 doi:10.20944/preprints202205.0121.v1

# Evolution of brains and computers: the roads not taken

Ricard Solé[1,2,3] and Luís F Seoane[4,5]
[1]ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Catalonia
[2]Institut de Biologia Evolutiva, CSIC-UPF, Pg Maritim de la Barceloneta 37, 08003 Barcelona, Catalonia
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA
[4]Departamento de Biología de Sistemas, Centro Nacional de Biotecnología (CSIC), C/ Darwin 3, 28049 Madrid, Spain
[5]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain

When computers start to become a dominant part of technology around the 1950s, fundamental questions about reliable designs and robustness were of great relevance. Their development gave rise to the exploration of new questions such as what made brains reliable (since neurons can die) and how computers could get inspiration from neural systems. In parallel, the first Artificial Neural Networks came to life. Since then, the comparative view between brains and computers has been developed in new, sometimes unsuspected directions. With the rise of deep learning and the development of connectomics, an evolutionary look at how both hardware and neural complexity have evolved or designed is required. In this paper, we argue that important similarities have resulted both from convergent evolution (the inevitable outcome of architectural constraints) and inspiration of hardware and software principles guided by toy pictures of neurobiology. Moreover, dissimilarities and gaps originate from the lack of major innovations that have paved the way to biological computing (including brains) that are completely absent within the artificial domain. As it occurs within synthetic biocomputation, we can also ask whether alternative minds can emerge from A.I. designs. Here we take an evolutionary view of the problem and discuss the remarkable convergences between living and artificial designs and what are the pre-conditions to achieve artificial intelligence.

Keywords: Evolution, brains, deep learning, embodiment, neural networks, artificial intelligence, neurorobotics

## I. INTRODUCTION

With the evolution of life came cognition (Levin & Dennett, 2020). As soon as cells were able to evolve into autonomous agents, the combination of receptors gathering signals and mechanisms of response to those signals rapidly transformed into rich molecular networks. Those networks provided the basis for the smaller scale of computation: survival requires exploiting resources in a reliable way that allows reproduction. Since this is a combination between growing and being robust against fluctuations over a minimal time window, computation was tied to predictive power (Friston, 2018; Jacob, 1998; Rao & Ballard, 1999; Seoane & Sole, 2018). It is this power what actually might foster the evolution towards brains (Llinás, 1988), large and small: in order to reduce the uncertainty of external fluctuations, prediction is a convenient faculty. If we follow the steps towards cognitive complexity that predate the emergence of brains, several key ingredients seem necessary. Looking at their evolutionary emergence will be relevant for our discussion concerning the space of possible cognitive networks. One of them was the invention of neurons: specialized cell types with a marked elongated, branched shape capable of establishing connections. In most cases, these are polar, unidirectional structures, with response functions that involve non-linear thresholds. The power of neurons became a reality as soon as groups of them became interconnected, leading to the first neural networks. Among the key innovations associated to these early assemblies, interneurons must have been a crucial step towards information processing beyond the sensor-actuator chain. With the Cambrian explosion of life, the rise of animals favored the development of sensory organs, learning and movement (Erwin & Valentine, 2013).

All these factors came together within a novel developmental design: brains emerged within bilateral animals, and those newcomers actively explored their worlds, moving around. A compelling proposal concerning the origins of brains is, in fact, the so-called *moving hypothesis*: it posits that the active exploration of the external world fostered the evolutionary path that led to brains (Llinás, 1988). In a novel biosphere dominated by predator-prey arms races, brains were an optimal solution to deal with information. If we fast forward in time, several important changes took place paving the way towards complex brains. This is particularly dramatic for human brains: a rapid expansion during evolution facilitated the addition of microcircuit modules to a multilayered neocortex (DeFelipe, 2011).

Turning our attention to machines, we can see how inventors and scholars have repeatedly drawn inspiration from Nature's cognitive systems. Some times through outright imitation, as in the case of mechanical automata (Fig. 1a). Others by focusing efforts on human-specific cognitive problems (e.g. chess, Fig. 1b), (Yuste, 2015). In yet other cases, through converging metaphors—e.g. from Cajal's flows within neurons and across neural circuits to technological networks that enable the flow of information (Fig. 1c). The exchange of ideas between computational designs and theoretical neuroscience has been constant. And prediction too has been a major force in the development of a large part of technology, particularly after the rise of Information Technology from the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



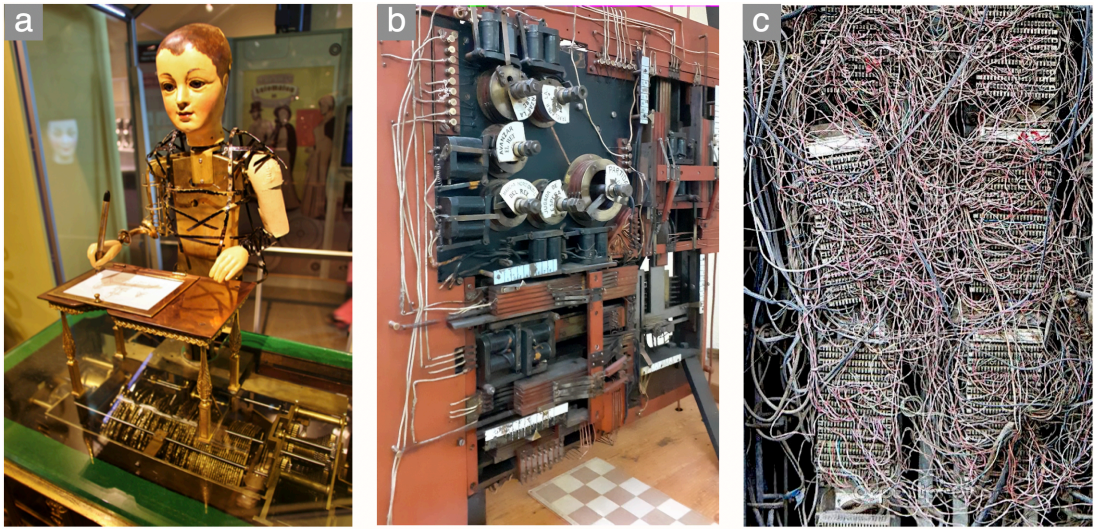


FIG. 1 **Technological metaphors used to describe diverse aspects of cognitive complexity before electronic computers**. (a) mechanical automata able to write using a set of connected gears that could be changed to execute diverse writing or drawing tasks. (b) Leonardo Torres-Quevedo 1910 prototype of his electromechanical chess-playing automaton. (c) Tangled network of interconnections in a telephone network board.

1950s (Valverde, 2016). In parallel with the development of the theory of computation, the first steps towards a theory of neural networks came to life, starting from early comparisons between brains and computers.

The first computers were plagued with problems associated to faulty units: vacuum tubes were prone to failure. Far from the reliable nature of brains, where neurons can die with no major consequences for the system-level performance, single-element failures could cause large disruptions. The comparative analysis between brains and computers (or computational analogies of brains) has been a recurrent topic since von Neumann's book *The computer and the brain* (1958). In the original formulation, the main problem was how to design reliable computers made of unreliable parts. Such approximation was largely forgotten within computer science with the rise of integrated circuits, although the problem became a central topic in the domain of neural networks. With the potential of *simulating* neural systems some of the early metaphors of memory involved strong analogies with magnetic materials. Such analogies would be developed in depth with the use of the statistical physics of spin glasses as computational substrates (Ackley et al., 1985; Hopfield, 1982). These similarities eventually provided the basis for the attractor picture that is now widespread.

Over the last decade, a new wave of excitement has emerged with the rise of Deep Learning networks (Kelleher, 2019). These descendants of the early multilayer neural networks developed in the 1990s have been accompanied with a considerable set of expectations (and hype). Because of their remarkable power to deal with specific problems far beyond the capacities of humans, claims have been repeatedly made suggesting that larger systems will eventually achieve cognitive skills similar (if not greater) than human brains, including consciousness or awareness[1]. However, as it has happened before many times (the winter-spring cycles of A.I.) artificially intelligent systems are still rather far from our general intelligence (Mitchell, 2019), and, indeed, they often appear brittle when taken even slightly out of their well-controlled closed worlds (Qu et al., 2020).

All this mimicry, inspiration and convergences bear some pressing questions: Do natural cognitive designs exhaust the space of the possible? Will every artificial cognitive solution ever found correspond to an earlier invention in nature? If so, then understanding natural cognition should be sufficient to learn everything that there is to know about intelligence and cognition. Might comprehending nature be necessary as well—i.e. does every cognitive solution respond to some natural challenge or feature that we need to understand in order to ultimately grasp cognition? If so, which is a minimal set of such challenges and features that can generate the range of cognitive designs? It is also possible that nature is not so restrictive and all-encompassing. This would leave a large elbow room for artificial cognition, human invention, and open-ended progress. Yet more wondrous questions are also put forward: What solutions might have been missed in the natural history of cognition? Are there artificial cognitive designs that cannot be reached by extant evo-

[1] See the recent stir caused by OpenAI's chief scientist, Ilya Sutskever, claiming that "it may be that today's large neural networks are slightly conscious". `https://lastweekin.ai/p/conscious-ai?s=r`.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



lutionary forces alone?

In this paper we argue that very relevant lessons can (and must) be obtained from a comparative analysis between evolved and designed cognitive systems. On the one hand, there are several non-trivial observations that suggest a limited repertoire of design principles that pervade and constrain the space of the possible: evolved and artificial architectures often converge. Secondly, there is a question regarding certain dynamical patterns exhibited by living systems that seldom appear in Artificial Neural Networks (ANN). Brains seem to operate close to critical states: is this a relevant trait to be considered when building artificial counterparts? Third, we will consider a list of attributes of human brains that define a gap between our species and any other living organism and we will see why A.I. systems might require to include evolutionary dynamics to get there.

## II. CONTINGENT VERSUS CONVERGENT EVOLUTION

Digital culture historian Kevin Kelly suggested in an essay on A.I. that future developments would allow us to create "artificial aliens" (Kelly, 2015). Specifically, Kelly conjectured that ongoing developments within this field will eventually create the conditions for new kinds of intelligences different from human ones:

> *Some traits of human thinking will be common (as common as bilateral symmetry, segmentation, and tubular guts are in biology), but the possibility space of viable minds will likely contain traits far outside what we have evolved. It is not necessary that this type of thinking be faster than humans, greater, or deeper. In some cases it will be simpler. Our most important machines are not machines that do what humans do better, but machines that can do things we can't do at all. Our most important thinking machines will not be machines that can think what we think faster, better, but those that think what we can't think.*

Such possibility is, from an evolutionary perspective, very appealing. The problem of how cognitive complexity emerged is a difficult one because behavior does not leave fossils (except in some limited and indirect fashion) and little can be said about intelligence. In this context, an alternative approach to standard comparative and phylogenetic approaches would be the study of "synthetic minds" resulting from engineering neural networks or evolvable robots (Solé, 2016). In a nutshell, by designing or evolving artificial alternatives to living matter, it could be possible perhaps to recreate the conditions for minds (and even consciousness) to emerge. In principle, it can be argued that multiple possibilities, perhaps including these "alien" minds pointed out by Kelly, might be found. Is that the case? Is there a space of endless possibilities inhabited by artificial intelligences orthogonal to those found in Nature?

Two extreme possibilities can be envisaged. In one, consistent with Kelly's picture, completely new forms of intelligence might be possible. This situation fits the picture of evolutionary change as a highly contingent process with many potential paths available. Contingency was particularly advocated by the late Stephen J. Gould (1990) who suggested that, if we would be able to re-run the tape of evolution, a completely different biosphere (and different minds) would be obtained. The human brain actually tells a story of tinkering associated to its nested origins.

However, the study of development reveals that very strong constraints might deeply limit the potential paths that can be followed. This is illustrated for example by the architecture of camera eyes that are found across many biological groups, from some single-cell organisms or jellyfish to cephalopods or vertebrates. A remarkable design principle is always at work despite their totally independent origins (Lane, 2009). If we look at the evolution of cognition in nature, what do we see? Do minds converge?

