3.2 Perceptual Inference

Mathematically, the process of perception can be modeled as abductive inference using Bayesian probability theory. The perceptual system chooses from a variety of "interpretations", each consistent with an input signal. A posterior probability can be assigned to each interpretation, given the input. Using Bayes' rule, it can be recovered from multiplying a likelihood function that assigns a probability to each input given a particular percept with the prior probability of that percept:

$$p(x|i) = p(i|x) \cdot \frac{p(x)}{p(i)} = \alpha \ p(i|x) \cdot p(x),$$
(3)

where $\alpha = p(i) = \sum_x p(i|x)p(x)$ is a normalization constant, which can be neglected when comparing different posterior probabilities given the same input. The way this scheme is (approximately) realized in a living organism can be quite complicated. In vision science, it is often assumed that the space of percepts is homomorphic to the space of states in the world, and thus the likelihood function could be given the interpretation of "a mapping" from world states to input states, e.g., the optical projection from a 3D world onto the retina of the eye. Such an approach to visual perception is typically called an "inverse-optics"[43] approach, since the task for the visual system is to undo the effects of optical projection. Given such an interpretation and knowing likelihood and prior probability, it is possible to calculate posterior probabilities using the Bayesian scheme above. To determine which percept x is selected based on the posterior distribution, one usually introduces a "loss function" which describes the consequences of making an error in the process of choosing an interpretation. The easiest and most principled loss function is given by a delta-function, centered around the maximum of the posterior distribution (the so-called "MAP" - estimate). Other, more involved choices are possible.[44]

However, in an evolutionary setting, the assumption that world states and scene interpretations generically are homomorphic has been shown to be wrong,[45] and we thus refrain from it. For a perceptual network we assume that the "sensory input" i corresponds to the initial state of a certain (sub-)set of nodes and "perceptual interpretations" x correspond to the stable states of the network. On the network level, the set of rules $\{R_j\}$ gives rise to a "perceptual strategy", and knowing the rules $\{R_j\}$ and the network's topology, we could compute posterior probabilities directly based on priors defined on inputs to maximize a quantity that represents a "fitness payoff" with respect to the problem at hand. What the network sees is not merely "given" by its input but by the utility of its strategy. We thus introduce a "payoff-function" defined on the space of inputs and perceptions:

$$F : I \, X \rightarrow [0,1], (i,x) \, \alpha \, r \in [0,1].$$
(4)

For convenience, one often chooses the interval [0,1] for payoff values, but in general any finite interval defined on the positive Reals will suffice. Perceptions, on this account, do not serve as truthful representations of input but as useful guides toward maximizing payoffs.

With this definition of a payoff-function one could calculate an "expected payoff" for a given set of updating rules $\{R_j\}$ and graph G:

$$\langle F \rangle = \sum_{i,x} F(i,x) p(i,x | \{R_j\}, G).$$
(5)

43 Palmer, *Vision Science,* 23.

44 Mamassian et al., "Bayesian Modelling of Visual Perception", 20.

45 Prakash, "Invention of Structure" 111, and references cited therein.

We want to infer the "average best" perceptual strategy implemented by a network. Since the probability is parametrized by the set of rules $\{R_j\}$, we need to adjust the rules in order to maximize the expected payoffs (assuming an otherwise fixed topology of the graph G):

$$\Rightarrow \{R_j\}_{\text{opt}} = \operatorname{argmax}\left(\langle F \rangle\right) = \operatorname{argmax}\left(\sum_{i,x} F(i,x)\, p(i,x|\{R_j\},G)\right). \tag{6}$$

From now on we will drop the conditional dependence on $\{R_j\}$ and G for simplicity when writing fitness payoffs.

3.3 Mutual Information

In our model we assume a particular form of the payoff function which equals the logarithm of the probability $p(i,x)$ divided by the probability of the marginal probabilities $p(i)p(x)$. This results in a mutual information (MI) between X and I:

$$\langle F \rangle \equiv MI(I,X) = \sum_{i,x} p\,(i,x)\,\log_2\left(\frac{p(i,x)}{p(i)p(x)}\right) = \sum_{i,x} p\,(x|i)p(i)\,\log_2\left(\frac{p(x|i)}{p(x)}\right) \tag{7}$$

$$= \sum_{i} p(i) D_{KL}\big(p(x|i)\big\|p(x)\big) = \big\langle D_{KL}\big(p(x|i)\big\|p(x)\big)\big\rangle_i. \tag{8}$$

In other words, the payoff-function is set equal to the difference in self-information between a posterior probability $p(x|i)$ and a marginal probability $p(x)$. More intuitively, F could be thought of as quantifying the information that is present in the asymptotic state about the initial state with the Kullback-Leibler divergence D_{KL}.[46]

The mutual information is bounded above by the entropy $H(I)$:

$$MI(I,X) \le H(I), \tag{9}$$

so the maximum of mutual information would be obtained in a case where the perceptual rule leads to a one-to-one mapping from inputs to perceptions. In any realistic setting (e.g. where the space of inputs largely exceeds the state of percepts) this will likely not be the case, and we have to relate input and perception probabilistically and optimize Eq. (8) instead. Already a noisy input (such as in the example of speech recognition above) would make this a stochastic problem, which would generally lower the maximal value of mutual information achievable in the network. To each sensation there exist several perceptual representations which all have a certain posterior probability, which is specified by the network's evolution under noise (here "noise" is taken to lead to a perceptual "misrepresentation" of the input).

