Cognitive Machinery and Behaviours

Bryan Fruchart and Benoit Le Blanc

ENSC, IMS laboratory, Bordeaux INP, CNRS, Talence, France
bfruchart@ensc.fr

Abstract. In this paper we propose to merge theories and principles explored in artificial intelligence and cognitive sciences into a reference architecture for human-level cognition or AGI. We describe a functional model of information processing systems inspired by several established theories: deep reinforcement learning mechanisms and grounded cognition theories from artificial intelligence research; dual-process theory from psychology; global-workspace theory, somatic markers hypothesis, and Hebbian theory from neurobiology; mind-body problem from philosophy. We use a formalism inspired by flow-graph and cybernetics representations. We called our proposed architecture IPSEL for Information Processing System with Emerging Logic. Its main assumption is on the emergence of a symbolic form of process from a connectionist activity guided by a self-generated evaluation signal. This theoretical work aims to provide a reference architecture for analysing behaviours of artificial systems. We also discuss artificial equivalents of concepts elaboration, common-sense and social inter-actions. The originality of this transdisciplinary work is that it stands at a general level of abstraction thus avoiding technical considerations. It can be considered as an artificial general intelligence design proposition althought its aim is to be an analysing tool for Human interactions with present and future artificial intelligence systems. We conclude by enunciating several conjectures on artificial agents' behaviours which could allow readers to explore new perspectives on Artificial General Intelligence.

Keywords: Human-level cognition, Artificial general intelligence, Cognitive modeling.

1 Introduction

Recent publications raised discussions on limits of the followed current approaches in artificial intelligence (AI) [Marcus and Davis, 2019]. These limits on artificial systems' capacities and the debates they raised aren't new. In fact, one could consider they are analogous to the indirect debate between Alan Turing exhibiting his Imitation game as a test of AI [Turing, 1950] and John Searle with his counterargument of the Chinese room [Searle, 1980]. Can machines understand humans? And can humans truly understand machines? Artificial information processing systems aim to simulate processes that are usually done by human cognition, thus we decided to model Human-like cognition as a reference architecture for artificial systems. To design our model, we took inspiration from various established cognitive science theories.

From the field of AI, we took inspiration from deep reinforcement learning (DRL) frameworks, grounded cognition theories and prior cognitive architectures. Systems that implement DRL have been shown to efficiently perform human-level tasks from sensory input computations [Everitt et al., 2018]. It is often said that conventional DRL alone cannot account for the way humans learn. It's too slow, requires very large datasets, doesn't generalize well, struggles to perform symbolic processing and lacks the ability to reason on an abstract level [Garnelo et al., 2016]. Recent reports, however, show that these issues can be overcome by architectural and modality modifications for narrowed environments [Dosovitskiy and Koltun, 2016; Wayne et al., 2018]. As an example, DeepMind researchers implement two different learning speeds for simulation of episodic memory and meta-learning. They concluded that "a key implication of recent work on sample-efficient deep RL is that where fast learning occurs, it necessarily relies on slow learning, which establishes the representations and inductive biases that enable fast learning." [Botvinick et al., 2019]. This architectural consideration of decomposing cognition into two modes has been largely explored by cognitive architects. On a recent review, Luliia Kotseruba and John Tsotsos present a broad overview of the last 40 years of research on cognitive architectures [Kotseruba and Tsotsos, 2018]. The most represented approach is Hybrid architecture, where a connectionist process cooperates with a symbolic one. This dual-process assumption seems to be the most promising one considering the natural synergy between these two approaches. It is often said that connectionist models efficiently perform inductive reasoning and classifications but lack symbolic and deductive abilities. On the other hand, deductive inference requires entities and rules, which are hard to a-priori define for complex and partially observable environments. An interesting idea concerning this constraint is to make the symbolic part emerge from the connectionist activity [Hopfield, 1982].

Hybrid systems' propositions are usually inspired by psychology research where William James proposed in 1890 to decompose human cognition into two subsystems which he named "Associative thinking" and "Reasoning thinking" [James et al., 1890]. Many works have been done around this principle, one of the most notable is the extended experiment conducted by Amos Tversky and the Nobel prize winner Daniel Kahneman on human economic decision making [Kahneman, 2011]. For them, cognitive processes can result from the production of two different systems. System 1 which is described as fast, unconscious and automatic, accounting for everyday decision and subject to errors. The System 2 is slower, conscious and effortful. For Kahneman, System 2 operates complex decision-making processes and is more reliable. To better understand the relation between these different cognitive modes, we have also been vastly inspired by the works of Carl Jung and Sigmund Freud, who were among the first to distinguish and study the unconscious part of our mind [Freud and Bonaparte, 1954; Jung, 1964].