A remarkable observation from a comparative analysis of brain structures is that radically different topologies seem to share very similar functionalities (McGhee, 2011). A perfect illustration is provided by birds versus mammalian brains. Their early neural organization diverged 340 Myr ago, evolving in completely independent ways. And yet, their obviously different neural structures do not generate radically different minds (Emery & Clayton, 2004; McGhee, 2011; Prior et al., 2008). Such a convergence of minds is supported by the common traits of behavioral patterns such as associative learning, predator avoidance or decision making mechanisms that indicate the presence of a common cognitive toolkit (Keijzer, 2017; Powell et al., 2017). These commonalities in terms of cognitive repertoires could in fact be shared by a-neural systems (van Duijn, 2017). Divergent architectural brain designs are found all over the tree of life. Dolphins for example have brains that depart from primate ones, showing a combination of archaic features combined with a massively expanded cortex (Morris, 2003). But despite the differences, they display complex intelligence, communication skills and social life. Similarly, the brains of octopuses (members of the class of cephalopods that includes squids and cuttlefish) provide a further instance of convergent traits despite their being invertebrates (Godfrey-Smith, 2016). This group has evolved brains with multilayered cortical maps as well as a unique 8-fold extra neural clusters that autonomously control the arms. Because of their shape-shifting nature and distributed neural autonomy of the arms, these organisms have been often labelled as "alien" but they perform cognitive tasks similar to those displayed by other animals.

Behind each evolutionary convergence we can often find shared external pressures (e.g. the need to deal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



with a same niche) or, more fundamentally for cognition, a deep mathematical principle or constraint. Such key operating principles demand that certain solutions are found over and again by biological brains and their technological counterparts. Let us study some of these commonalities and the crucial computational principles that bring about such design convergences.

### A. Threshold units

The new wave of computing machines towards the mid XX century provided the right technological context to simulate logic elements similar to those present in nervous systems (Shannon, 1938). Theoretical developments within mathematical biology by Warren McCulloch and Walter Pitts revealed one first major result: it is possible to define units of cognition (*neurons*) under a logical framework (McCulloch & Pitts, 1943; Rashevsky, 1946, 1960). These formal neurons were described in terms of threshold units, largely inspired by the state-of-the-art knowledge of real excitable cells. Not surprisingly, the early works of Walter and Pitts had to do with threshold neurons and their mathematical description (Pitts, 1942). Over the last decades, major quantitative advances have been obtained by using a combination of neuron-inspired models with multilayer architecture and novel hardware improvements combined with massive use of training data. The growing understanding of single-neuron dynamics suggests that deeper computational complexity might be at work (Gidon et al., 2020)—but let us focus for a moment on the simplest models.

The pictorial conceptualization behind the McCulloch-Pitts model is sketched in Fig. 2a-b. The *formal neuron* shown here (Fig. 2b) is a simple Boolean system. Its state takes one of two values: $S_i \in \Sigma \equiv \{0, 1\}$ (a description of neurons as spins, $S_i \in \Sigma \equiv \{-1, +1\}$, is often convenient to derive potential energies for formal neural circuits). These two states are commonly associated to neurons resting (inactive) or firing (sending signals to others). Formal neurons react to incoming signals from a set of $N$ presynaptic units. Its response is a sudden activation if a weighted sum of the inputs is larger than a threshold (Hertz et al., 1991; Peretto, 1992; Rojas, 2013). While activation is all-or-nothing, weights, $\omega_{ij}$, are continuous and tell us how much the state of presynaptic neuron $j$ affects postsynaptic neuron $i$ (thus modeling the strength of connections). They can also be positive or negative, hence implementing excitation and inhibition. In the McCulloch-Pitts approach, postsynaptic neuron $S_i$ integrates incoming signals as:

$$S_i(t+1) \;=\; \sigma\left(\sum_{j=1}^{N} \omega_{ij} S_j(t) - \theta_i\right). \qquad (1)$$

The additional parameter, $\theta_i$, defines the neuron's threshold. The non-linear function $\sigma(x)$ is 1 if its argument is positive and 0 otherwise. Thus $S_i$ fires if the weighted sum of presynaptic inputs is larger than its threshold. The non-linearity introduced by $\sigma(\cdot)$ implements the all-or-none neural response. Alternative implementations use smooth step functions—e.g. $\sigma(x) = 1/(1+\exp(-\beta x))$, where $\beta$ (an inverse temperature) controls how much the non-linearity approaches the step function as $\beta \to \infty$.

McCulloch and Pitts crucially show that formal threshold neurons can build any logic Boolean circuit. A direct consequence is that brains, or at least their Boolean representation, can execute at least any logic operation that computers can perform. The elegant picture emerging from the McCulloch-Pitts model of a formal neuron is a powerful one. They broke new ground by showing that there was a neural analog to logic circuits and provide an important message concerning the power of brains as computational systems. Is this enough to get close to the complexity of brains? The development of ANN has revealed the enormous potential of so called semi symbolic artificial intelligence, but their achievements are only a tiny subset of the possible.

Are there alternative ways of designing cognitive networks that are not grounded in threshold-like units? This is a particularly relevant question, since advances in cognitive neuroscience and in particular artificial neural networks are indeed inspired by basic units exhibiting such kind of behavior. There is in fact another instance of threshold-like networks that have evolved in living systems: the webs of gene-gene interactions that rule the dynamics of cells. These so called gene regulatory networks have also a long tradition that starts in the aftermath of cybernetics and is grounded in a picture of gene-gene interactions similar to the McCulloch-Pitts model (Bornholdt, 2008; Glass & Kauffman, 1973; Kurten, 1988; Luque & Solé, 1997). Here, too, information exchanges are mediated by mechanisms that have a different molecular nature but share a fundamental commonality: responses are typically mediated by threshold-like response functions.

### B. Hierarchical processing of sensory inputs

The second half of the XX century saw quick advances regarding natural visual processing—from Hubel and Wiesel's identification of V1 neurons responding to light bars tilted at different angles (Hubel & Wiesel, 1959, 1962), to our modern understanding of visual cortical regions meshed in a complex network of hierarchical and parallel processes (Churchland & Sejnowski, 1994; Livingstone & Hubel, 1988; Rolls & Deco, 2007). We now have a rather complete view of how static visual stimuli are reconstructed: In the retina, edges are localized by ganglion cells that integrate information about spatial gradients of light (Levick, 1967; Russel & Werblin, 2010). This continues in the primary visual systems, where different neurons (like the ones discovered by Hubel and Wiesel) respond to bars of light of different sizes and at

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



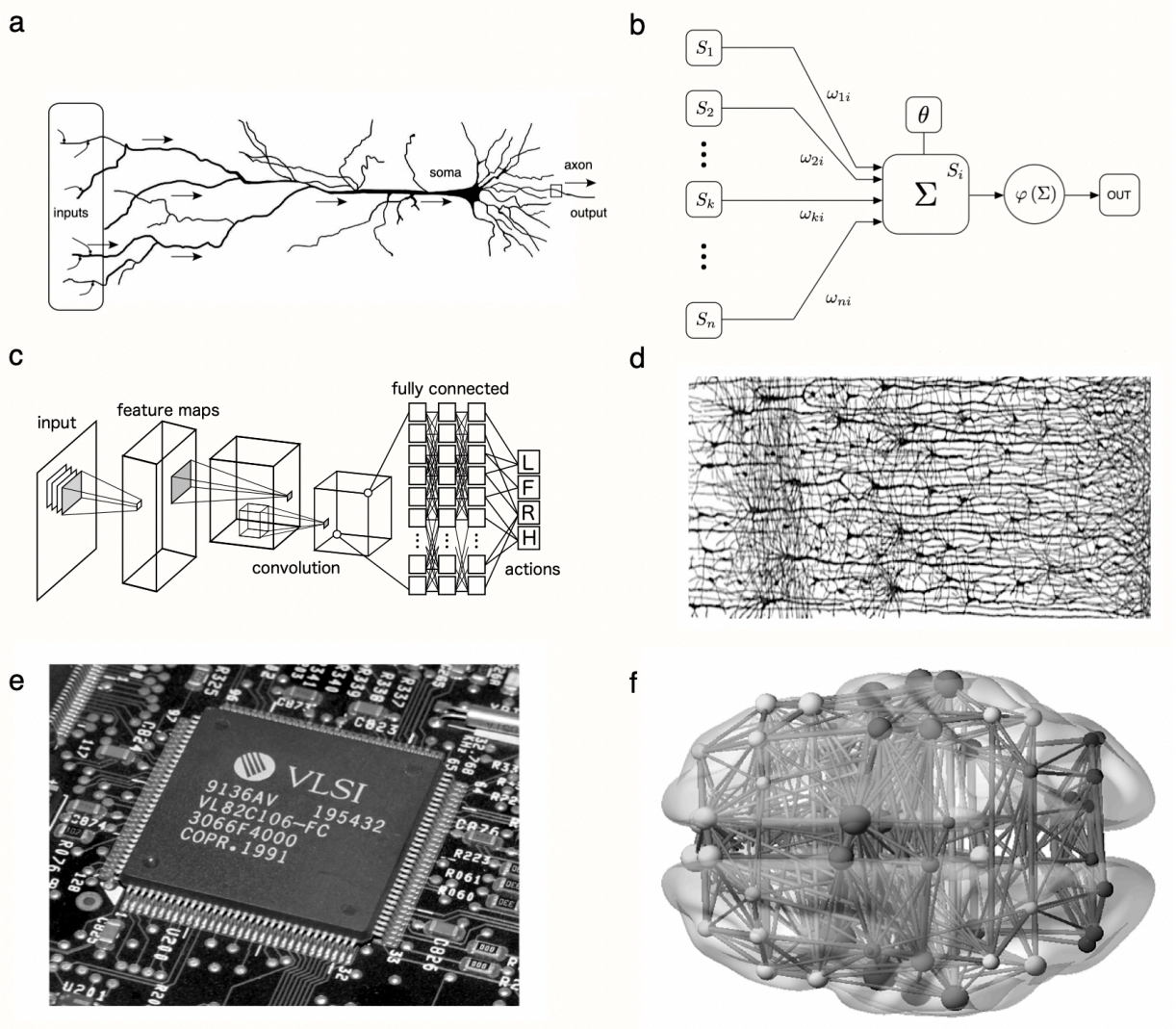


FIG. 2 **Convergent design principles in living and computational systems**. (a) A pyramidal neuron, to be compared with a neuron toy model suggested by McCulloch and Pitts (b) which preserves the minimal components at the logic level. (c) Complex tasks can be achieved by using layered ANN structures, characteristic of deep neural networks, which reminds us of the circuits found in the brain cortex (d). In both VLSI circuits (e) and brain connectomes (f) a nested hierarchy has been shown to exist, displaying common statistical laws, such as Rent's rule. This rule establishes that the amount of connections $C$ between elements in a sub-system of size $N$ with the rest of the system scales as a power law $C \sim N^p$, with $p \sim 0.8$ in both neural and VLSI circuits.

different locations across the visual field. These are building blocks of visual percepts that are latter assembled, in higher visual cortices, into geometrical shapes and, eventually (as color is added), into the refined objects that we perceive.

This process is paralleled by modern technology. By design, edge detection has long been mimicked by filters for image processing (Marr & Hildreth, 1980; Marr, 1982), and a hierarchical organization was capitalized by early ANN (Fukushima, 1988)—a choice inherited by state-of-the-art Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012; Yamins & DiCarlo, 2016). Not so much by design, learning algorithms shape the computational roles and organization of layers across these hierarchies, often leading to the convergence of individual receptor fields, e.g., of biological neurons and units of parallel computer vision systems (Nelson & Bower, 1990). Some modern CNN show a broader convergence with the biological visual processing pathway (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2020; Yamins et al., 2014; Zhuang et al., 2021). After CNN were trained in an object recognition tasks, individual units were identified whose activity correlated with that of real neurons in the ventral visual stream of humans performing the same task. This suggests that both natural and artificial systems converge on some computational steps necessary for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



visual recognition. Similar convergences are observed for auditory processing, as activity of CNN trained on auditory tasks can predict neural activity in the corresponding cortices (Kell et al., 2018), or for networks that predict fMRI and MEG responses during language processing tasks (Caucheteux & King, 2022). Cutting-edge developments in artificial visual systems incorporate ideas from natural language processing such as transformers or local context embedding (Thiry et al., 2021; Trockman & Kolter, 2022). Researchers are still struggling to understand precisely how these mechanisms operate, or why they achieve such high performances. Might these ideas actually depart from computational principles common across neural systems? While this possibility remains open, note that a hierarchical mode of operation seems a constant even in the most novel architectures.