On this account the "goal" of perception is being able to distinguish as best as possible between different inputs. Perception, thus understood, amounts to the ability to recognize differences in the world (or more generally: perception is the ability to recognize "differences that make a difference"[47] for fitness). Finding a set of rules and a topology defined *for individual agents* which maximize the mutual information between perception and input to *the network* is the "perceptual problem" that needs to be solved by our model. For this, we can borrow techniques to search through the rule-set using algorithms from machine learning. The setup of the problem suggests evolutionary programming techniques such as genetic algorithms[48] as discussed below.

---

46 See Amari, *Information Geometry*, 11. Alternatively, to relate this to other concepts in unsupervised machine learning, maximizing the mutual information introduced here corresponds to maximizing an expected log-likelihood for the probabilities $p(i,x)$ parametrized by the rule set $\{R_j\}$ and defined on the graph G. This formulation makes it in principle amenable to optimization techniques involving calculating gradients of a "free-energy" function. However, in the more general setup where arbitrary (non-differentiable and non-continuous) payoff functions are used, such methods are no longer straightforwardly applicable.

47 Bateson, *Mind and Nature*.

48 Holland, *Adaption in Natural and Artificial Systems*.

4 Preliminary Experiment: Perceptual Strategies

4.1 Model and Methods

We tested the performance for 3-bit inputs distributed on the first 3 nodes of a perceptual network with $N = 16$ agents on a graph G with topology A. The system was initialized in the state $X=(i[1], i[2], i[3], 0, ..., 0)$ and was evolved according to a set of rules $\{R_j\}$. The attracting sets were recorded for each input. This allowed us to construct a conditional probability $p(x|i)$ defined on the space of inputs, which later informed mutual information (i.e. fitness payoffs). For each input bit, we introduced uncertainty by including a probability ($p = 0.05$) to flip an input bit during the initial phase of the evolution.

We then used a genetic algorithm to determine the best set of rules in terms of fitness. Each rule is specified by the parameters μ_1 and μ_2 (see Eq. (10) below), which thus defined the "genes" of an agent. We evolved the network for 50 generations, each comprising 25 individuals. The best 20% were kept ($\lambda = 0.2$), the remaining 80% were generated using a fitness-weighted sexual recombination procedure. We included a fixed mutation probability ($p = 0.01$) of randomly shifting the values of μ_1 and μ_2 by ±1. To make our model biologically more plausible, we assumed a refractory period of 1 for each agent, which means that an agent – after it has been excited – cannot be excited immediately in the next round.

Since we wish to work with a perceptual rule which could handle changes in the size of the neighborhood structure,[49] we chose a rule that assigns the next state of an agent solely dependent on the number of messages it receives, independent of source. In other words, at the basic level our agents do not have the representational capacity to label messages with the "name" of their sender.[50] More concretely, we assume a rule of the following type:

$$x'_j = \begin{cases} 1 \text{ if } m_j \in [\mu_{1,j}, \mu_{2,j}] \text{ and } x_j = 0 \\ 0 \text{ else} \end{cases}, \tag{10}$$

with $1 \le \mu_{1,j} \le \deg(j)$ and $\mu_{1,j} \le \mu_{2,j} \le \deg(j)$ where messages m_j are computed as $m_j = (Ax)_j$ and deg(j) is the degree of the j-th agent. In total there are $1/2 \cdot (\deg(j) + 1) \deg(j)$ possible rules per agent. If we consider a network of N perceptual agents, each potentially following a different rule, there are:

$$N_P = \prod_{j=1}^{N} \frac{(\deg(j)+1)\deg(j)}{2} \approx \left(\frac{\left(\langle \deg \rangle + 1\right)\langle \deg \rangle}{2} \right)^N, \tag{11}$$

different possibly strategies of the network (assuming the degree per agent is roughly constant); an enormous space of strategies. So, even though the rule expressed in Eq. (10) is quite simple, the rule space scales exponentially with the size of the network.

In this investigation, we optimized the set of rules $\{R_j\}$ for a fixed topology defined by the graph G which either resembled (i) a 4×4 grid, (ii) a small world/scale free (SW/SF) network (with initial states forming a fully-connected community), constructed using the Barabási-Albert method,[51] or (iii) a complete graph where every agent is connected to every other agent. We also tested our results against randomly generated (Erdös-Renyi) networks with sparse $\left(\langle \deg \rangle = N/10\right)$ and dense $\left(\langle \deg \rangle = N-1\right)$ average degree distributions. For simplicity, links were assumed to be bi-directional, i.e. agent j both sends and receives a message from agent k.

We also compared the evolved rule-set to a variant of the "majority rule" where an agent transitions from 0 to 1 if and only if at least half of its neighbors are in state 1, else it goes to 0. The majority rule has been found to be very efficient in solving the "density classification task" for a cellular automaton when defined on small-world

---

49 Technically, an object-oriented programming framework allows for this in contrast to a representation using matrices of pre-defined sizes. But for simplicity we refrained from this option.

50 Note that it is plausible that networks with higher representational capacity do develop such a "labeling" which might not be given at the elementary level of single ("1-bit") agents.

51 Albert and Barabási, "Statistical Mechanics of Complex Networks", 71.

graphs.[52] The density classification task is highly global but involves only a single rule at the local level. It could thus be regarded as benchmark test for local and parallelized computational architectures.

Another fixed rule which we used for comparison is the "max-thresh" rule, where the threshold for updating to 1 is set to the degree of the agent:

$$x'_j = \begin{cases} 1 \text{ iff } m_j = \deg(j) \text{ and } x'_j = 0 \\ \qquad\quad 0 \text{ else} \end{cases}. \tag{12}$$

So, for example, if an agent has 3 other neighbors to communicate with, the state transition takes place if and only if $m_j = 3$.