Converging neurobiology studies associate reasoning or declarative cognitive functions with distributed brain activities. This assumption finds an echo in the words of the Global Workspace Theory [Dehaene et al., 1998]. Functional brain imaging shows that conscious cognition is associated with the spread of cortical activity, whereas unconscious cognition tends to activate only local regions [Baars, 2005]. Experi-

mental reports stressed the notion of Free will by observing unconscious initiative before voluntary action [Libet, 1985] giving us our intuition on how both systems are architectured. These large-scale considerations on the brain activity have started to operate a shift in the way that cognitive scientists analyse cognition. Vinod Menon studied psychopathology and wrote: "The human brain is a complex patchwork of interconnected regions, and network approaches have become increasingly useful for under-standing how functionally connected systems engender, and constrain, cognitive functions." [Menon, 2011]. However, brain processing does not rely only on electrical activity; information flows are encoded into electrical-chemical potentials. Emotions, which are associated with chemical neurotransmitters, play an undeniable role in human behaviours whether they are conscious or unconscious. To integrate this part in our model, we took inspiration from the Somatic Marker hypothesis formulated by Antonio Damasio [Damasio et al., 1991]. Lastly, we also considered the Hebbian theory described by neurobiologists. Named after Donald Hebb, this cell assembly theory modelized the synaptic plasticity of the brain. Recent publication in artificial intelligence shows how recurrent neural networks with Hebbian plastic connections provide a powerful novel approach to the learning-to-learn problem [Miconi, et al., 2018].

All these theories have been thoroughly discussed by their corresponding discipline. Because of the space limitation, we cannot reference or further develop these discussions or reports. Moreover, this is not the objective of this paper. In this introduction we have presented what are the transdisciplinary sources that have inspired our architecture. The formalism we use to represent networks' activities is described in part 2 along with the definition of our architecture. Discussion of its attributed capacities will be presented in part 3. More specifically, we will discuss conceptual reasoning, common-sense knowledge and social interactions such as language. As a conclusion, we conjecture on behaviours of agents that would be architectured with our proposition.

2 Information processing systems with emerging logic (IPSEL)

2.1 Cognition as a flow graph

The considered system is represented as a graph of processing units connected together. Processing units are represented by nodes and their connections by weighted and oriented arcs called routes. We define three types of units. Source units which represent sensor organs of the system; sink units representing motor organs; and routing units which are graph nodes that are neither sources nor sinks. All together they form a network in which flows are spreading. These flows are called Action Potential Flows (AP-flows). When a unit or a route is crossed by AP-flows we say it is activated. AP-flows have the property of being persistent for an undefined amount of time. When activated, routing units can emit part of a signal called Emotional Response Signal (ERS). Considered all together, ERSs represent the internal response of the whole network being crossed by AP-flows. We give no constraint on how ERS are

produced, it can be generated by one group of units or the generation can be distributed amongst all units. A schematic representation is given in figure 1.

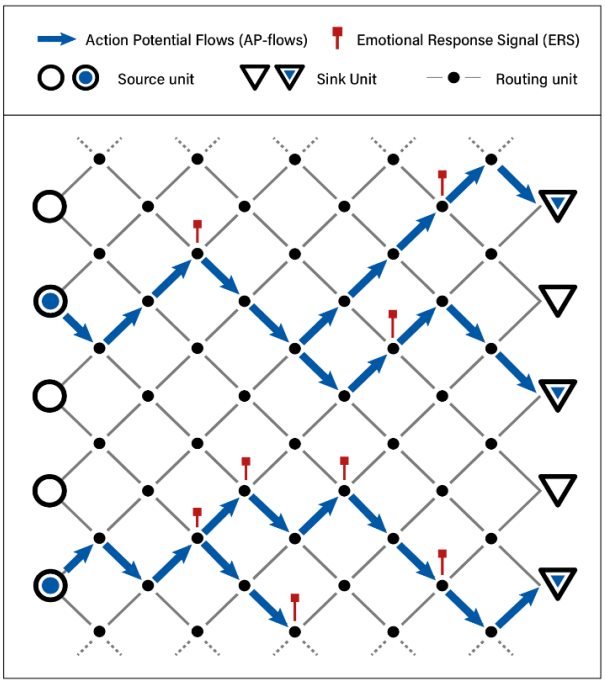

Fig. 1. An arbitrary cognition flow graph. Only few units and routes have been displayed for clarity.