What fundamental mathematical principles might underlie the evolutionary convergences just outlined? A possibility is that both brains and CNN are tapping into some essential, objective structure of input stimuli. Using models of interacting spins, Stephens et al. (2013) derive effective *statistical physics* of natural images, and find that edge-detecting filters (like the ones in the retina and early layers of CNN) are the simplest, most salient features. Might similar statistical relevance in input signals explain successive layers as we trade simplicity for salience? A way to test this is by applying Stephens's approach repeatedly, at several levels of the visual hierarchy—as implemented, e.g., for language features (Seoane & Solé, 2020a).

### C. Wiring cost universals

A very different kind of convergent design involves the presence of optimal wiring principles in both brain and Very Large Scale Integrated (VLSI) circuits. Some mathematical regularities emerge from demanding a high efficiency under strong packing constraints (Christie & Stroobandt, 2000). Vertebrate brains, and the human brain in particular, seem to present these regularities as well (Bassett et al., 2010). This close relationship between integrated circuits and neural systems is captured by the so-called Rent's rule, which defines a power law in networks that exhibit hierarchical modularity. Assuming that we partition the system into sub-systems of size $N$, the rule establishes that the number of connections $C$ linking the subset with the rest of the system scales as

$$C = \langle k \rangle N^p, \tag{2}$$

where $\langle k \rangle$ gives the average number of links per node, whereas $0 \leq p \leq 1$ is the so-called *Rent's exponent*. This is characteristic of fractal objects (where the basic pattern is repeated at different scales) and thus an indication of the presence of hierarchical order. When this method was applied to neural networks, a striking convergence was found. Both the neural web of the nematode *C. elegans* and human cortical maps shared a Rent's exponent $p$ close to what is the expected value for an optimally efficient hierarchical system. Such convergence, that shares many properties in common with VLSI circuits, illustrates the role played by cost constrains in promoting convergent designs (Moses et al., 2008, 2016).

State-of-the-art design of hardware networks and microchips is pushing the limits regarding space and other constraints—e.g. some circuits must operate within tolerable latencies (Howard et al., 2019). Machine learning techniques (e.g. reinforcement learning) might soon take over the manufacturing of new architectures (Mirhoseini et al., 2021). Will these new generations of designs follow Rent's rule as well? If this and other similar regularities emerge out of the underlying task (and not from the design process), we propose that convergences and (more interestingly) deviations from such laws would indicate whether the landscape of computational designs is being expanded in actual novel ways.

### D. A few building blocks of dynamical systems enable complex cognition

Studying Recurrent Neural Networks (RNN) and certain cortical circuits as dynamical systems suggests that simple mathematical principles underlie complex cognition as well. Attractors, saddle nodes, and limit cycles of such complex, high-dimensional systems constitute a dynamic backbone that guides neural activity towards low-dimensional manifolds. Forces behind cognitive processes become readily accessible—thus opening "the black box" (Sussillo & Barak, 2013).

The phase diagram in Fig. 3a shows an attractor (filled circle), a saddle node (empty circle), and an unstable fixed point (shaded circle) surrounded by a limit cycle (red). This depicts, qualitatively, the phase diagram of a leaky-integrate-and fire integrator neuron. In real neurons, their membrane potential is changed by currents injected from pre-synaptic axons, while recovery is driven by ion fluxes across the membrane that reset the neuron to its resting state (attractor). Noise or injected currents (of either sign) move the system around the attractor, sometimes towards the saddle node, which separates two dynamical regimes: at its left, the system returns to resting; at its right, dynamics are thrust around the limit cycle—a spike. Note how the dynamics around the saddle node are *attractive* along the vertical direction and repulsive along the horizontal one. For such integrator neurons, the saddle node mediates the "decision" of whether to spike or not by channeling trajectories along a single line that condenses all relevant information.

Early experiments on stimulation of real axons found that spiking behaviors always fell within a few classes (Hodgkin, 1948; Hodgkin & Huxley, 1952). It turns out that there are just a few ways in which attractors, saddle nodes, and limit cycles can interact in the phase diagrams of spiking neurons (FitzHugh, 1961; Izhikevich, 2000, 2007; Nagumo et al., 1962; Rinzel & Ermentrout,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



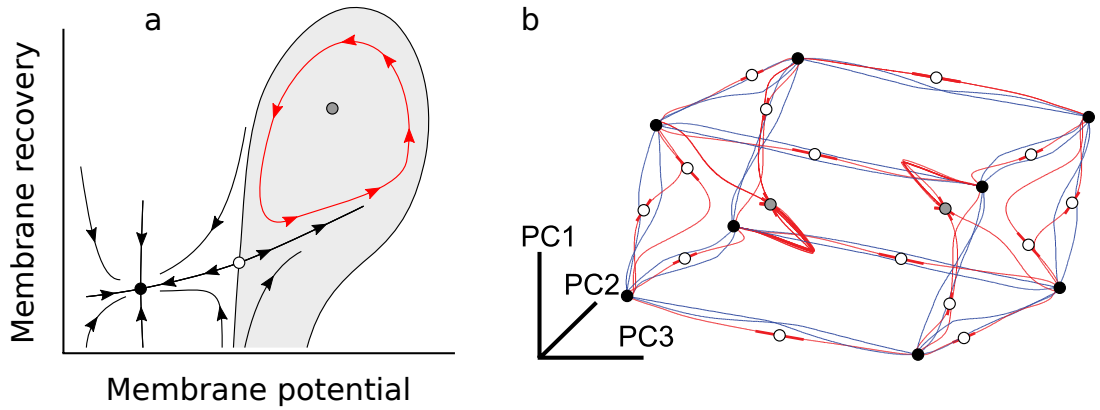


FIG. 3 **Attractors, saddle nodes, and repellers in the phase diagram of a dynamical system**. (a) Phase diagram of a spiking neuron. A saddle node attracts trajectories along the vertical direction and splits the horizontal one in two diverging trajectories. Thus, it mediates the decision of whether to spike or not. (Plot inspired by (Izhikevich, 2007).) (b) Attractors of the dynamics of a RNN partition the phase space into attractor basins that store $2^3$ bits of information. Saddle nodes mediate trajectories that diverge towards each attractor depending on the patterns of bits that need to be stored. (Panel adapted from (Sussillo & Barak, 2013).)

1998). These explain integrator and resonator neurons (the later responding to latency, rather than intensity, of input pulses), which pretty much exhaust the range of observed spiking classes (Izhikevich, 2007). Such simple decisions (whether to spike, whether to do it repeatedly, whether to do it at a fixed or variable firing rate) are thus mediated by a small dynamical alphabet.

Sussillo and Barak (2013) studied whether similar elements mediate decision making in RNN trained to perform different tasks. These networks are much more complex that spiking neurons, and their activity moves around high-dimensional spaces. They localized fixed points (both attractors and saddle nodes) numerically, and found that they partition the phase space according to a learned task. Fig. 3b shows the backbone of a RNN that stores the last input ($+1$ or $-1$) of three independent neurons (thus it must store $2^3$ patterns, each in a different attractor). In these and other RNN, saddle nodes tend to be stable along most orthogonal directions except one. Thus, as for the integrator neuron, they channel all unimportant degrees of freedom towards one-dimensional trajectories (depicted in Fig. 3b) that capture the relevant variables for the cognitive process.

Similar principles underlie some decision making processes in humans. Collective dynamics of cortical networks appear confined to low-dimensional manifolds (Gallego et al., 2017; Gardner et al., 2022; Gao & Ganguli, 2015; Yuste, 2015), suggesting a channeling of activity towards useful cognitive regions. By combining neural data on a context-dependent decision task with RNN models, similar dynamical backbones can be uncovered (Kaufman et al., 2014; Mante et al., 2013).

### E. Learning, learning to learn—meta-learning

Learning is a problem in which we saw impressive, yet irregular progress. The puzzle shows two distinct scales, from the microscopic to overarching principles, which we cannot fit yet in an all-encompassing picture. We know the details of memory formation in individual synapses (Kandel, 2007), which build associations using Hebbian ("fire together, wire together") reinforcement (Hebb, 1949). We also know several algorithms to implement similar plasticity in formal models of neural networks, with different strategies often giving rise to distinct cognitive frameworks (Ackley et al., 1985; Hopfield, 1982; Rumelhart et al., 1985). In these algorithms, global information about system (brain or ANN) performance is precisely deployed to individual synapses (Lillicrap et al., 2020). In the cornerstone *backpropagation algorithm* (Rumelhart et al., 1985), the chain rule of the derivative computes the causal contribution of each synaptic weight towards a global output error—thus connections can be precisely corrected. This method is incredibly successful, as shown by recent ANNs with super-human performance in complex games (Mnih et al., 2015; Moravčík et al., 2017; Silver et al., 2016, 2017).

While we understand functioning microscopic mechanisms and overarching principles for learning, we do not comprehend how learning descends from the complex, symbolic framework to the synaptic scale. There are serious issues concerning how the brain could implement the algorithms that solve these problems in machines (Lillicrap et al., 2020). Naive backpropagation would demand exact, symmetric backward paths to every synapse. Recent advances suggests solutions enabled by the generalizing and plastic abilities of the brain (Bellec et al., 2020; Lillicrap et al., 2020). Under certain circumstances, if we deliver *fake error derivatives* (as a

 doi:10.20944/preprints202205.0121.v1



fixed, randomly weighted combination of global errors) to synapses, they can first adapt themselves to making these randomly weighted signals useful, and then proceed with learning as usual (Lillicrap et al., 2016, 2020).

Problems become more pressing in Reinforcement Learning (RL) (Sutton & Barto, 2018). In RL, scenarios and rewards change as a response to agent actions. Feedback might arrive late in time. The problem of credit assignment affects not only individual synapses, but also representations of earlier, fading states. To implement RL, some form of time travel (see below), even if deeply unconscious, must take place. Evidence shows that, in mice, hippocampal place cells replay recent walks through a maze in reverse—a purported mechanism (and elegant solution) to assign credit to all states that contributed to the eventual reward (Foster, 2017; Foster & Wilson, 2006; Penagos et al., 2017).

The prefrontal cortex (PFC) seems a paramount site of RL in humans brains. Within a broader network (including basal ganglia and the thalamus), PFC implements meta-learning and generalization by combining two coupled loops of RL (Subramoney et al., 2021; Wang et al., 2018). An outer RL loop allows PFC to learn broad classes of models, while the internal loop allows PFC to converge upon a model within the class that explains current input. This convergence happens within very few examples—also for inputs derived from models within the family but not previously shown to the network. Furthermore, this convergence happens without further modifications to PFC weights. This implies that memory about a current model is not stored as synapse changes, but as fading states of the network dynamics. Hence, this mode of learning happens at the speed of network operation, not at the slower rate of network learning.

## III. THE GAP

The goal of building an artificial intelligence, as defined by the founders of the field, is far from being achieved. Great expectations were generated by breakthroughs in a diverse range of domains involving specific problem-solving. But the achievement of human-like intelligence, or some alternative cognitive system able to deal with some of the crucial ingredients associated to the gap, is far from close. Larger multilayered networks are not enough to get there, and one direction to look has to do with our initial discussion concerning evolutionary convergence. In this context, although evolutionary algorithms are used in A.I., evolutionary arguments are too often ignored within the A.I. community. In this section we consider some key qualitative traits that make human minds singular (Suddendorf, 2013), and how these features might constrain machine intelligence expectations. Tab. I summarizes the most salient similarities and differences, thus outlining our comparative approach to brains versus machine evolution.

### A. Language

Language is a complex system itself, and has a network organization that includes multiple interacting layers (Niyogi, 2006; Solé et al., 2010). Other species posses complex communication systems, but none of those shows recursivity—actually, no other species seems able to process recursively organized sequences (Dehaene et al., 2015). This feature ultimately confers open-ended expressive capabilities to human language (Bickerton, 1990).

It has been argued that language has not evolved as a communication means, though, but as an advanced representation system (Berwick & Chomsky, 2016; Hauser et al., 2002). This might have been triggered when facing, with an advanced brain, some evolutionary pressures common to other eusocial animals (Bickerton, 2014). Such pressures would demand displacement: the ability of a signal to represent arbitrary events not immediately present. This problem is solved by many species (e.g. ant pheromones indicate a non-present reward), but Bickerton argues that this feature, planted in the much richer hominid brain, would kick-start an irreversible process towards full-fledged language. Alternative, more gradualistic views of language evolution assign perhaps even greater roles to evolutionary pressures (Arbib et al., 2005; Berwick & Chomsky, 2016; Corballis, 2009; Rizzolatti & Arbib, 1998).