4.2 Results

Results for some exemplary networks are given in Table 2. The rules which evolved after 50 generations of the GA favor "interface strategies",[53] which in this context means that generically there is no similarity between input states and asymptotic states of the network (or the individual perceptual agents). The perceptual states of the network do not mirror any structure in the input other than a probabilistic relationship given by the posterior $p(x|i)$ which informed fitness payoff (= mutual information).

By contrast, the (fixed) majority rule mirrors the structure of the messages of each agent.[54] While it gave satisfying solutions for some topologies, it was generically not able to compete with a collection of rules $\{R_j\}$ on a random topology which implemented an interface strategy. The max-thresh rule generically solved the problem badly except for particular topologies. Again, interface strategies quickly out-competed the max-thresh rule.

Table 2. Results for some randomly initialized networks, defined on a fixed topology visualized in the left column (input states are highlighted in red). The perceptual rules for the fittest strategy after the genetic algorithm are displayed in the middle column by the values of μ_1 and μ_2 for each agent. Averaged values for mutual information (fitness) for the initial and final population after the GA and results for the majority and max-thresh rules are given on right column (the maximally achievable value is 2.14.)
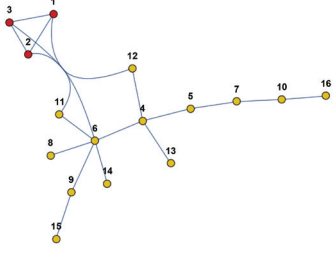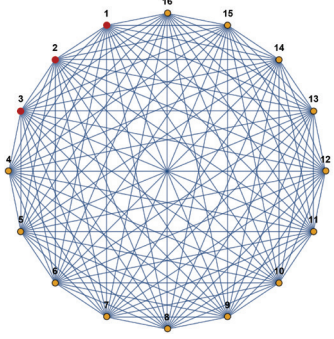
Network, N = 16	Perceptual rule after GA	Mutual information (std. deviation)

Grid, average degree $\langle \deg \rangle = 3$

$$\mu_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 4 \\ 4 \\ 3 \\ 1 \\ 1 \\ 3 \\ 3 \\ 1 \\ 3 \\ 1 \\ 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 3 \\ 3 \\ 2 \\ 2 \\ 4 \\ 4 \\ 3 \\ 3 \\ 1 \\ 4 \\ 3 \\ 2 \\ 3 \\ 3 \\ 1 \end{bmatrix}$$

MI_0	MI_{GA}	MI_{maj}	$MI_{max\,thresh}$
0.45 (0.47)	1.57 (0.27)	0.36	0.36

52 Watts and Strogatz, "Collective Dynamics of 'Small World' Networks", 442.
53 Hoffman et al., "Interface Theory of Perception", 184.
54 More precisely, the majority rule says that the order of the states of the agent is homomorphic to the number of messages it receives. If the "number of messages" is considered to be an external "resource" for the agent, then the majority rule corresponds to a "critical realist strategy" in Hoffman et al., "Interface Theory of Perception", 184.

Table 2. (*Continued*)

Small world/scale free, average degree $\langle \text{deg} \rangle = 2.25$

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 3 \\ 2 \\ 4 \\ 2 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

MI_0	MI_{GA}	MI_{maj}	$MI_{max\,thresh}$
0.33 (0.39)	2.08 (0.17)	0.79	1.10

Complete network, average degree $\langle \text{deg} \rangle = N - 1 = 15$

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \\ 12 \\ 14 \\ 7 \\ 13 \\ 4 \\ 2 \\ 10 \\ 14 \\ 1 \\ 1 \\ 1 \\ 3 \\ 11 \\ 8 \end{bmatrix}, \mu_2 = \begin{bmatrix} 13 \\ 14 \\ 12 \\ 15 \\ 14 \\ 15 \\ 5 \\ 4 \\ 11 \\ 14 \\ 6 \\ 10 \\ 5 \\ 12 \\ 11 \\ 15 \end{bmatrix}$$

MI_0	MI_{GA}	MI_{maj}	$MI_{max\,thresh}$
0.38 (0.54)	2.10 (0.08)	0.0	0.0

Different behavior was observed for different topologies. SW/SF topologies lead to a quick emergence of "fit" strategies with comparatively low computational cost. For grids we found that the average fitness payoff increased more slowly compared to SW/SF graphs, although it sometimes gave good results, depending on the initial configuration of the network. While good and fast converging results have been obtained for complete graphs due to the intrinsic refractory period of the rule, computational time much increased (several hundreds of seconds compared to ca. ten seconds for small grids and SW/SF topologies). The SW/SF topology thus promises to offer a good compromise between speed and fitness.

Similar results have been obtained for randomized graphs, where networks with sparse degree distribution behaved similar to the SW/SF networks (i.e. generically quick convergence and high fitness) and networks with dense degree distributions behaved similar to complete networks (i.e. high fitness but high run-time)

4.3 Discussion and Outlook

In general, we found that the use of genetic algorithms leads to quick convergence on interface strategies embodied by a perceptual network. In the model outlined in the previous section, one could straightforwardly model perceptual inferences as strategies to compute mutual information between past and future states of a network of agents.

While we have optimized rules of the individual agents, the topology of the network has been fixed. In a next step, topology could be optimized as well and we sketch how one could proceed. First, realize that each column of the adjacency matrix could be interpreted as an "action" that each agent takes, i.e. as determining the set of other agents which it will communicate with in the next step. Second, the bi-directionality between nodes could be relaxed. Technically this means considering directed graphs, where the direction of each arrow specifies the direction of communication. Third, and similar as before, a genetic algorithm could be used to update the adjacency matrix, where each column is encoded by a "chromosome" associated to each agent of the network with "genes" representing the locations where messages will be sent to. Occasionally during the recombination process, a random mutation could flip a bit (i.e. a gene), that is, create or destroy a new link in the adjacency.