We associate source units' activities as a process which transforms environment interactions into AP-flows. It is continuous and said to be the system's perception of its environment. AP-flows then spread into the network and eventually reach sink units. Sink units' activities are responsible for transformation of AP-flows into environment interactions. This process is said to be the system's behaviour. The function that connects perception to behaviour is called cognition and is represented by the structure of the network.

We named this form of representation a Cognition Flow Graph (CFG). Structural information of the graph is represented by the arcs' weight. They are probabilities of the type "probability of the route to be activated if connected unit is activated". Altogether, the arcs' weights form a probability distribution over pairs of units. We call it the intuitive probability distribution (ID) of the cognition flow graph. There are two mechanisms that allow ID editing. The first is called Hebbian Learning (HL). It has the function to grow connections between unconnected units that have simultaneous activities. It changes the ID probabilities from 0 to something greater than 0. The second mechanism, called Reinforcement Learning (RL), increase or decrease the arcs' weights to optimize emotional response signals of the structure.

The particular routings of AP-flows through the structure determine, for a set of activated source units, which sinks units will be activated. Thus, we say that perception is processed by cognition to produce behaviours. Cognition is performing a computation on perception with intuitive probability distribution as instructions.

2.2 IPSEL functional architecture

Different natures of routing imply different natures of computation and thus different natures of behaviours. In this part we regroup various kinds of routing and define abstract systems that represent their consequent computation.

We distinguish two types of routing possibilities. Direct paths: on these paths, AP-flows have a unique routing possibility. And indirect paths: on these paths AP-flows have multiple routing possibilities, which implies a notion of network and allow flow cycles. Considered all together, indirect paths form a network of networks.

Behaviours engendered by activities on direct paths are called direct behaviours. We represent them as being the production of a system called Direct System (S0). Activities on indirect paths can have two modalities. When one indirect path is considered it is said to be a local activity. When the activities are considered over a combination of indirect paths, involving potentially unconnected distant networks, it is said to be a global activity. Behaviours engendered by local activities are called intuitive behaviours and are the production of the Intuitive system (S1). Combinations of local activities form global activities which embody a computation attributed to the deliberative system (S2). We postulate that S2 emerges from S1 because of the relations between local and global activities.

While experiencing its environment, the structure of indirect paths' networks changes because of RL and HL. At the local scale, preferred routings will emerge and form local patterns of activities. At the global scale, unconnected networks will develop connections because of HL, and preferred routings between these localities will emerge because of RL. We define the notion of concept which, in our formalism, means a combination of local routing patterns that have developed inter-local routes at the global scale. Because of concept formation, local networks can now be activated by flows coming from the global activity. This kind of global flow activities is said to emerge since it requires a previous step of local structure self-organisation. At the local scale, AP-flows are continuous and form a global configuration at any time. However not all global configurations imply activated concepts. Thus, from a global point of view, concepts appear in an ordered sequence. Once again, due to HL, RL and the persistence of AP-flows, concepts that appear close in the sequence will develop and reinforce inter-connections. Since the sequence is ordered, it can also be viewed as the emergence of probabilistic causality relations between concepts. We define a second probability distribution over pairs of concepts called the conceptual probability distribution (CD). ID represents S1 knowledge whereas CD represents S2 knowledge.

Environment perception penetrates the system through sensory organs where they are transformed into Action potential flows called messages. These flows propagate through the structure and activate direct and indirect paths. Propagation on direct

paths will engender direct behaviours seen as production of the direct system S0. Propagation in indirect paths activates local networks and engenders intuitive behaviours seen as production of the intuitive system S1, and it is determined by the intuitive distribution. At the global scale, activated concepts engender new local flow propagations and can be inferred from a previously activated concept. When it is the case, the appearing sequence of concepts is said to be the production of the declarative system (S2) and is determined by the conceptual distribution. The cycle of flow propagation between local and global configurations is said to be the reasoning behaviour of the system. It can also be viewed as communication between S1 and S2. AP-flows that activate sink units for behaviour productions are called commands.