The final stage of language is irreducibly complex, and it hardly leaves intermediate fossils in evolution nor in development. The closest to a language fossil is a debated *proto-language* form that arises in individuals who are trained late, or that emerges as a chimera of cohabitating tongues (Bickerton, 1990, 2014). But this seems an unstable cognitive solution: children of proto-language speakers readily complete it into a full-fledged tongue. This again suggests an irreversible evolution as language complexity crossed some threshold—as argued by Bickerton.

As it evolved, language coopted, or tapped into, circuitry for sequence representation (Dehaene et al., 2015); auditory processing, including complex stimuli such as music (Koelsch, 2009); and motor control (Armstrong et al., 1995; Corballis, 2009; Rizzolatti & Arbib, 1998)—among others (Berwick & Chomsky, 2016; Bickerton, 2014). It also sprawled a semantic network present all across the neocortex (de Heer et al., 2017; Huth et al., 2012). The most prominent regions for language implementation sit usually at the left hemisphere, around the Sylvian fissure (Blank et al., 2016; Catani et al., 2005; Fedorenko & Kanwisher, 2009; Fedorenko et al., 2012; Fedorenko & Thompson-Schill, 2014; Geschwind, 1972), thus have ready access to vital cortices (auditory, motor, etc.). This neural substrate appears to be very similar across individuals and languages (Ayyash et al., 2022). Besides these commonalities in their neural implementation, different tongues seem to share other universal traits such as their productivity (Chomsky, 1986) and some ac-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



| | **Human Brains** | **NH vertebrate brains** | **Deep AN Networks** | **EVOL-neurorobotics** |
|---|---|---|---|---|
| Wiring | Hierarchical-nested | Hierarchical-nested | Feed-forward | FF, programmed |
| Basic units | Neurons | Neurons | Threshold units | Threshold units |
| Internal dynamics | Critical | Critical | Point attractors | Sensorimotor control |
| Time travel | Yes | Limited | None | None |
| Generalisation | Yes | Limited | No | No |
| Language | Syntactic | Simple | None | Proto-grammar |
| Meta-learning | Yes | Limited | Learning To learn | None |
| Mind readers | Yes | Limited | No | Emotion detector |
| Right $\neq$ wrong | Yes | Yes | Built Ethics | Built Ethics |
| Extended mind | Vast | Limited | No | Embodiment |
| Social Learning | Dominant | Limited | No | Imitation learning |

TABLE I **Comparative analysis of (i) human and (ii) non-human (NH) vertebrate brains, (iii) Deep Artificial Neural Networks (DANN) and (iv) evolved neurorobotic agents**. This table highlights the current chasm separating living brains from their computational counterparts. Each item in the non-human is intended to reflect a characteristic quality, which does not reflect the whole variability of this group (which is very broad). For the DANN and robotics columns, there is also large variability and our choice highlights the presence of the given property at least in one instance. As an example, the wiring of neural networks in neurobotic agents is very often feedforward, but the most interesting cases studies discussed here incorporate cortical-like, reentrant networks.

counts of efficiency (Gibson et al., 2019). Notwithstanding the purported universalities, linguistic codes throughout the world present an astonishing variety (Evans & Levinson, 2009).

Another feature common to all tongues across this huge diversity is ambiguity—a rather counter-intuitive trait. Animal communication codes are not ambiguous, as mistaken calls can be fatal (Bickerton, 1990). Computer languages cannot accept polysemous instructions either. And yet, ambiguity is ever present in human language (Solé & Seoane, 2015). A minimal model of communication codes that simultaneously optimize conflicting features suggests that ambiguity enables large expressive power with smaller vocabularies (Ferrer i Cancho & Solé, 2003; Seoane & Solé, 2018).

Ambiguity also enables semantic accessibility. Seman-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



tic networks connect words that are related through their meaning (usually, by being synonyms). They can be derived, e.g., from curated linguistic corpora (Fellbaum, 1998; Miller, 1995; Seoane & Solé, 2018) or from free-association experiments (Goñi et al., 2011). Semantic networks present a scale free and small world structure *provided that polysemy is included*. Scale free graphs are dominated by a few large hubs (central concepts that link to many others), while most words only have a few connections. This places some constraints on how language webs can be implemented in neural hardware (Solé et al., 2010; Steyvers & Tenenbaum, 2005), suggesting that a statistical regularity hides a relevant constraint of language evolution. Small world is defined by networks with a high clustering (i.e. abundant triplets of interrelated concepts—thus having abundant local connections) and small average distance between nodes. Thus, polysemy makes semantic networks easy to navigate through a search by association (Motter et al., 2002; Solé & Seoane, 2015).

Most approaches to implement artificial language attempt to manually hard-wire some overall computational, syntactic traits, or to infer grammars from large corpora. But alternatives exist that take seriously the relevance of Darwinian evolution in the origins of language. Notably, Luc Steels's *Talking Heads* experiment (1999a; 1999b; 2003; 2016) allowed to developed setups with embodied robots that converse about an external world. Steels capitalizes on Fluid Construction Grammars (Steels, 2011), a framework that includes ambiguity while managing combinatorial explosions—a key aspect of syntax. As robots exchange appreciations about their external world, their grammars, syntax, and semantics mature and their understanding of the environment becomes sharper.

It might be possible to build human-like language by design. But, if Bickerton's suggestion of an irreversible evolution under the appropriate circumstances are true, setting up evolutionary frameworks for artificial minds might ease the work. Alternatively, since artificial cognitive systems are different from humans, we can wonder what effects such evolutionary pressures might have on them—what kinds of communicative or representation systems such dynamics might bring about.

### B. Time travel

We are *time travelers*, able to locate ourselves in time by storing past events in multiple layers of detail while being able to imagine multiple alternative futures (Buonomano, 2017). The past is reached thanks to episodic memory (which is autobiographical in nature) and mixed evidence suggests that animals might have a rather limited capacity to remember personal episodes (Roberts, 2002; Suddendorf & Corvallis, 2007). No current artificial system has this ability; although many features such as goal-directed behavior, planning or causation require time representation.

The powerful capacity of brains to explore possible futures is not reducible to simple rules of forecasting, which were likely present in early stages of brain evolution. This ability seems enhanced (if not enabled) by the language capacity—the displacement property (to represent scenarios not currently available) is naturally extended into imagining futures. While non-human animals have a limited ability to plan the future, it does not come close to the human capacity that language brings about. In evolutionary terms, predicting future events has been a major force towards cognitive complexity: reducing environmental uncertainty can largely counterbalance the costs of a cognitive apparatus (Llinás, 1988; Seoane & Sole, 2018). Past recollection and generation of possible futures seem intimately connected, as same areas that are used to recall past events have been coopted to plan future events and ultimately to create alternative futures (Schacter et al., 2007).

So far, the time dimension of cognition is barely represented within neurorobotics, where research focuses mainly on the spatial extent of sensory information. The reason is the preeminent role played by information processing associated to sensory devices, whereas the role of time is limited to find out the next possible action. Implementing temporal cognition is being recognised as a missing component in neurorobotics (Maniadakis & Trahanias, 2011). In any case, early work on time representation in connectionist models already indicates that recurrent networks might be a necessary condition (Elman, 1990) and a precursor component of complex language use by a symbolic mind.

We expect Reinforcement Learning to be the branch that most early explores time traveling—as policies extend in time. In recent breakthroughs, RL agents first elaborate internal representations of their external world (Ha & Schmidhuber, 2018; Kaiser et al., 2019). This allows a limited forecasting, and even *dreaming* worlds over longer periods (Ha & Schmidhuber, 2018). This way, policies are better informed, improving performance. These models base their internal representations (and limited time travel) in the simplest correlations between sensory information (pixels in a screen over time). Meanwhile, human mental models include proper objects and agents causally related. It is the extrapolation of such causal relationships that enable our rich time travel experience.

### C. Mind reading

We can identify emotions and states of mind of others thanks to a set of specialized systems that evolved as part of our lineage's social nature. Face recognition processes are devoted ample, specialized regions in our brain (Kanwisher et al., 1997), along with a system of mirror neurons that respond to actions of others as if they were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



our own (Corballis, 2009; Ramachandran, 2012; Rizzolatti et al., 1996). Although mirror neurons are shared with other species, the consequences for humans are enormous. The capacity for reading minds was a crucial component in our evolution as a cooperative species: knowing the mind of others provides an effective way of making decisions relevant to group needs. And such mechanisms, likely to be the target of evolution, open the door to another remarkable trait: self-consciousness. Developing a theory of mind might inevitably create the conditions for "knowing" your own individual, distinct nature. In this view, self-consciousness would be a side effect of mind reading.

Thus, here too, evolutionary dynamics has played a central role in developing mechanisms for social interactions that are likely to be a major requirement to develop advanced artificial cognition. So far, emotions are detected by robotic agents able of visual pattern recognition and supervised learning. If evolution and social interactions are a requirement for self-awareness, the often discussed possibility of generating conscious A.I. might necessarily demand interactions among communicating agents. Within robotics, this requires the generation of internal representations that encode information about internal states of other agents—i.e. about information of the external world that is not readily available to sensory systems.

### D. Right from wrong

We have a moral mind, and evidence indicates that there is some hardwired tendency to make moral decisions right from early life. Cooperating moral minds have been helpful in fostering communities, and thus generate meaning under the context of social interactions (Tomasello & Vaish, 2013).

Building moral machines is a hot topic within both A.I. and robotics. Here, too, moral brains are evolved systems (Dennett, 1995; Harris, 2011), whereas machines require an explicit programming of moral norms (Gordon, 2020). While the goal of evolving machines that will avoid harming humans has been a recurrent topic within fictional works such as Asimov's rules of robotics[2] it becomes a pressing issue as autonomous robots are being deployed (Wallach & Allen, 2008). This connects inevitably with time travel and mind reading: moral decisions imply choices and understanding their implications for others. That means having a theory of mind, representing counterfactuals, future paths (as it occurs with the famous trolley experiment) and the implications of these actions.

### E. Extended mind

A major ingredient for the success of humans in their ecological conquest of the biosphere is related to their remarkable capacity for understanding and manipulating their environments. We are somehow limited by our senses or physiology, but none of these limitations really matters, since all of them can be overcome by the appropriate technological interface. This is part of what Clark and Chalmers dubbed *the extended mind* (Clark & Chalmers, 1998). Extended cognition starts with the embodied nature of agents and ends in the specific interfaces that they use to connect with their worlds. It can be found in very simple organisms. One very interesting example is provided by spiders, as their spiderwebs define a powerful example of niche-constructed structures that outsource information processing by means of an externalised structure (Japayasu & Laland, 2017). In this case, in which small brains are involved, external structures allow to reduce environmental uncertainty. Insect nests would be another successful example (Bonabeau et al., 1997). In this case, nests act as both engineered structures regulated in a self-organized fashion, and a vehicle for ecological engineering.

Little is found in their synthetic counterparts: beyond embodiment, robot-driven manipulation of their environments is almost absent. In silico counterparts based on reinforcement learning are the closest experiments in this direction, although limited to simulated environments (Gupta et al., 2021; Ha, 2019). A very recent, promising proposal uses the web as an embodied environment to improve artificial question-answering (Nakano et al., 2021).

### F. Social learning

Thanks to language and mind reading, and fostered by extended cognition, humans massively involve themselves in complex social interactions. One particularly relevant aspect of this is social learning: the extraordinary capacity to learn from others and being able to transmit information through teaching. A great deal of this is connected with imitation, which is unevenly distributed in non-human vertebrates. Songbirds or cetaceans display high levels of imitation, while non-human primates have rather limited skills (Hauser et al., 2002). Some authors suggest that this is a crucial attribute that is needed to create true human-like machines (Schaal, 1999). Social learning has been a very active domain within robotics, particularly within the context of human-robot interactions. A broad range of social robot scenarios can be defined (Fong et al., 2003), from ant-like robots to potential socially-intelligent agents (the latter within the domain

[2] As noted by Brooks (2003) and Mitchell (2019) Asimov's Three Rules of Robotics illustrate the non-trivial character of moral decisions. Because of the importance of context (or the environment), apparently well-established programmed rules can conflict with each other and create unexpected, and sometimes undesirable outcomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



of speculation). A specially relevant development in this area deals with the design of human-shaped robots able to learn facial expressions and react to them in meaningful ways (Breazeal & Brian, 2002; Breazeal, 2003).