What would one expect to happen? As previously indicated, small world/scale free topologies present a good compromise between speed and computational cost and are thus expected to be the favored outcomes of a learning (evolutionary) process which optimized the topology (actions) of the network. Such networks already hint at the presence of sub-networks of different sizes and different perception-action capacities, e.g. "hubs" which integrate a lot of incoming information and/or send a lot of messages into the network. It is furthermore likely that a single "node" (network) on a coarser level is itself composed of many interacting sub-networks ("hubs within hubs"). This would indicate that our model is consisting of a hierarchy of heterogenous agents, a fact which is, however, observer-relative and depends on the coarse-graining chosen (on the elementary level the network is just composed of 1-bit agents).

A further developed model could be conceived along several lines. First, a more sophisticated – and heterogenous – model of the refractory dynamics could be used. Agents could have varying periods between excitable states. Second, a non-synchronous updating scheme could be employed. Ideally, the network would feature a combination between asynchronous updating and synchronous updating schemes for several groups of agents. Third, probabilistic updating could be used. In our model, probabilistic inference was restricted to initial uncertainty, the rules and topologies were otherwise deterministic. The natural interpretation of probabilities in the adjacency regards the interaction probability between adjacent nodes. Edges with probability close to 1 represent nodes that almost always exchange signals, whereas edges with probability close to 0 represent nodes that almost never communicate. In a probabilistic setting with low fluctuations around a central value, the network topology might play an even bigger role than in the previous subsection. We conjecture that SW/SF graphs are robust with respect to small fluctuations, as has been demonstrated for many empirical networks which approximate such a topology.[55] This would set SW/SF topologies further apart from, e.g., lattice structures.

The results have not been assessed rigorously in terms of statistics and should only convey a preliminary sense of what is possible with this method of inquiry.

5 Ontology

We have demonstrated in the last section how to couple computational optimization techniques with a simple toy model to address the following two questions: 1) how can a model premised on the experiences of individual agents lead to stability of networks and 2) how could the dynamics of a network of such agents

---

[55] Albert and Barabási, "Statistical Mechanics of Complex Networks", 49ff. It is of particular interest in the present context that both signal transduction networks in biological cells and functional networks in mammalian brains exhibit SW/SF structure; see Agrawal, "Extreme self-organization in networks constructed from gene expression data"; and Sporns and Honey, "Small worlds inside big brains".

be interpreted as perceptual inference? We furthermore showed how our networks could be optimized to efficiently solve a perceptual task. Our results are consistent with findings from research on "complex adaptive systems".[56] We also found consistency with previous work on interface strategies from perception research.

The first question above is ontological. Its solution involved the ability of finite networks to reach an asymptotic state, either stationary or as part of an attracting set, often in very short time. The answer to the second question involved the use of optimization techniques applied to a generic and initially randomized model. Given the particular architecture of our system, we used a technique known as genetic algorithms. Conceptually, genetic algorithms resemble biological learning on a genetic time scale. But nothing specific depends on this choice of optimization technique. Other techniques such as gradient descent, energy minimization or simulated annealing seem in principle viable, given the right architecture (e.g. continuously valued state functions and topologies). It is worth noting that these procedures all operate on relations between nodes: it is the *network itself* that evolves. "Learning" in this sense is a collective, network-level, not an individual, node-level phenomenon.

While our preliminary results demonstrate how philosophical thinking and artificial intelligence research could thus be fruitfully coupled, there are other questions pertaining to agency that have not been addressed. Below, we will select some of these questions and discuss them briefly. Importantly, the goal of this section *is not* to argue that by doing AI research long-standing philosophical puzzles could be resolved. Instead, AI provides us with a constructive methodology by which we could investigate philosophical questions. Investigating such questions might subsequently *enrich* many of the methods in AI research, in particular when interested in artificial *general* intelligence – e.g. for autonomous robotics – rather than in building applications systems for well-defined, highly-constrained task environments.

Exemplary questions pertain to the status of embodied mentality and action, learning, and consciousness. We conjecture that this combination of philosophy and AI research allows general principles of intelligence to emerge from a "quasi-empirical" investigation under philosophical scrutiny.

5.1 Decisions and Agency

We modeled our networks on psychological agents. The elementary entities of our networks are agents that integrate and process information (they "perceive" it). While networks produce stable patterns and solve particular tasks, they would intuitively not yet count as true psychological agents. One distinguishing property of such agents is their ability to make decisions related to their (conscious) perceptions. The original formulation of Hoffman and Prakash and Fields et al.[57] featured an explicit mode of processing which resembled a basic notion of "decision" that each agent makes, based on its current experience. While introducing an explicit decision-making mechanism into our networks is in principle straightforward, there are at least two preconditions that have to be met: i) the decision space must be smaller or equal in size than the perceptual space if a (rational) decision is to be solely based on an agent's perception, and ii) such a decision-mechanism should involve a probabilistic (undetermined) component. A way to implement such a mechanism in our basic model would thus be to assign to each elementary agent a probability to decide on one of two outcomes. On the network level this would correspond to 2^N possible states, each with a possibility that is the product of individual decision probabilities. For $N = 10$ elementary agents, this would already generate 1024 possible states reachable through such a decision-mechanism. Assuming that each agent decides probabilistically with (the same) probability p on one of two possible outcomes and assuming, again, that only the total number k of previous experiences (e.g. (0,0,1,1) being equivalent to (1,0,0,1)) is important, the probabilities on the network level would be given by the binomial distribution:

$$p_{decision} = \binom{N}{k} p^k \left(1 - p\right)^{N-k}.$$

56 Mitchell, *Complexity*.
57 Hoffman and Prakash, "Objects of Consciousness", 6; Fields et al., "CA-networks", 188.

One sees that already some very austere assumptions (elementary agents decide equally; only the number of previous experiences matters) could generate a huge complexity on the network level. This would be even more drastic when those assumptions were relaxed. Even single biological cells, for example, have memory not just of the number but the content and consequences of previous experiences.[58] To model such systems, some elements of the perceptual space X must be set aside to represent memories instead of current percepts.[59]