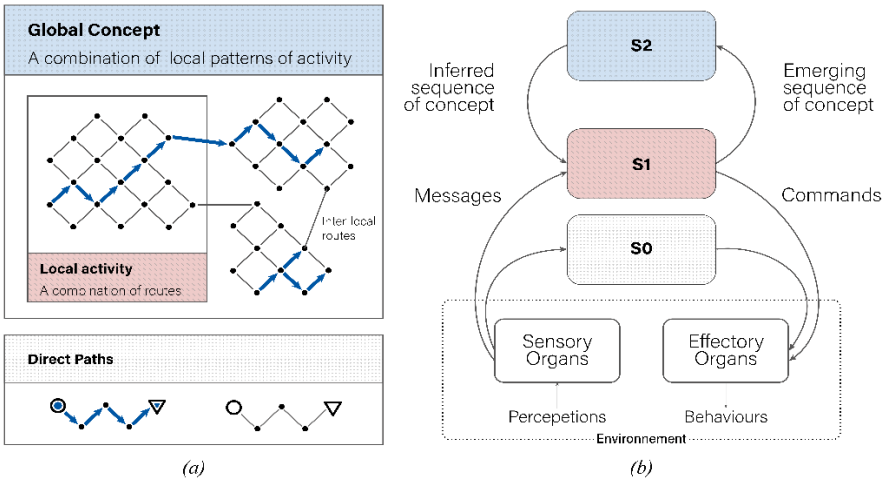

Fig. 2. (a) distinction between direct, local and global activities. (b) IPSEL functional architecture.

IPSEL agents alternate between three natures of behaviours (direct, intuitive or reasoning), corresponding to what is required for environmental interactions. Perceptions that activate direct paths engender direct behaviours. Other perceptions engender intuitive behaviours. Occasionally, internal flow propagations instantiate concepts and trigger concept inferences at the global scale. The inferred concept sequence acts as new sources of flows for local activities. It is the reasoning behaviour of the agent and engenders further intuitive behaviours.

3 Discussion

In our model, a concept is a combination of simultaneous local patterns of activities. We can state that the more a route is activated, the more it may be reinforced. For this reason, invariance on perceptions will engender more reinforcement for their own connected networks. Invariance on combination of simultaneous local patterns of activities will engender inter-local connections thanks to the Hebbian mechanism and develop concepts. From the global perspective, the first criteria of invariance on per-

ceptual patterns is the fact that they continuously change over time. Thus, we could suppose that Time would be one, if not the first, of the primary concepts an IPSEL agent may internally represent with structure differentiation. Through the integration of the concept of Time, the structure can now characterize further perceptions. All perceptions do not change evenly throughout Time and are modulated by the body position and sensor orientation. Therefore, invariance on perceptions through Time engender the formation of the concept of Space. With the ability to represent Time and Space concepts, the structure can now form a concept of Object which is perception's invariances through Space and Time. Time, Space and Object are the three primary concepts. From that point, the system may differentiate objects from one another to form more elaborate concepts, again, by representing the invariance of its perception through already acquired concepts. Depending on its sensors' position, the system could form the concept of its own body, as it may be the most invariant object of perception. Geometric forms, colours, symbols and so on, are all internal representations of invariant perceptions through Space/Time/Object. Progressively, the structure represents its perceptual environment with concepts. IPSEL agent's world representation is thus, totally subjective.

From the global point of view, concepts appear in sequences. Because of Hebbian and reinforcement mechanisms, concepts which are close in the sequence will develop and differentiate inter-concepts connexions. Through the same dynamism in which local patterns of a common concept can activate each other, inter-concepts' connexions enable concept inference. Sequences of concepts can now be internally simulated, therefore the perceptual environment they have originated from, can be simulated. This environment simulation is valuable for the structure, as it gives it the capacity to represent past or future configurations and their associated emotional responses. This allows the system to remember and to predict.