What is required to move forward beyond imitation and emotion detection? Here we might need mind reading features and embodied interactions in order to properly create complex social exchanges. An important contribution in this direction is the work of Arbib and co-workers aimed at building agents that are explicitly equipped with a mirror neuron system (Arbib et al., 2014). In their model, which avoids the extra complexities of language, they consider gestures as the communication channel between virtual simulations of interacting agents. In this *dyadic* brain model, two agents interact by combining both (simulated) visual and manual interactions. In a nutshell, the model allows two agents to learn how to exchange, predict and react to each other's gestures in a ritualised manner. These are very promising steps towards linking social learning with mind reading. As suggested by some authors, these might have been crucial ingredients for major cognitive transitions in human evolution (Breazeal, 2003; Ramachandran, 2000), and the transition would be, once again, a qualitative shift.

We expect progress in 'mind reading', 'right from wrong', and 'social learning' to go hand by hand, as some of these lines can be preconditions or even foster advances for others. While there is still a relevant gap to the complexity of these traits in humans, some progress is being made with agents that use reinforcement learning in simulated games or virtual environments (Ecoffet & Lehman, 2021; Jaques et al., 2019; Rabinowitz et al., 2018).

## IV. A SPACE OF COGNITIVE COMPLEXITY

The set of properties displayed by human brains that define a cognitive gap can be represented in a unified way by means of a *morphospace*—i.e. a three-dimensional space that allows a mapping of all given case studies within a finite domain. By using such a space, we can locate the different systems and compare them. Additionally, the presence of voids (i.e. empty volumes lacking any candidate system) can provide evidence for constraints or forbidden evolutionary paths. This approach was first introduced within the context of morphological traits of shells (Raup, 1966) and has been later on widely used within Paleobiology and evolutionary biology (McGhee, 2006, 2011; Niklas, 2004; Tyszka, 2006), and in other different contexts including network science (Avena-Koenigsberger et al., 2015; Corominas-Murtra et al., 2013; Goñi et al., 2013) and computational neuroscience (Arsiwalla et al., 2017; Duong-Tran et al., 2021; Ollé-Vila et al., 2020; Seoane, 2019, 2020, 2021). Morphospaces provide us with a global picture of possible designs and how they relate to each other (whether they are distant or close) in a feature space. By making reasonable assumptions about relationships between features, we can still make some qualitative assessments about our systems of interest (Arsiwalla et al., 2017; Dennett, 2017; Seoane, 2019, 2020).

A morphospace of cognitive complexity is outlined in Fig. 4. We propose three axes:

1. **Computational complexity**: This needs to be understood as some measure over the tasks performed by each kind of agent. That would include memory, learning, decision making and other cognitive traits.

2. **Degree of autonomy**: This is a crucial attribute of adaptive complexity. We can define autonomy as "the property of a system that builds and actively maintains the rules that define itself, as well as the way it behaves in the world" (Ruiz-Mirazo & Moreno, 2012).

3. **Interactions between agents**: This third and no less relevant dimension might enable cognition capabilities that transcend the individual. Tight interactions between agents might be a pre-requisite for (or a consequence of) eusociality (Wilson, 2012), as they might enable a switch of the selective focus of Darwinian selection.

At the bottom-front of the morphospace we locate the region where artificial systems exist, whereas most living organisms populate the left vertical, high-autonomy wall. The bottom surface of this space includes all those systems that lack the social component—they need little interaction with others to sprawl their cognitive phenotype. Here we have several kinds of robots as well as mechanical automata, protocells and solitary organisms. A most obvious feature of our plot is that living and artificial systems appear separated by a gap that grows bigger as systems become more complex or more socially interactive. The divide reflects a fundamental difference between biological and artificial systems: the pressure of Darwinian selection and evolution that promote autonomy (as discussed in (Dawkins, 1976) in terms of selfishness) (Dennett, 2017; Haig, 2020). Composed replicative units are more complex, thus can support the propagation of their selves with enhanced internal computation that enables to predict ever more complex environments (Seoane & Sole, 2018). Due to evolution, this computational prowess must further protect autonomy—thus closing a reinforcing loop that necessarily pushes biological replicators towards the left wall of our morphospace.

Most artificial agents have not come together through Darwinian selection. Instead, they are typically designed or programmed to perform functions under environment-free scenarios, with some exceptions. The closest to large autonomy are embodied neurorobots (again, embodiment as a driver for true cognitive complexity) that are capable of sensing external cues, move in space and react in simple ways. Gray Walter's tortoise or Braitenberg's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



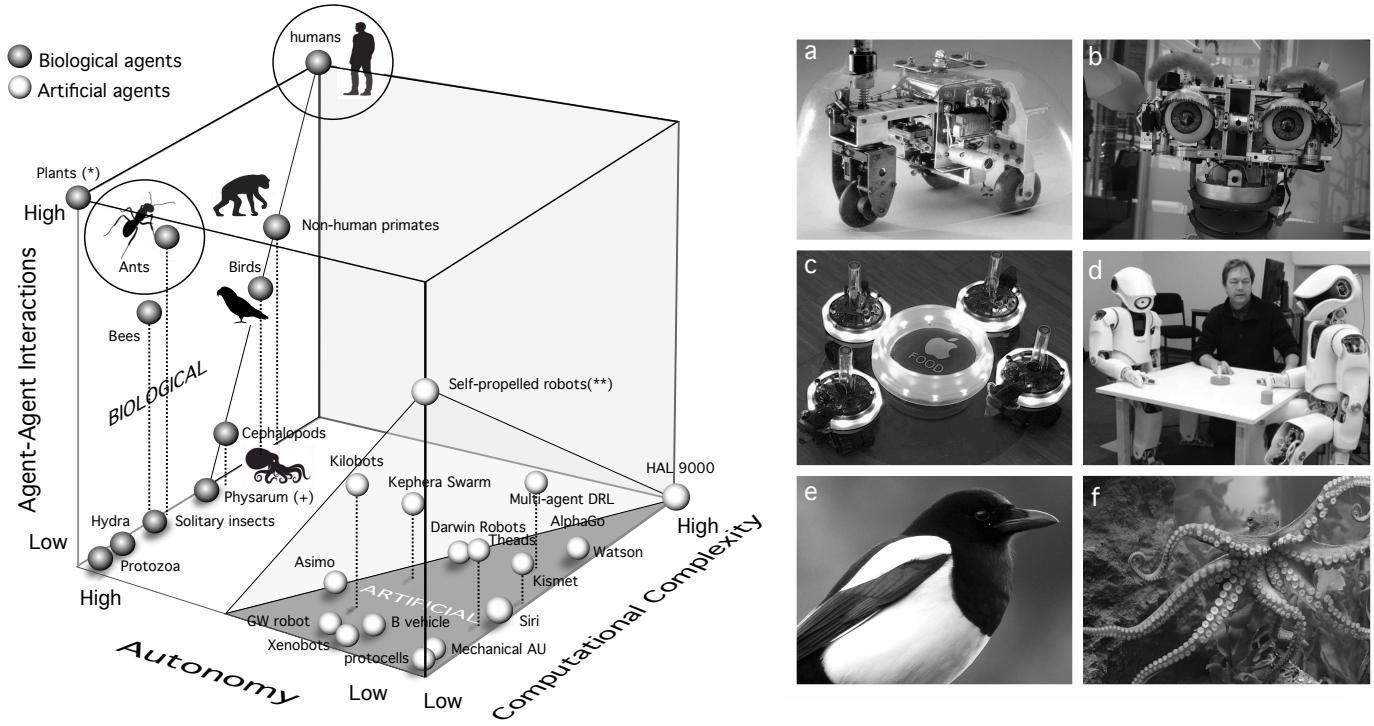


FIG. 4 **A Morphospace of cognitive complexity.** Autonomous, computational and social complexity constitute the three axes of this space. Human brains are located in the upper corner, scoring with maximal autonomy, computational complexity and agency. The examples shown here include both natural and artificial systems, as indicated. Plants (*) are located in the upper left corner since ecological interactions are known to play a key role, some of them by means of chemical communication exchanges. Current A.I. implementations cluster together in the high computation and low social complexity regime, with variable degrees of interaction-based rules (as it occurs with multi-agent Deep Reinforcement learning, DRL). Simple embodied systems displaying low computational complexity include mechanical automata, xenobots or Braitenberg vehicles. Another limit case here is provided by self-propelled robots (**) which are randomly moving, bullet-shaped agents carrying no sensors nor internal states, that interact physically leading sometimes to collective swarming patterns. The boundaries of this artificial subset (dark gray) are limited in the Autonomy direction by a boundary where several instances of mobile neurorobotic agents are located (such as Asimo, Kephera robots or different versions of Edelman's Darwin's robots). The left wall of high autonomy is occupied by living systems displaying diverse levels of social complexity. This includes some unique species such as Physarum (+) that involves a single-celled individual. On the right panels, six different particular case studies are highlighted: (a) Gray-Walter tortoise (a simple cybernetic mobile robot), (b) Kismet social robot able to detect and respond to emotions from visual cues, (c) Swarms of Kephera robots with evolvable neural networks, sensors and lights to communicate information, (d) talking heads experiment using humanoid robots (image courtesy of Luc Steels, in the image). Two examples of animal minds that share common complexities with human behavior, despite their marked differences in brain architecture are (e) magpies and (f) octopuses (image by Kelly Tarlton).

vehicles (GW robot and B vehicle in Fig. 4, left) are early precursors: they are electromechanical robots with a minimal sensorimotor architecture that allows them to respond to simple cues (approaching or avoiding lights or returning to the charging source). One great leap has been provided by the development of robots able to learn through experience using ANN, combining movement, visual processing and motor responses[3] (Seth et al., 2004).

As we move up from this surface and consider the role played by agent interactions, we also see a rather uneven distribution of case studies. Within the artificial domain, with few exceptions, interactions are lim-

[3] Darwin robots in particular have been developed under a set of design principles that are inspired by cortical architectures. In their implementation, these simulated cortical areas mimic reentrant neuroanatomic connections (an important feature that pervades high-level brain properties, including consciousness). Moreover, each area contains neuronal units that represent both activity levels and the timing of the activity of groups of neurons. As it occurs with real brains, neuronal synchronization allows neural binding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



ited to some particular features associated to the way they have been trained to perform very specific types of tasks (such as playing games). Two important exceptions are Luc Steel's Talking Heads (Steels, 2003) and swarm robotic systems such as Kilobots (Rubenstein et al., 2012; Slavkov et al., 2018). In the later, computational complexity is externalised while autonomy is required to making decisions associated to their relative location to others. One corner in this domain involves self-propelled robots, that have been extensively studied as a class of active matter (Deblais, 2018). They have zero computational complexity (they move by means of vibrations) and their random behavior in isolation discards non-zero autonomy since they have no real sensing of the external world. Another, isolated corner is represented by fictional entities (such as Asimov's robots or HAL9000) that would be pre-programmed to behave as intelligent agents without being exposed to social interactions.

This departure is highlighted in Fig. 5, where an additional space of possible cognitions is shown. Here the cognitive complexity dimension is completed by system size (how many agents define the group) and the role played by extended cognition (EC). Here we move beyond the boundaries of single multicellular agents and consider groups of agents of diverse kinds, from small primate groups to the massive, multi-queen ant colonies known as "supercolonies". The case study of ants is particularly relevant in terms of their ecological impact on the biosphere by means of an active modification of their environments, only comparable to the impact of humans. As E. O. Wilson points out, had humans failed to colonise the planet, the biosphere would be dominated by social insects (Wilson, 2012). However, in stark contrast with human or vertebrate brains (see Tab. I) ants are equipped with small brains and the cognitive power comes from the *collective intelligence* resulting from agent-agent interactions as well as with their capacity to build large-scale structures (Fig. 5a) that are several orders of magnitude larger than the individual size. These have inspired the development of simple robotic agents that build structures, such as the artificial termites (eTermites in the morphospace) in Fig. 5b (Werfel et al., 2014). Humans on the other hand can act as ecosystem engineers and exploit their EC on multiple scales, from single individuals to large collectives.