Real psychological agents have a body that acts as a boundary which lets them interact with their environment. In previous sections, we made the simplifying assumption that asymptotic states of the network, after manually putting the network into an initial state, directly informed our payoff-function. In a more realistic scenario, this is no longer the case: networks are not manually put into an initial state and asymptotic states are not directly read out. As a remedy, one could supply the network with effector organs by which it could interact with other networks or its environment. Similarly for sensations, one could introduce sensory organs to "inform" the network. One could thus think of embedding such networks into an artificial "body"[60] and placing them into an artificial environment. In other words, one would have created an artificial agent – a robot – whose control architecture is a perceptual network. Couplings should be *local*. However, control will be *determined globally* via the (asymptotic) state of the network. One could look into biology to infer a basic set of "actions" for such agents. These would likely include "movement", "consumption", "fight", "flight", and "reproduction". Such different behaviors should be afforded by the effector organs which one gives to the artificial agents. Similar considerations apply to sense organs that let agents "see" (and induce a sense of space), "hear" (induce a sense of succession) and "feel" (induce a sense of touch). A subset of states of the control-structure (the network) would then correspond to "emotions", "moods", "internal perceptions", or "epistemic feelings."[61]

However, before we could tackle particular problems within this framework, we need to (re-)consider some basic ontological questions. First, providing networks artificial bodies, while in principle adequate, seems to be an *ad hoc* solution from an ontological standpoint. The embodiment as described above seems to be a solution to a particular problem (that of making our networks more "agent-like"), but not a principled one. It is not necessitated by our model. The ontological puzzle related to agency, therefore, is whether one could derive "bodies" within the sparse ontology underlying our model.

A possible way to approach this puzzle is as follows. Instead of "injecting" sensations and actions into the network by constructing an artificial agent with sense and effector organs, one would want to identify such organs with certain regularities in the network's dynamics. In this setting, an artificial agent would correspond to a sub-network embedded in a larger network, and the basic set of actions identified above would correspond to particular (local) sub-patterns in its state. Sensations would correspond to the messages that the sub-network receives from other sub-networks, including an "environment" sub-network with slowly varying state. "Movement" through the network could be construed as adjusting the connectivity between sub-networks, i.e. sub-network scale learning, though how such a process would result in a sense of "space" remains to be worked out. Sub-networks could seemingly "consume" or even "kill" each other by changing their respective states (recall that we chose the all-0-state to represent a stationary state of making no experience, i.e. "death"). "Reproduction" could be understood as pattern-emergence after (local) interaction between sub-networks, though again, how much duplication is needed to constitute "reproduction" remains to be determined. Flight-or-fight is a particular decision-strategy on the level of sub-networks. The senses of space, succession, and touch would, in a further development of this approach, correspond to representational strategies which let sub-networks control their actions. Whether such an approach is feasible within a minimal ontology remains to be seen.

---

58 Baluška and Levin, "Cognition throughout biological systems"; Keijzer, "Evolutionary convergence".

59 Fields et al., "CA-networks".

60 Note that this wouldn't yet count as a "living body", which would furthermore require that its bodily integrity follows from a self-organized process of self-assembly and direction of energetic flows; cf. Boden, "Is Metabolism Necessary?", 236.

61 In AI and particularly developmental robotics, these issues are considered under the rubric of "intrinsic motivation"; for a review, see Oudeyer and Kaplan, "What is intrinsic motivation?". The general issue of affect has become significant enough to warrant its own specialty journal from IEEE; for the relevant historical development, see Picard, "Affective Computing".

5.2 Fitness and Learning

Fitness plays an important role in our models, since fitness, and not "truth" (i.e. correspondence between network-states and inputs), determines whether a particular network is apt or not. Fitness was calculated from a payoff-function defined for asymptotic states and inputs (cf. Section 3.3). This is, again, an *ad-hoc* assumption which needs to be revisited. AI architectures too could be distinguished along the way fitness (or "cost" or an "objective function") is defined from the outset. To date, truly general AI frameworks that do not rely on a pre-given function (e.g. a "cost"-function as in supervised learning) or a "truth"-criterion encoded in the detection or reproduction of statistical regularities in the input (e.g. feature extraction as in unsupervised learning) are still in their infancy. Moreover, not every strategy which is good in isolation is good when competing against other strategies. Such an approach to learning thus suggests concepts from evolutionary game-theory and reinforcement learning. Reinforcement learning is the most "biological" learning paradigm in AI research, and is particularly suited to model environmental interactions and games. The key to any such optimization is "learning from experience". Feedback is not directly given to the network by a cost function but as a response, detected via perception, to the outputs generated by the environment. How to implement systems that learn autonomously by exploring their environments is a central question of developmental robotics.[62]

An important fact, related to the difference between unsupervised and reinforcement learning, is that the environment itself may consist of (one or more) other agents. If the environment is large and effectively independent from these agents' actions, fitness consequences can often be written as a result of a (time-independent or only slowly time-dependent) payoff function. If this is not the case, however, the fitness consequences of the agent's outputs are themselves dependent on the strategies taken by the other agents. This difference is similar to that between optimization problems which can be exactly specified mathematically without knowledge of the current states of the network (e.g. finding the shortest path between two points on a map) as opposed to the much more intriguing problems where only a rough goal is specified in advance, but the actual feedback is determined dynamically (e.g. navigating a self-driving car, where the desired location is known but "success" depends on many, in part unpredictable factors). Problems of the latter type are generally considered in terms of "cooperative" or "distributed" systems within AI and robotics.[63] They also arise in the development of operating systems or interpreters.