Internal intuitive representation is inspired by the philosophy of Carl Jung [Jung, 1964]. The notion of concept emerging from perceptual experience is inspired and well developed in other terms by grounded cognitivists [Barsalou, 2010]. Characterization of environmental perception through Time and Space consideration is mainly inspired by the philosophy of Arthur Schopenhauer [Schopenhauer, A. 1891]. Objects' definition and relationships for environmental representations is inspired from Rudolf Car-nap's book "The logical structure of the world" [Carnap, 1967]. Recent reports show that the symbolic nature of computation is attainable through a connectivism mechanism with the help of some structural specifications [Lample and Charton, 2019]. Other artificial neural network models consider expressive probabilistic circuits with certain structural constraints that support tractable probabilistic inference [Khosravi et al., 2019]. In the neurobiology field, a neural basis for the retrieval of conceptual knowledge has been proposed from empirical reports [Tranel et al.; 1997] and strong evidence for a neural realization of distributional reinforcement learning have been presented [Dabney et al., 2020].

Common-sense is defined by Cambridge online dictionary as "The basic level of practical knowledge and judgment that we all need to help us live in a reasonable and safe way", or for Marvin Minsky "the ability to think about ordinary things the way people can" [Singh and Minsky, 2003]. For an IPSEL agent, common sense would be

the system's knowledge represented by its differentiated structure. It would have several forms: intuitive when local pat-terns are considered, giving the agent a sort of "common-sense" about which behaviours to produce for a given set of perceptions; conceptual when it states how concepts are linked together, and how objects they represent may inter-act with each other. In both cases these knowledges are embodied in the structure and are thus mostly acquired by individual experience. Experience is relative to the system's perceptual modalities, therefore its common-sense is subjective. For example, distinguishing between north and south magnetic poles appears to be common sense for a homing pigeon whereas most humans require a tool for achieving this distinction. In a broad sense, in the IPSEL paradigm, we would define common-sense as knowledge acquired by experience.

For an IPSEL agent, all behaviours are either direct or intuitive, even if sometimes the intuitive behaviour is triggered by inferred conceptual sequences produced by the declarative system S2. If multiple agents are interacting with each other, they can learn intuitive synchronized behaviours that would externally be seen as communication. They can also learn common symbols that refer to subjective concepts, hence allowing the development of communication language as commonly defined. For that reason, we say that an IPSEL agent has two communication modalities: intuitive where words of a speech are intuitive learned behaviours and conceptual when symbols or combination of symbols refer to concepts. These communications can be of various forms since words of a speech can be of multiple natures such as body-movement, sound, smell or visual pattern (in other words, everything that can be perceived by both agents involved). The various modalities of speech, intuitive or deliberative, have been explored and theorized by psychiatrists Sigmund Freud [Freud and Bonaparte, 1954]. Complex social behaviours have been shown to emerge from artificial multi-agents' interactions with reinforcement learning [Baker et al., 2019].

4 Conclusion

In 1950, Alan Turing proposed a test to evaluate machine intelligence. It has been greatly debated and the community had a hard time defining intelligence and other terms associated to it. As Searl pointed out, symbols don't carry out meaning and symbolic computation isn't enough to catch the idea behind it. It is maybe for this reason that Turing included two humans in his original description of the imitation game. Two humans, when they communicate, can use overtone, common-sense, metaphors, irony, abstraction, that is to say, many language forms that not only rely on grammatically correct symbolic sentences but also on a shared world representation and socio-cultural knowledge. Beside the great achievement of artificial intelligence techniques, machines still struggle to catch these deeper aspects and are only efficient in narrowed environments. Consequently, machine behaviours, trustworthy AI, ethical AI and explainable AI are all new topics of interest for the community.

In this paper, we proposed a reference architecture that gives a functional description of what could be an information processing system that displays human-level cognition. We briefly discussed the artificial pendants of conceptual reasoning, com-

mon-sense and meaningful language. Inspired by transdisciplinary established theories, the model is not a technical description of artificial general intelligence. Instead, it must be considered as a tool for characterizing current and future AI systems behaviours, and Human-AI cooperation.

In our theory, our IPSEL agent builds its knowledge through perceptual experiences. Throughout different phases of development its inner structure self-organizes and enables the emergence of an inner dialog between inner representations and sensory perceptions. This inner dialog is guided by a self-generated signal we have called the system's emotional response. From this perspective, we conjecture that IPSEL's agent is emotionally rational and its knowledge is subjective. For an IPSEL agent, the ability to succeed at the Turing test, would require that the system is granted with the same modality of sensors as humans and has had an individual experience of the world that is close to a human's one. At the end, even with these requirements, nothing assures us that the specific tested agent will pass the test. But are we sure that all humans uniformly would?