The large, empty void in the space is a reminder of the enormous distance from humans to any other species, as well as the lack of machine learning models of agents displaying EC. Some steps in this direction have been made, which are inspired in some key examples such as framing (Fig. 5c). Using a deep reinforcement learning system, set of agents can discover rule of cooperation that might recapitulate the early steps towards managing ecosystems before the emergence of agriculture (Perolat et al., 2017) (Fig. 5d). Finally, one remarkable example of EC is provided by the spiderwebs created by spiders having very small brains (Fig. 5e) that act as effective auditory sensors with a total surface that can be up to $10^4$ times larger than the individual spider. By building this externalized cognitive apparatus, individuals are released from body size constraints (Zhou et al., 2022). Although these structures can be generated by an evolutionary model of spider behavior (Fig. 5f) (Dawkins, 1977), they require a pre-specified set of rules of construction (and thus do not emerge *de novo* as an evolutionary innovation. Here again, the evolutionary emergence of the innovation represented by the spiderweb is far from the current state of the art of A.I. Crossing the empty land with to cognitive agents displaying complex extended mind remains a challenge.

Solving the problem of constructing a true A.I., as suggested by the structure of the morphospace, will require much more than cumulative innovations. As it occurs with evolution, new innovations might require major evolutionary transitions (Solé, 2016). These results are further expanded and summarized in Tab. I, where different features associated to cognitive complexity are presented for the main four groups of systems discussed here, namely human and non-human vertebrate brains as well as deep networks and neurorobotic agents.

## V. DISCUSSION

Can machines ever achieve true intelligence? In a recent paper entitled "Building machines that learn and think like people" (Lake et al., 2017) it has been argued that, for ANN to rapidly acquire generalisation capacities through learning-to-learn, some important components are missing. One is to generate context and improve learning by building internal models of intuitive physics. Secondly, intuitive psychology is also proposed as a natural feature present since early childhood (children naturally distinguish living from inanimate objects) which could be obtained by introducing a number of Bayesian approximations. Finally, compositionality is added as a way to avoid combinatorial explosions. In their review, Lake et al. discuss these improvements within the context of deep networks and problem-solving for video games, and thus consider the programming of primitives that enrich the internal degrees of freedom of the ANN. These components would expand the flexibility of deep nets towards comprehending causality. (See also the life-long work of Jürgen Schmidhuber for important developments over the last decades in the meta-lerning or learning-to-learn paradigms[4]) Lake et al. also point at several crucial elements that need to be incorporated, being language a prominent one. So far, despite groundbreaking advances in language processing, the computational counterparts of human language are very far from true language abilities. These improvements will no doubt create better

[4] https://people.idsia.ch/$\sim$juergen/metalearning.html.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2022 doi:10.20944/preprints202205.0121.v1



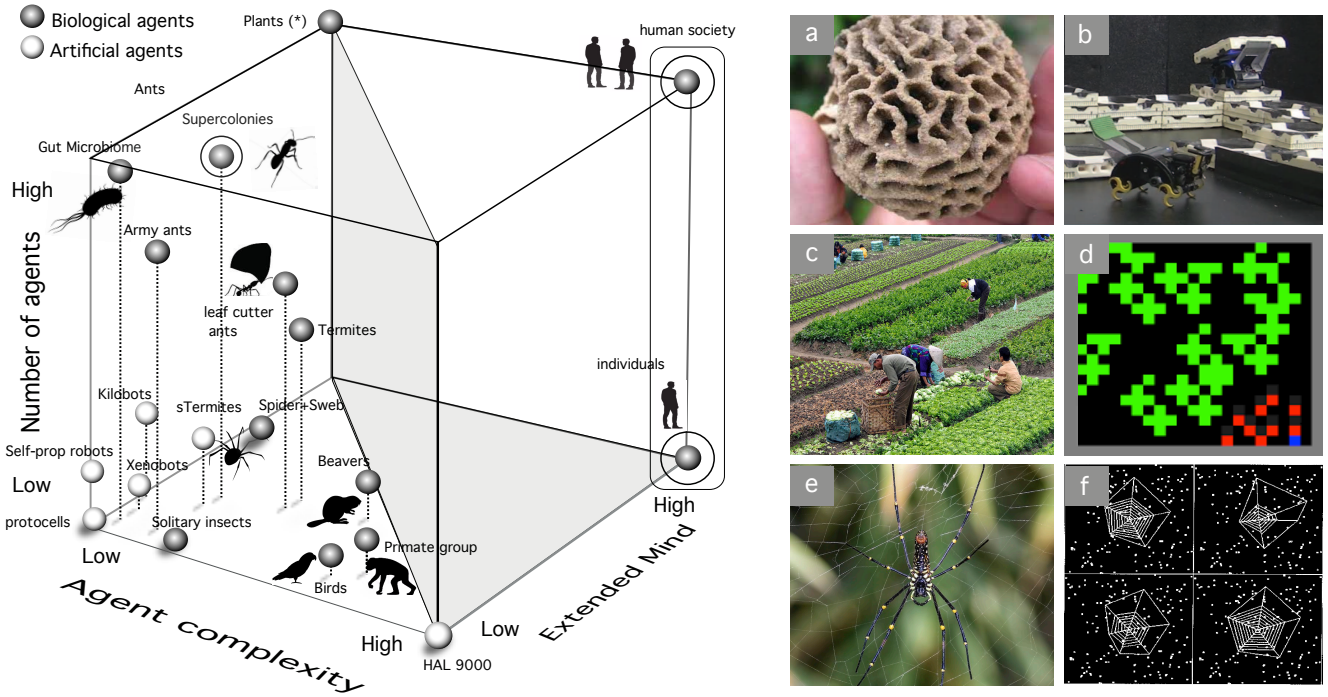


FIG. 5 **A Morphospace of extended minds.** A space of cognitive agents (left) can also be constructed by considering again the agent complexity axis (as in the previous figure) along with the number of agents involved in a given group as well as the role played by extended cognition (EC). The later includes the presence of active mechanisms of niche construction, ecosystem engineering or technological skills. The corner occupied by plants (*) involves (for example, in a forest) small computational power, massive populations and an "extended mind" that needs to be understood in terms of their interconnectedness and active modification of their environments, particularly soils. While ants in particular rival humans in their shear numbers and ecological dominance, the massive role played by externalized technology in humans makes them occupy a distinct, isolated domain in our cube (with both individuals and groups able to exploit EC). Some remarkable examples of EC are displayed on the right panels. Termites create complex spatial structures (a: fungus-growing chamber) and have inspired some swarm robotic systems (b) able to construct structures. Farming has been one of the ways humans have engineered their environments (c) and some deep multi-agent reinforcement learning models show how a collective of agents interacting with a given, limited resource can give rise to cooperative strategies (d). The dominant role of EC in some species beyond humans is illustrated by cobweb spiders (e). They are equipped with a tiny brain, but their spiderwebs act as sophisticated antennas, which allow for a powerful sensing and response to their environments. In this context, efficient cobwebs can be evolved using artificial evolution (f). The gap separating humans from the rest of natural and artificial systems is highlighted by the empty volume on the right, which needs to be explored by future models of artificial cognition.

imitations of thinking, but they are outside an embodied world where—we believe—true complex minds can emerge by evolution.

Are there alien minds? Yes and no. An affirmative answer emerges from the obvious: artificial systems do not need to follow biological constraints or Darwinian evolutionary paths. Being designed by humans or evolved within computers using ad hoc optimisation procedures, the final outcome can depart from biology in multiple ways. A deep network can outperform humans in a very specific task using a training algorithm based on feed-forward convolutional nets that, although inspired by experiments, lack the re-entrant loops that might be crucial to achieve true intelligence and awareness. Robotic agents can have behavioral patterns of response to complex environments, but the cognitive skills are externalised: algorithms are being executed in a rather powerful computer that resides somewhere else outside the body. But these are systems where agency plays a minor role. Perhaps the really relevant question is: Are there autonomous alien minds?

If convergent designs are an indication that there is a limited repertoire of possible neural architectures and cognitive autonomous agents, the future of A.I. is in the evolutionary arena. That means that the roads not taken necessarily cross the land of embodiment: as it occurs with naturally evolved systems, moving in an uncertain world was one of the engines of brain evolution. Moreover, another crucial component of evolutionary innovations is the emergence of new forms of cooperation. When cognitive agents are involved, that means evolving communication and dealing with information (Baluška & Levin, 2016). What kind of interesting phenomena can be observed using these two ingredients? Evolved

 doi:10.20944/preprints202205.0121.v1



robotic systems illustrate fairly well the ways in which evolutionary dynamics simultaneously link some of these components of cognition. As an example, robotic agents moving on a landscape where both positive and negative inputs (sources of charge and discharge, respectively) are located on given spots develop communication along with cooperative strategies that improve group fitness (Floreano et al., 2007; Mitri et al., 2009). Each robot is equipped with a set of sensors and lights and start foraging with a random configuration. A feedforward ANN allows evolving the interactions between sensors and lights and to generate communication among robots that allows for cooperation and altruism. Finding and avoiding positive and negative scenarios create the conditions for increasing group fitness. However, crowding also triggers cheating and deception (a familiar trait of evolution): robots can also evolve into lying to each-other. Despite the simple nature of the players, a combination of some key evolvable features can lead to unexpected insights.

As pointed out in the introduction, the paths that lead to brains seem to exploit common, perhaps universal properties of a handful of design principles and are deeply limited by architectural and dynamical constraints. Is it possible to create artificial minds using completely different design principles, without threshold units, multilayer architectures or sensory systems like those that we know? Since millions of years of evolution have led, through independent trajectories, to diverse brain architectures and yet not really different minds, we need to ask if the convergent designs are just accidents or perhaps the result of our constrained potential for engineering designs. Within the context of developmental constraints. evolutionary biologist Pere Alberch wrote a landmark essay that can further illustrate our point (Alberch, 1989). It was entitled "The Logic of Monsters" and presented compelling evidence that, even within the domain of theratologies, it is possible to perceive an underlying organization: far from a completely arbitrary universe of possibilities[5] there is a deep order that allows to define a taxonomy of "anomalies". Within our context, that would mean that the universe of alien minds might be also deeply limited.

## Author contributions

Original conceptualization: R.Solé. Both authors elaborated the research and wrote the manuscript.

[5] Since failed embryos are not the subject of selection pressures, it can be argued that all kinds of arbitrary morphological "solutions" could be observed.

## Funding

RS was supported by the Spanish Ministry of Economy and Competitiveness, grant FIS2016-77447-R MINECO/AEI/FEDER, and an AGAUR FI-SDUR 2020 grant. LFS is founded by the Spanish National Research Council (CSIC) and by the Spanish Department for Science and Innovation (MICINN) through a Juan de la Cierva Fellowship (IJC2018-036694-I). LFS developed his work at the Spanish National Center for Biotechnology (CNB at CSIC), a "Severo Ochoa" Center of Excellence funded by MICINN grant SEV 2017-0712.

## Acknowledgments

R Solé thanks Xerxes Arsiwalla, Jordi Delgado, Gemma de las Cuevas, Marti Sanchez-Fibla, Patrick Fraser and Jordi Piñero for useful discussions about brains and minds. Special thanks to the Santa Fe Institute for hosting his visit where the original conceptualisation of this work emerged at the Cormac McCarthy Library. We also thank Ephraim Winslow and Thomas Wake for enlightening ideas.

## References

Ackley DH, Hinton GE, Sejnowski TJ. 1985. A learning algorithm for Boltzmann machines. *Cognitive Sci.* **9**(1), pp.147-169.

Alberch P. 1989. The logic of monsters: evidence for internal constraint in development and evolution. *Geobios* **22**, pp.21-57.

Arbib MA. 2005. From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behav. Brain Sci.* **28**(2), pp.105-124.

Arbib M, Ganesh V, Gasser B 2014. Dyadic brain modelling, mirror systems and the ontogenetic ritualization of ape gesture. *Phil. Trans. R. Soc. B* **369**, p.20130414.

Armstrong DF, Stokoe WC, Wilcox SE. 1995. *Gesture and the nature of language*. Cambridge University Press.

Arsiwalla XD, Solé R, Moulin-Frier C, Herreros I, Sanchez-Fibla M, Verschure P. 2017. The morphospace of consciousness. arXiv preprint arXiv:1705.11190.

Avena-Koenigsberger A, Goñi J, Solé R, Sporns O. Network morphospace. *J. R. Soc. Interface* **12**(103), p.20140881.

Ayyash D, Malik-Moraleda S, Gallée J, Affourtit J, Hoffman M, Mineroff Z, Jouravlev O, Fedorenko E. 2022. The universal language network: A cross-linguistic investigation spanning 45 languages and 11 language families. bioRxiv: https://doi.org/10.1101/2021.07.28.454040.