5.3 Mind and consciousness

Finally, the "gold standard" for counting as agent is to have mind and, ultimately, consciousness. Since perceptual networks are equivalent to Turing complete systems,[64] it is expected that the framework could inform us which non-Turing-computable processes (if any) need to be invoked to study mind and consciousness. Some have argued that mentality evolved as a feature of complex biological systems to solve various control problems (such as self-regulation, synchronization, or preventing over-adaptation to specialized problems) that could not be solved in an algorithmic ("robotic") fashion alone.[65] Other claims regarding the non-computational nature of consciousness involve, for example, mathematical problem solving, understanding, or the process of "comprehension"[66] – all instances of "insightful" problem solving.

---

62 For a range of approaches, see the papers in Baldassarre and Mirolli, *Intrinsically motivated learning*.

63 For an overview, see Parker, "Distributed intelligence".

64 This can be inferred from the fact that perceptual networks could in principle be conceived such as to reproduce other frameworks, which are known to be computationally universal. For example, the perceptual mappings embodied by a node in a perceptual network could be used to implement a universal logic gate (such as the NAND). Given the capacity to arbitrarily route signals, Turing completeness could be inferred from there. Note that this is very different from actually showing how particular computations are carried out by such systems. Note also that computational universality is not a very rare property: Even very simple systems can be computationally universal (Wolfram. *A New Kind of Science*, 716-717).

65 Jordan, "The Wild Ways of Conscious Will"; Augustyn, "Biological Mentality".

66 Penrose, *The Emperor's New Mind*; Searle, "Mind, brains, and programs"; Faggin, "Requirements".

All these proposals emphasize a particular aspect that is deemed necessary for consciousness and postulate a concept of consciousness that would intuitively be called "mind-like" as opposed to a mechanical (purely algorithmic) form of inference ("when encountering this symbol, then do this, else do that" or its stochastic extensions). Perceptual networks could quasi-empirically provide insights here. If one could state these claims precisely in the language of the model, it should be demonstrable whether or not a solution to such a (control or insight-related) problem could be found using perceptual networks. And if not, exactly which capacity should be introduced at the basic level in order to do so could be determined.[67] Even if finding such a solution involved conceptually straightforward optimization techniques, it is plausible that any solution (i) is highly improbable and could otherwise be only reached accidentally (a "measure-zero event"), (ii) is emergent with respect to any local rule-following behavior of nodes, and (iii) would not be described as a "number-crunching" procedure but rather as an "creative act" of the network (due to, e.g., the plasticity of network topology). Such a solution might, for example, involve global coherence that results from an unpredictable local adjustment of the individual strategies of elementary units which happens "on the fly": an interplay between global and local constraints in order to act "mindfully".

This relates to the previous notions of learning or to "habituation". When a network randomly tries out some solution to a problem and gets feedback on whether it was successful or not, this, together with a mechanism for learning, could lead to "knowing" when to act a certain way. One could furthermore conceive of networks that, while not able to generally solve such a problem, can implement a learning strategy that finds a "good enough" solution to the problem of choosing the right strategy when necessary. (And in particular the biological setting suggests that it is really only about "making it better than my competitor".) It naturally follows from the fact that such a learning process depends on the individual history of experiences that strategies that have been successful in one situation will perform poorly in a different situation, and this is exactly what we also see in nature: a way of thinking that was apt in a particular situation, is completely useless in another. Conscious beings can learn to (intuitively?) choose the right way of conceptualizing things and have the right kind of "intuition" or "insight". Machines, at this date, cannot.

These questions raise two issues that are seldom addressed, but deserve further consideration. First is the ("computational") role of the environment. Humans and other organisms routinely store information in the environment, e.g. in the form of books, but also houses, agricultural plantings, etc. This information is relatively stable and does not have to be remembered. Only the means of accessing it must be remembered. From the present perspective, what is important is that such environmental information effectively imposes a top-down constraint on the behavior of not just the encoding agent but other agents sharing the same environment. When the environment itself is allowed to process (i.e. alter in some way) this information, this constraint could be conceived of as computational process.[68] Second, and related, is that from a practical perspective, we typically do not *want* AI systems or other artifacts to be "creative" in the sense of responding unpredictably to environmental states. The exploration of "free" systems that behave in unpredictable ways has been largely confined to simulations.

We see from these brief discussions that AI research could profit enormously from a more "biological" approach to intelligence and a thorough discussion of conceptual issues from philosophy in order to advance the state of artificial general intelligence, in particular with respect to agency and the potential non-computational nature of consciousness.

---

67 For example, it could be the case that a basic capacity of "understanding" has to be granted to even the elementary units of the system to solve such problems. Similarly, when able to use irrational numbers as weights, RNNs have been shown to be able to have a capacity of hypercomputation (Siegelmann, "Neural and Super-Turing Computation"), that is, they would exceed the computational powers of Turing machines. Right now, this seems to be more of a hypothetical rather than actual possibility, but it illustrates the point that "what can be done" with a particular architecture depends on the capacities we "inject" into it (cf. Copeland, "Hypercomputation", 462). In particular, restricting "computational states" to states that can be measured at finite resolution or otherwise determined by observation collapses the computational capability back to Turing; see Fields, "Consequences of nonclassical measurement".

68 This issue is partially addressed within the extended mind framework in cognitive science, e.g. Clark and Chalmers, "Extended mind". It is more thoroughly addressed by the idea of niche construction in evolutionary biology, e.g. Keijzer, "Evolutionary convergence".