References

1. [Baars, 2005] Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research, 150, 45-53.
2. [Baker *et al*., 2019] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528.
3. [Barsalou, 2010] Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. Topics in cognitive science, 2(4), 716-724.
4. [Botvinick et al., 2019] Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. Trends in cognitive sciences.
5. [Carnap, 1967] Carnap, Rudolf (1967). The Logical Structure of the World. Berkeley: University of California Press.
6. [Dabney *et al*., 2020] Dabney, W., Kurth-Nelson, Z., Uchida, N. et al. A distributional code for value in dopamine-based reinforcement learning. Nature (2020).
7. [Damasio *et al*., 1991] Damasio, A. R., Tranel, D., & Damasio, H. C. (1991). Behavior: theory and preliminary testing. Frontal lobe function and dysfunction, 217.
8. [Dehaene *et al*., 1998] Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the national Academy of Sciences, 95(24), 14529-14534.
9. [Dosovitskiy and Koltun, 2016] Dosovitskiy, A., & Koltun, V. (2016). Learning to act by predicting the future. arXiv preprint arXiv:1611.01779.
10. [Everitt *et al*., 2018] Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. arXiv preprint arXiv:1805.01109.
11. [Freud and Bonaparte, 1954] Freud, S., & Bonaparte, P. M. (1954). The origins of psychoanalysis (Vol. 216). London: Imago.
12. [Garnelo *et al*., 2016] Garnelo, M., Arulkumaran, K., & Shanahan, M. (2016). Towards deep symbolic reinforcement learning. arXiv preprint arXiv:1609.05518.

13. [Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8), 2554-2558.
14. [James *et al*., 1890] James, W., Burkhardt, F., Bowers, F., & Skrupskelis, I. K. (1890). The principles of psychology (Vol. 1, No. 2). London: Macmillan.
15. [Jung, 1964] Jung, C. G. (1964). Man and his symbols. Laurel.
16. [Jung and Haier, 2007] Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. Behavioral and Brain Sciences, 30(2), 135-154.
17. [Kahneman, 2011] Kahneman, D. (2011). Thinking, fast and slow. Macmillan.
18. [Khosravi *et al*., 2019] Khosravi, P., Choi, Y., Liang, Y., Vergari, A., & Van den Broeck, G. (2019). On Tractable Computation of Expected Predictions. In Advances in Neural Information Processing Systems (pp. 11167-11178).
19. [Kotseruba and Tsotsos, 2018] Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: core cognitive abilities and practical applications. Artificial Intelligence Review, 1-78.
20. [Lample and Charton, 2019] Lample, G., & Charton, F. (2019). Deep learning for symbolic mathematics. arXiv preprint arXiv:1912.01412.
21. [Libet, 1985] Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and brain sciences, 8(4), 529-539.
22. [Marcus and Davis, 2019] Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon.
23. [Menon, 2011] Large-scale brain networks and psychopathology: a unifying triple network model. Trends in cognitive sciences, 15(10), 483-506.
24. [Miconi *et al*., 2018] Miconi, T., Clune, J., & Stanley, K. O. (2018). Differentiable plasticity: training plastic neural networks with backpropagation. arXiv preprint arXiv:1804.02464.
25. [Schopenhauer, 1891] Schopenhauer, A. (1891). The world as will and idea (Vol. 1). Library of Alexandria.
26. [Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. Behavioral and brain sciences, 3(3), 417-424.
27. [Singh and Minsky, 2003] Singh, P., & Minsky, M. (2003). An architecture for combining ways to think. In IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No. 03CH37502) (pp. 669-674). IEEE.
28. [Tranel *et al*., 1997] Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. Neuropsychologia, 35(10), 1319-1327.
29. [Turing, 1950] Turing, A. M (1950). Computing Machinery and Intelligence. Mind 49: 433-460
30. [Wayne *et al*., 2018] Wayne, G., Hung, C. C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., ... & Gemici, M. (2018). Unsupervised predictive memory in a goal-directed agent. arXiv preprint arXiv:1803.10760.