Baluška F, Levin M. 2016. On having no head: cognition throughout biological systems. *Front. Psychol.* **7**, 902.

Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger DR, Moore SW, Bullmore ET. 2010. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* **6**(4), p.e1000748.

Bellec G, Scherr F, Subramoney A, Hajek E, Salaj D, Legenstein R, Maass W. 2020. A solution to the learning dilemma

for recurrent networks of spiking neurons. *Nat. Commun.* **11**(1), pp.1-15.
Benenson Y. 2012. Biomolecular computing systems: principles, progress and potential. *Nat. Rev. Genet.* **13**, 455-468.
Berwick RC, Chomsky N 2016. *Why only us: Language and evolution*. MIT press.
Bickerton D. 1990. *Language and Species*. University of Chicago Press.
Bickerton D. 2014. *More than nature needs*. Harvard University Press.
Blank I, Balewski Z, Mahowald K, Fedorenko E. 2016. Syntactic processing is distributed across the language system. *Neuroimage* **127**, pp.307-323.
Bonabeau E, Theraulaz G, Deneubourg JL, Aron S, Camazine S. 1997. Self-organization in social insects. *Trends. Ecol. Evol.* **12**(5), pp.188-193.
Bornholdt S. 2008. Boolean network models of cellular regulation: prospects and limitations. *J. R. Soc. Interface* **5**(1), pp.S85-S94.
Bray D. 1990. Intracellular signalling as a parallel distributed process. *J. Theor. Biol.* **143**, 215-231.
Breazeal C, Brian S. 2002. Robots that imitate humans. *Trends Cog. Sci.* **6**, 481-487.
Breazeal C. 2003. Toward sociable robots. *Robot. Auton. Syst.* **42**(3-4), pp.167-175.
Brooks RA. 2003. *Flesh and Machines: How Robots Will Change Us*. Vintage.
Buonomano D. 2017. *Your Brain Is a Time Machine: The Neuroscience and Physics of Time*. WW Norton & Company.
Catani M, Jones DK, Ffytche DH. 2005. Perisylvian language networks of the human brain. *Ann. Neurol.* **57**(1), pp.8-16.
Caucheteux C, King JR. 2022. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**(1), pp.1-10.
Chomsky N. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.
Christie P, Stroobandt D. 2000. The interpretation and application of Rent's rule. *IEEE T. VLSI Syst.* **8**(6), pp.639-648.
Churchland PS, Sejnowski TJ. 1994. *The computational brain*. MIT press.
Clark A. 1997. *Being there*. Cambridge, MA: MIT Press.
Clark, A. and Chalmers, D., 1998. The extended mind. *Analysis*, **58**, pp.7-19.
Corballis MC. 2009. Language as gesture. *Hum. Movement Sci.* **28**(5), pp.556-565.
Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C. 2013. On the origins of hierarchy in complex networks. *Proc. Nat. Acad. Sci.* **110**(33), pp.13316-13321 (2013).
Dawkins R. 1976. The selfish gene. Oxford University Press.
Dawkins, R., 1997. Climbing Mount Improbable. WW Norton and Company.
Deblais A, Barois T, Guerin T, Delville PH, Vaudaine R, Lintuvuori JS, Boudet JF, Baret JC, Kellay H. 2018. Boundaries control collective dynamics of inertial self-propelled robots. *Phys. Rev. Let.* **120**(18), p.188002.
de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE. 2017. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**(27), pp.6539-6557.
Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C. 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* **88**(1), pp.2-19.
DeFelipe J. 2011. The evolution of the brain, the human nature of cortical circuits, and intellectual creativity. *Front. Neuroanat.* **5**, p.29.
Dennett DC. 1995. *Darwin's dangerous idea*. Simon & Schuster.
Dennett DC. 2017. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company.
Phylogenetic origins of biological cognition: convergent patterns in the early evolution of learning. *Interface Focus*, **7**, 20160158.
Duong-Tran D, Abbas K, Amico E, Corominas-Murtra B, Dzemidzic M, Kareken D, Ventresca M, Goñi J. 2021. A morphospace of functional configuration to assess configural breadth based on brain functional networks. *Netw. Neurosci.* 5(3), pp.666-688 (2021).
Ecoffet A, Lehman J. 2021. Reinforcement learning under moral uncertainty. In International Conference on Machine Learning (pp. 2926-2936). PMLR.
Elman JL. 1990. Finding structure in time. *Cogn. Sci.* **14**(2), pp.179-211.
Emery NJ, Clayton NS. 2004. The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science* **306**, 1903-1907.
*The Cambrian Explosion*. Genwodd Village, Colorado: Roberts and Company.
Evans N, Levinson SC. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**(5), pp.429-448.
Fedorenko E, Kanwisher N. 2009. Neuroimaging of language: why hasn't a clearer picture emerged?. *Lang. Linguist.* **3**(4), pp.839-865.
Fedorenko E, Nieto-Castanon A, Kanwisher N. 2012. Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* **50**(4), pp.499-513.
Fedorenko E, Thompson-Schill SL. 2014. Reworking the language network. *Trends Cogn. Sci.* **18**(3), pp.120-126.
Fellbaum C. ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
Ferrer i Cancho R, Solé RV. 2003. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci.* **100**(3), 788–791 (2003).
FitzHugh R. 1961. Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1**(6), pp.445-466.
Floreano D, Mitri S, Magnenat S, Keller L. 2007. Evolutionary conditions for the emergence of communication in robots. *Curr. biol.* **17**(6), pp.514-519.
Fong T, Nourbakhsh I, Dautenhahn K. 2003. A survey of socially interactive robots. *Robot. Auton. Syst.* **42**(3-4), pp.143-166.
Foster DJ. 2017. Replay comes of age. *Annu. Rev. Neurosci.* **40**, pp.581-602.
Foster DJ, Wilson MA. 2006. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**(7084), pp.680-683.
Friston K. 2018. Does predictive coding have a future? *Nat Neurosci.* **21**(8), pp.1019-1021.
Fukushima K. 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks* **1**(2), pp.119-130.
Gallego JA, Perich MG, Miller LE, Solla SA. 2017. Neural manifolds for the control of movement. *Neuron* **94**(5), pp.978-984.
Gardner RJ, Hermansen E, Pachitariu M, Burak Y, Baas NA,

Dunn BA, Moser MB, Moser EI. 2022. Toroidal topology of population activity in grid cells. *Nature*, pp.1-6.

Gao P, Ganguli S. 2015. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, pp.148-155.

Geschwind N. 1972. Language and the brain. *Sci. Am.* **226**(4), pp.76-83.

Gibson E, Futrell R, Piantadosi SP, Dautriche I, Mahowald K, Bergen L, Levy R. 2019. How efficiency shapes human language. *Trends Cogn. Sci.* **23**(5), pp.389-407.

Gidon A, Zolnik TA, Fidzinski P, Bolduan F, Papoutsi A, Poirazi P, Holtkamp M, Vida I, Larkum ME. 2020. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* **367**(6473), pp.83-87.

Glass L, Kauffman SA. 1973. The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* **39**(1), pp.103-129.

Godfrey-Smith P. 2016. *Other minds: The octopus and the evolution of intelligent life*. London: William Collins.

Goñi J, Arrondo G, Sepulcre J, Martincorena I, Vélez de Mendizábal N, Corominas-Murtra B, Bejarano B, Ardanza-Trevijano S, Peraita H, Wall DP, Villoslada P. 2011. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cogn. Process.* **12**(2), pp.183-196.

Goñi J, Avena-Koenigsberger A, Velez de Mendizabal N, van den Heuvel MP, Betzel RF, Sporns O. 2013. Exploring the morphospace of communication efficiency in complex networks. *PLoS One* **8**(3), p.e58070 (2013).

Gordon JS. 2020. Building moral robots: ethical pitfalls and challenges. *Sci. Eng. Ethics* **26**(1), pp.141-157.

Gould SJ. 1990. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company.

Gupta A, Savarese S, Ganguli S, Fei-Fei L. 2021. Embodied intelligence via learning and evolution. *Nat. Commun.* **12**(1), pp.1-12.

Ha D. 2019. Reinforcement learning for improving agent design. *Artif. Life* **25**(4), pp.352-365.

Ha D, Schmidhuber J. 2018. World models. arXiv preprint arXiv:1803.10122.

Haig D. 2020. *From Darwin to Derrida: selfish genes, social selves, and the meanings of life*. Mit Press.

Harris S. 2011. *The moral landscape: How science can determine human values*. Simon and Schuster.

Hauser MD, Chomsky N, Fitch WT. 2002. The faculty of language: what is it, who has it, and how did it evolve?. *Science* **298**(5598), pp.1569-1579.

Hebb DO. 1949. *The Organization of Behavior*. New York: Wiley & Sons.

Hertz J, Krogh A, Palmer RG. 1991. *Introduction to the theory of neural computation. Vol I*. Addison-Wesley, Redwood City, CA, USA (1991).

Hodgkin AL. 1948. The local electric changes associated with repetitive action in a non-medullated axon. *J. Physiol.* **107**(2), p.165.

Hodgkin AL, Huxley AF. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**(4), p.500.

Hopfield JJ. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554-2558.

Hopfield JJ. 1994. Physics, computation and why biology looks so different. *J. Theor. Biol.* **171**(1), 53-60.

Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV. 2019. Searching for mobilenetv3. In *Proc. IEEE Int. Conf. Comput. Vis*, pp. 1314-1324.

Hubel DH, Wiesel TN. 1959. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**(3), p.574.

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. **160**(1), p.106.

Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**(6), pp.1210-1224.

Izhikevich EM. 2000. Neural excitability, spiking and bursting. *Int. J. Bifurcat. Chaos* **10**(06), pp.1171-1266.

Izhikevich EM. 2007. *Dynamical systems in neuroscience*. MIT press.

Jacob F. 1998. *Of flies, mice, and men*. Harvard University Press.

Japayasu HF, Laland KN. 2017. Extended spider cognition. *Anim. Cogn.* **20**, 375-395.

Jaques N, Lazaridou A, Hughes E, Gulcehre C, Ortega P, Strouse DJ, Leibo JZ, De Freitas N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In International Conference on Machine Learning (pp. 3040-3049). PMLR.

Jonas E, Kording KP. 2017. Could a neuroscientist understand a microprocessor? *PLoS Comput. Biol.* **13**(1), p.e1005268.

Kaiser L, Babaeizadeh M, Milos P, Osinski B, Campbell RH, Czechowski K, Erhan D, Finn C, Kozakowski P, Levine S, Mohiuddin A. 2019. Model-based reinforcement learning for atari. arXiv preprint arXiv:1903.00374.

Kandel ER. 2007. *In search of memory: The emergence of a new science of mind*. WW Norton & Company.

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17** (11), 4302–11.

Kaufman MT, Churchland MM, Ryu SI, Shenoy KV. 2014. Cortical activity in the null space: permitting preparation without movement. *Nature Neurosci.* **17**(3), pp.440-448.

Evolutionary convergence and biologically embodied cognition. *Interface Focus*, **7**, 20160123.

Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**(3), pp.630-644.

Kelleher JD. 2019. *Deep learning*. MIT press.

Kelly K. 2015. What do you think about machines that think? Edge, retrieved: https://www.edge.org/response-detail/26097.

Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**(11), p.e1003915.

Koelsch S. 2009. Neural substrates of processing syntax and semantics in music. Music that works, pp.143-153.

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. *Adv. Neur. In.* 25.