Acknowledgements

References

Acedo, Luis. "A cellular automaton model for collective neural dynamics." *Mathematical and Computer Modelling*, 50: 5-6 (2009), 717–725.

Agrawal, Himanshu. "Extreme self-organization in networks constructed from gene expression data." *Physical Review Letters,* 89: 268702 (2012).

Amari, Sun-ichi. *Information Geometry and its Applications*. Tokyo: Springer Japan, 2016.

Augustyn, Kenneth A. (ed). "Biological Mentality.", *Journal of Cognitive Science,* 19:2 (2018), 99 – 284.

Baldassarre Gianluca, and Mirolli, Marco (eds). *Intrinsically motivated learning in natural and artificial systems*. Berlin, Heidelberg: Springer, 2013.

Baluška, František, and Levin, Michael. "On having no head: Cognition throughout biological systems." *Frontiers in Psychology,* 7:902 (2016).

Bateson, Gregory. *Mind and Nature: A Necessary Unity*. New York: Duton, 1979.

Boden, Margaret A. "Is metabolism necessary?" *British Journal for the Philosophy of Science,* 50 (1999), 231–248.

Boden, Margaret A. *The Creative Mind: Myths and Mechanisms*. London: Routledge, 2003.

Cangelosi Angelo, and Schlesinger, Matthew. *Developmental robotics: from babies to robots*. Cambridge, MA: MIT Press, 2015.

Chater, Nick. *The Mind is Flat*. London: Allen Lane, 2018.

Chemero, Anthony. *Radical Embodied Cognitive Science*. Cambridge MA: MIT Press, 2009.

Clark, A. and Chalmers, D. The extended mind. *Analysis,* 58:1 (1998), 7-19.

Copeland, B. Jack "Hypercomputation." *Minds and Machines*, 12:4 (2002), 461–502.

Copeland, B. Jack. *Artificial Intelligence: A Philosophical Introduction*. Hoboken, NJ: Wiley-Blackwell, 1993.

Dennett, Daniel C. *The Intentional Stance*. Cambridge MA: MIT Press, 1987.

Dretske, Fred I. *Knowledge and the Flow of Information*. Cambridge MA: MIT Press, 1981.

Dreyfus, Hubert. *What Computers Can't Do*. Cambridge MA: MIT Press, 1972.

Faggin, Federico. "Requirements for a Mathematical Theory of Consciousness." *Journal of Consciousness*, 18 (2015), 421.

Fields, Chris, Hoffman, Donald D., Prakash, Chetan, and Singh, Manish. "Conscious agent networks: Formal analysis and application to cognition." *Cognitive Systems Research,* 47 (2018), 186–213.

Fields, Chris. "Consequences of nonclassical measurement for the algorithmic description of continuous dynamical systems." *Journal of Experimental and Theoretical Artificial Intelligence,* 1 (1989), 171-178.

Fields, Chris. "If physics is an information science, what is an observer?" *Information,* 3 (2012), 92-123.

Friston, Karl. "Life as we know it." *Journal of the Royal Society Interface,* 10 (2013), 2013.0475.

Froese, Tom, and Ziemke, Tom. "Enactive artificial intelligence: Investigating the systemic organization of life and mind." *Artificial Intelligence,* 173 (2009), 466-500.

Gardener, Martin. "The fantastic combinations of John Conway's new solitaire game 'life'". *Scientific American,* 223 (1970), 120-123.

Gebali, Fayes. *Analysis of Computer Networks,* second edition*,* Cham: Springer, 2015.

Gibson, James J. *The Ecological Approach to Visual Perception,* Boston: Houghton Mifflin, 1979.

Goertzel B, Lian R, Arel I, de Garis H, and Chen S. "A world survey of artificial brain projects, part II: biologically inspired cognitive architectures." *Neurocomputing,* 74 (2010), 30–49.

Greil, Florian, and Drossel, Barbara. "Dynamics of Critical Kauffman Networks under Asynchronous Stochastic Update." *Physical Review Letters*, 95:4 (2005), 177–4.

Grossberg, Stephen. "Nonlinear neural networks: Principles, mechanisms, and architectures". *Neural Networks,* 1 (1988), 17-61.

Hebb, Donald O. *The Organization of Behavior.* New York: Wiley, 1949.

Hegselmann, Rainer, Mueller, Ulrich, and Troitzsch, Klaus G. (eds.). *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View,* Dordrecht: Kluwer, 1996.

Heidegger, Martin. *Being and Time.* Oxford: Basil Blackwell, 1965.

Hoffman, Donald D., and Prakash, Chetan. "Objects of consciousness." *Frontiers in Psychology*, 5:77 (2014).

Hoffman, Donald D., Prakash, Chetan, and Singh, Manish. "The Interface Theory of Perception", *Psychonomic Bulletin and Review*, 22 (2015), 1480-1506.

Hofstadter, Douglas R., and Mitchell, Melanie. "The copycat project: A model of mental fluidity and analogy-making." In *Fluid Concepts and Creative Analogies*, edited by Douglas R. Hofstadter, and the Fluid Analogies Research Group, 205-26. Cambridge, MA: MIT/Basic Books, 1995.

Holland, John H. *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* Cambridge MA: MIT Press, 1992.

Husserl, Edmund. *Logical Investigations*. London: Routledge, 2001.

Jordan, J. Scott. The wild ways of conscious will: what we do, how we do it, and why it has meaning, *Frontiers of Psychology,* 4:574 (2013).

Kauffman, Stuart. *The Origins of Order.* Oxford: Oxford University Press, 1993.

Keijzer, Fred A. "Evolutionary convergence and biologically embodied cognition." *Interface Focus,* 7 (2017), 20160123.

Knill, David C. and Richards, Whitman A. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press, 1996.

Latour, Bruno. *Reassembling the Social: An Introduction to Actor-Network-Theory,* Oxford: Clarendon Press*,* 2005.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. "Deep learning". *Nature*, 521: 7553 (2015), 436–444.

Lucas, John R. "Minds, Machines, and Gödel." *Philosophy,* 36 (1961), 112-127.

Macy, Michael W., and Willer, Robert. "From Factors to Factors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology,* 28:1 (2002), 143–166.

Mamassian, Pascal, Landy, Michael, and Maloney, Laurence T. "Bayesian Modelling of Visual Perception." in *Probabilistic Models of the Brain: Perception and Neural Function,* edited by R.P.N. Rao et al., 13–36. Cambridge MA: MIT Press, 2002.

McClelland, James L., Rumelhart, David E. and the PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: Psychological and biological models. Cambridge MA: MIT Press/Bradford Books, 1986.

McCulloch, Warren S., and Pitts, Walter H. "A logical calculus of ideas immanent in nervous activity." *Bulleting of American Biophysics,* 5 (1943), 115-133.

Merleau-Ponty, Maurice. *Phenomenology of Perception.* London: Routledge & Kegan Paul, 1965.

Mitchell, Melanie. *Complexity: A Guided Tour*, Oxford: Oxford University Press, 2009.

Mumford, David. "Pattern theory: the mathematics of perception." *International Congress of Mathematicians. Vol 3,* 1-3 (2002), 1- 21.

Newell, Allan, and Simon, Herbert A. "The logic theory machine: A complex information processing system." *IRE Transactions on Information Theory* 2:3 (1956), 61-79.

Oudeyer, Pierre-Yves, and Kaplan, Frederic. "What is intrinsic motivation? A typology of computational approaches." *Frontiers in Neurorobotics* 1:6 (2009).

Palmer, Stephen E. *Vision Science: Photons to Phenomenology.* Cambridge MA: MIT Press, 1999.

Parker, Lynne E. "Distributed intelligence: Overview of the field and its application in multi-robot systems." *Journal of Physical Agents,* 2:2 (2008), 5-14.

Penrose, Roger. *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford: Oxford University Press, 1989.

Picard, Rosalind W. "Affective Computing: From laughter to IEEE." *IEEE Transactions on Affective Computing,* 1:1 (2010), 11-17.

Prakash, Chetan. "On Invention of Structure in the World: Interfaces and Conscious Agents." *Foundations of Science*, 134: 3 (2019), 105–16.

Réka, Albert, and Barabási, Albert-Lázsló. "Statistical mechanics of complex networks." *Reviews of Modern Physics,* 74: 1 (2012), 47–97.

Rovelli, Carlo. "Relational quantum mechanics." *International Journal of Theoretical Physics*, 35:8 (1996), 1637–1678.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. "Learning representations by back-propagating errors." *Nature,* 323 (1986), 533-536.

Russell, Stuart, and Norvig, Peter. *Artificial Intelligence: A Modern Approach*. Third Edition. Upper Saddle River NJ: Prentice Hall, 2010.

Samuelsson, Björn, and Troein, Carl. "Superpolynomial Growth in the Number of Attractors in Kauffman Networks." *Physical Review Letters*, 90:9 (2002), 589–4.

Sartre, Jean-Paul. *Being and Nothingness: An Essay on Phenomenological Ontology*. New York: Washington Square Press 1984.

Schelling, Thomas C. "Dynamic models of segregation." *The Journal of Mathematical Sociology*, 1:2 (1971), 143–186.

Schmidhuber, Jürgen: "Deep learning in neural networks: an overview." *Neural Networks* 61 (2015), 85–117.

Searle, John R. "Minds, brains, and programs." *Behavioral and Brain Sciences*, *3* (1980), 417–457.

Searle, John R. "The causal powers of the brain: The necessity of sufficiency" *Behavioral and Brain Sciences,* 13:1 (1990),164

Siegelmann, Hava T. "Neural and Super-Turing Computing." *Minds and Machines*, 13 (2003), 103–114.

Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature,* 529: 7585 (2016), 484–489.

Sipper, Moshe. "Simple + parallel + local = cellular computing." In *Parallel Problem Solving from Nature – PPSN V, LNCS 1498* edited by A.E. Eiben et al., 653-662. Berlin, Heidelberg: Springer, 1998.

Sloman, Aaron: *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind*. Harvester UK: Harvester Press, 1978.

Smolensky, Paul. "On the proper treatment of connectionism." *Behavioral and Brain Sciences,* 11 (1988), 1 – 74.

Sporns, Olaf, and Honey, Christopher J. "Small worlds inside big brains." *PNAS,* 103: 51 (2006), 19219-19220.

Steels, Luc and Brooks, Rodney A. (eds). *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1995.

Turing, Alan. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life* plus *The Secrets of Enigma.* Edited by B. Jack Copeland, Oxford: Clarendon Press, 2004.

Von Neumann, John, and Burks, Arthur. *Theory of Self-reproducing Automata.* Urbana Champaign IL: Illinois University Press, 1966.

Warren, Richard M. "Perceptual restoration of missing speech sounds." *Science,* 167:3917 (1970), 392-393.

Watts, Duncan J., and Strogatz, Steven H. "Collective dynamics of 'small-world' networks." *Nature,* 393:6684 (1998), 440–442.

Wolfram Research, Inc. *Mathematica 11.2*. Champaign IL, 2017.

Wolfram, Stephen. *A New Kind of Science,* Champaign IL: Wolfram Media, 2002.