Kurten KE. 1988. Correspondence between neural threshold networks and Kauffman Boolean cellular automata. *J. Phys. A* **21**(11), p.L615.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. Building machines that learn and think like people. *Behav.*

*Brain Sci.* **40**.
Lane, N., 2009. *Life ascending: the ten great inventions of evolution*. Norton and Co. New York,
Levick WR. 1967. Receptive fields and trigger features of ganglion cells in the visual streak of the rabbit's retina. *J. Physiol.* **188**(3), p.285.
Levin M, Dennett DC. 2020. Cognition all the way down. *Aeon Essays*.
Lillicrap TP, Cownden D, Tweed DB, Akerman CJ. 2016. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7**(1), pp.1-10.
Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. 2020. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**(6), pp.335-346.
Livingstone M, Hubel D. 1988. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240**(4853), pp.740-749.
Llinas RR. 2001. *I of the vortex: from neurons to self*. Cambridge, MA: MIT Press.
Luque B, Solé RV. 1997. Phase transitions in random networks: simple analytic determination of critical points. *Physical Review E* **55**(1), p.257.
Maniadakis M, Trahanias P. 2011. Temporal cognition: a key ingredient of intelligent systems. *Front. Neurorobotics* **5**, p.2.
Mante V, Sussillo D, Shenoy KV, Newsome WT. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**(7474), pp.78-84.
Marr D, Hildreth E. 1980. Theory of edge detection. *P. R. Soc. London B* **207**(1167), pp.187-217.
Marr D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.
McGhee GR. 2006. *The geometry of evolution: adaptive landscapes and theoretical morphospaces*. Cambridge University Press.
McGhee GR. 2011. *Convergent evolution: limited forms most beautiful*. MIT Press.
McCulloch WS, Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**(4), pp.115-133.
Miller GA. 1995. WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39-41.
Mirhoseini A, Goldie A, Yazgan M, Jiang JW, Songhori E, Wang S, Lee YJ, Johnson E, Pathak O, Nazi A, Pak J. 2021. A graph placement methodology for fast chip design. *Nature* **594**(7862), pp.207-212.
Mitchell M. 2019. *Artificial intelligence: A guide for thinking humans*. Penguin UK.
Mitri S, Floreano D, Keller L. 2009. The evolution of information suppression in communicating robots with conflicting interests. *Proc. Natl. Acad. Sci.* **106**(37), pp.15786-15790.
Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S. 2015. Human-level control through deep reinforcement learning. *Nature* **518**(7540), pp.529-533.
Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, Davis T, Waugh K, Johanson M, Bowling M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**(6337), pp.508-513.
Morris SC. 2003. *Life's solution: inevitable humans in a lonely universe*. Cambridge university press.
Moses, M.E., Forrest, S., Davis, A.L., Lodder, M.A. and Brown, J.H., 2008. Scaling theory for information networks. J. Royal Soc. Interface, **5**(29), pp.1469-1480.
Moses M, Bezerra G, Edwards B, Brown J, Forrest S. 2016. Energy and time determine scaling in biological and computer designs. *Phil. Trans. R.Soc. B* **371**(1701), p.20150446.
Motter AE, De Moura AP, Lai YC, Dasgupta P. 2002. Topology of the conceptual network of language. *Phys. Rev. E* **65**(6), p.065102.
Nagumo J, Arimoto S, Yoshizawa S. 1962. An active pulse transmission line simulating nerve axon. *Proc. IRE* **50**(10), pp.2061-2070.
Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, Hesse C, Jain S, Kosaraju V, Saunders W, Jiang X. 2021. WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.
Nelson ME, Bower JM. 1990. Brain maps and parallel computers. *Trends Neurosci.* **13**(10), pp.403-408.
Niklas KJ. 2004. Computer models of early land plant evolution. *Annu. Rev. Earth Planet. Sci.* **32**, pp.47-66 (2004).
Niyogi P. 2006. *The computational nature of language learning and evolution*. Cambridge, MA: MIT press.
Ollé-Vila A, Seoane LF, Solé R. Ageing, computation and the evolution of neural regeneration processes. *J. R. Soc. Interface* 17(168), p.20200181 (2020).
Penagos H, Varela C, Wilson MA. 2017. Oscillations, neural computations and learning during wake and sleep. *Curr. Opin. Neurobiol.* **44**, pp.193-201.
Peretto P. 1992. *An introduction to the modeling of neural networks*. Cambridge University Press.
Perolat, J., Leibo, J.Z., Zambaldi, V., Beattie, C., Tuyls, K. and Graepel, T., 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems*, 30.
Pinero J, Solé R. 2019. Statistical physics of liquid brains. *Philos. T. R. Soc. B* **374**(1774), p.20180376.
Pitts W. 1942. Some observations on the simple neuron circuit. *Bull. Math. Biophys.* **4**, 121-129.
Powell, R., Mikhalevich, I., Logan, C. and Clayton, N.S., 2017. Convergent minds: the evolution of cognitive complexity in nature. *Interface Focus*, **7**, 20170029.
Prior H, Schwarz A, Güntürkün O. 2008. Mirror-induced behavior in the magpie (Pica pica): evidence of self-recognition. *PLoS Biol.* **6**(8), p.e202.
Qu X, Sun Z, Ong YS, Gupta A, Wei P. 2020. Minimalistic attacks: How little it takes to fool deep reinforcement learning policies. *IEEE Trans. Cogn. Develop. Syst.* **13**(4), pp.806-817.
Ramachandran VS. 2012. *The Tell-Tale Brain: Unlocking the Mystery of Human Nature*. Random House.
Ramachandran VS. 2000. Mirror neurons and imitation learning as the driving force behind "the great leap forward" in human evolution. *Edge*, **29**.
Rao RP, Ballard DH. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**(1), pp.79-87.
Rashevsky N. 1946. The neural mechanism of logical thinking. *Bull. Math. Biophys.* **8**(1), pp.29-40.
Rashevsky N. 1960. *Mathematical biophysics: physico-mathematical foundations of biology*. (3rd Ed.) Dover P., Inc., New York, NY, USA.
Raup DM. 1966. Geometric analysis of shell coiling: general problems. *J. Paleontol.*, pp.1178-1190.
Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M. 2018. Machine theory of mind. In Inter-

national conference on machine learning (pp. 4218-4227). PMLR.
Rinzel J, Ermentrout GB. 1998. Analysis of neural excitability and oscillations. In *Methods in Neuronal Modeling*. Koch C, Segev I (eds.). MIT Press, Cambridge, Mass, pp.251-292.
Rizzolatti G, Arbib MA. 1998. Language within our grasp. *Trends Neurosci.* **21**(5), pp.188-194.
Rizzolatti G, Fadiga L, Gallese V, Fogassi L. 1996. Premotor cortex and the recognition of motor actions. *Cognitive Brain Res.* **3**(2), pp.131-141.
Roberts WA. 2002. Are animals stuck in time? *Psychol. Bull.* **128**(3), p.473.
Rojas R. 2013. *Neural networks: a systematic introduction*. Springer.
Rolls ET, Deco G. 2007. *Computational neuroscience of vision*. Oxford university press.
Rubenstein M, Ahler C, Nagpal R. 2012. A low cost scalable robot system for collective behaviors. In *2012 IEEE international conference on robotics and automation* (pp. 3293-3298).
Ruiz-Mirazo K, Moreno A. 2012. Autonomy in evolution: from minimal to complex life. *Synthese* **185**(1), pp.21-52.
Rumelhart DE. Hinton GE, Williams RJ. 1985. Learning internal representations by error propagation. Technical Report (DTIC Document, 1985).
Russell TL, Werblin FS. 2010. Retinal synaptic pathways underlying the response of the rabbit local edge detector. *J. Neurophysiol.* **103**(5), pp.2757-2769.
Schaal S. 1999. Is imitation learning the route to humanoid robots?. *Trends Cogn. Sci.* **3**(6), pp.233-242.
Schacter DL, Addis DR, Buckner RL. 2007. Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* **8**(9), pp.657-661.
Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Geiger F, Schmidt K. 2020. Brain-score: Which artificial neural network for object recognition is most brain-like?. BioRxiv, p.407007.
Shannon CE. 1938. A symbolic analysis of relay and switching circuits. *Electr. Eng.* **57**(12), pp.713-723.
Seoane LF. 2019. Evolutionary aspects of reservoir computing. *Philos. T. R. Soc. B* **374**(1774), p.20180377 (2019).
Seoane LF. 2020. Fate of Duplicated Neural Structures. *Entropy* **22**(9), p.928.
Seoane LF. 2021. Evolutionary paths to lateralization of complex brain functions. arXiv preprint arXiv:2112.00221.
Seoane LF, Solé R. 2018. The morphospace of language networks. *Sci. Rep.* **8**(1), pp.1-14.
Seoane LF, Solé RV. 2018. Information theory, predictability and the emergence of complex life. *Roy. Soc. Open Sci.* **5**(2), p.172221.
Seoane LF, Solé R. 2020. Criticality in pareto optimal grammars?. *Entropy* **22**(2), p.165.
Seth AK, McKinstry JL, Edelman GM, Krichmar JL. 2004. Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. *Cereb. Cortex* **14**(11), p.1185-1199.
Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587), pp.484-489.
Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y. 2017. Mastering the game of go without human knowledge. *Nature*, **550**(7676), pp.354-359.
Slavkov I, Carrillo-Zapata D, Carranza N, Diego X, Jansson F, Kaandorp J, Hauert S, Sharpe J. 2018. Morphogenesis in robot swarms. *Sci. Robot.* **3**(25), p.eaau9178.
Solé RV. 2016. Synthetic transitions: towards a new synthesis. *Phil. Trans. R. Soc. B* **371**, 20150438.
Solé R, Moses M, Forrest S. 2019. Liquid brains, solid brains. *Phil. Trans. R. Soc. B* **374**(1774), p.20190040.
Solé R, Seoane LF. 2015. Ambiguity in language networks. *Linguist. Rev.* **32**, 5-35.
Solé RV, Corominas-Murtra B. Valverde S, Steels L. 2010. Language networks: Their structure, function, and evolution. *Complexity* **15**(6), pp.20-26.
Steels L. 1999. *The Talking Heads Experiment*.
Steels L. 1999. *The talking heads experiment: Words and Meanings*. VUB Artificial Intelligence Laboratory.
Steels L. 2003. Evolving grounded communication for robots. *Trends Cogn. Sci.* **7**(7), pp.308-312.
Steels L. ed. 2011. *Design patterns in fluid construction grammar*. John Benjamins Publishing.
Steels L. 2016. Agent-based models for the emergence and evolution of grammar. *Phil. Trans. R. Soc. B* **371**(1701), p.20150447.
Stephens GJ, Mora T, Tkačik G, Bialek W. 2013. Statistical thermodynamics of natural images. *Phys. Rev. Let.* **110**(1), p.018701.
Steyvers M, Tenenbaum JB. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Sci.* **29**(1), pp.41-78.
Subramoney A, Scherr F, Maass W. 2021. Reservoirs learn to learn. In *Reservoir Computing* (pp. 59-76). Springer, Singapore.
Suddendorf T. 2013. *The gap: The science of what separates us from other animals*. Constellation.
Suddendorf T., Corvallis, 2013. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioural. Brain Sci.* **30**, pp.299-313.
Sussillo D, Barak O. 2013. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**(3), pp.626-649.
Sutton RS, Barto AG. 2018. *Reinforcement learning: An introduction*. MIT press.
Thiry L, Arbel M, Belilovsky E, Oyallon E. 2021. The unreasonable effectiveness of patches in deep convolutional kernels methods. arXiv preprint arXiv:2101.07528.
Trockman A, Kolter JZ. 2022. Patches Are All You Need? arXiv preprint arXiv:2201.09792.
Tomasello M, Vaish A. 2013. Origins of human cooperation and morality. *Ann. Rev. Psychol.* **64**, pp.231-255.
Tyszka J. 2006. Morphospace of foraminiferal shells: results from the moving reference model. *Lethaia* **39**(1), pp.1-12 (2006).
Valverde S. 2016. Major transitions in information technology. *Phil. Trans. R.Soc. B* **371**(1701), p.20150450.
von Neumann J. 1958. *The computer and the brain*. Yale university press.
Wallach W, Allen C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Nat. Acad. Sci.* **111**(23), pp.8619-8624.
Yamins DL, DiCarlo JJ. 2016. Using goal-driven deep learning

 doi:10.20944/preprints202205.0121.v1



models to understand sensory cortex. *Nat. Neurosci.* **19**(3), pp.356-365.

Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, Hassabis D, Botvinick M. 2018. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**(6), pp.860-868.

Werfel, J., Petersen, K. and Nagpal, R., 2014. Designing collective behavior in a termite-inspired robot construction team. *Science*, **343**, pp.754-758.

Wilson EO. 2012. *The social conquest of earth.* W.W. Norton & Co., New York, NY, USA.

Yuste R. 2015. From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**(8), pp.487-497.

Zhou, J., Lai, J., Menda, G., Stafstrom, J.A., Miles, C.I., Hoy, R.R. and Miles, R.N., 2022. Outsourced hearing in an orb-weaving spider that uses its web as an auditory sensor. *Proc. Natl. Acad. Sciences USA*, **119**, p.e2122789119.

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, Yamins DL. 2021. Unsupervised neural network models of the ventral visual stream. *Proc. Nat. Acad. Sci.* **118**(3).