AI & SOCIETY (2020) 35:995–1007
https://doi.org/10.1007/s00146-020-00942-y

OPEN FORUM

The missing G

Erez Firt[1]

Received: 23 August 2019 / Accepted: 6 January 2020 / Published online: 18 January 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract
Artificial general intelligence (AGI) is not a new notion, but it has certainly been gaining traction in recent years, and academic as well as industry resources are redirected to research in AGI. The main reason for this is that current AI techniques are limited as they are designed to operate in specific problem-domains, following meticulous preparation. These systems cannot operate in an unknown environment or under conditions of uncertainty, reuse knowledge gained in another problem domain, or autonomously learn and understand the problem-domain. We shall call AI systems capable of such feats artificial general intelligent (AGI) systems. The three tasks of this paper are to provide a working definition of the term AGI, examine the "missing G", i.e., the set of abilities that current AI systems lack and whose implementation will result in a basic AGI system, and consider different approaches, including a hybrid one, to a comprehensive solution for an AGI.

Keywords Artificial general intelligence · Learning · Understanding · Creativity · Abductive reasoning

1 Introduction

Artificial General Intelligence (AGI) is not a new notion, but it has certainly been gaining traction in the last decade or so. Conferences hosting researchers from different fields (computer and neuro-sciences, logic and mathematics, philosophy etc.) are dedicated to the study of AGI,[1] university courses invite expert guest speakers to discuss different aspects related to the concept,[2] and academic as well as industry resources are invested in or redirected to research in AGI. One of the main reasons for this is that AI researchers have realized that current AI techniques are limited. To be sure, Deep Learning networks and their variants (e.g., Convolutional Neural Networks) have reached outstanding achievements in various tasks and even surpassed the predictions of many experts; nevertheless, applications of these techniques are considered "narrow" AI. Why? Because these systems are designed to operate in specific problem-domains following meticulous preparation—usually by being fed predefined models and millions of training examples as input, before they can start working—and even then, their accuracy, although averagely high, is not guaranteed. These systems cannot operate in an unknown environment or under conditions of uncertainty. They cannot use knowledge gained in another problem domain. They cannot autonomously learn and produce a model of the world and act accordingly. They cannot understand the problem-domain and realize what model or parameters should be used and extracted from the environment. We shall call AI systems capable of such feats Artificial General Intelligent (AGI) systems, or at least the first phase of AGI systems. The first task of this paper is to provide a working definition of the term AGI, according to the above outline.

The main task of this paper is to examine the "missing G", i.e., the set of features or abilities that current AI systems lack and whose implementation will result in a basic AGI system; in other words, to highlight the features that separate the narrow from the general. These features include unsupervised learning, understanding, abductive reasoning and creativity. We shall explore the need for and the meaning of these capabilities, expound any philosophical aspect arising from the implementation of such capabilities and examine how they are implemented by or fit into different approaches, which offer comprehensive solutions for constructing AGI systems. Other issues, though they may be relevant in some way to the construction of AGI systems, are beyond the scope of this paper: the kind of hardware needed to construct AGI systems, whether quantum computing is

✉ Erez Firt
erezfirt@gmail.com

[1] Philosophy Department, University of Haifa, Haifa, Israel

[1] See Conference Series on Artificial General Intelligence. https://agi-conference.org/ Accessed Nov 2018.

[2] See MIT course page https://agi.mit.edu/. Accessed Nov 2018.

a prerequisite, or whether at all computational power is a crucial factor. Furthermore, we shall not try to speculate or determine the time of the first appearance of such systems—as Yann LeCun[3] stated in one of his public lectures, "[N]o one in their right mind would tell you it's going to be less than 20 years, and if someone tells you it's more than 20 years, what it means is that they have no idea how long it's going to take."

The structure of this paper is as follows: In the second section, we introduce a "good-enough" definition of AGI; in the third and fourth sections, we discuss unsupervised learning, understanding, abductive reasoning and creativity; in the last section, we shall consider two approaches to a comprehensive solution for an AGI system—the brain emulation approach and the cognitive architecture approach. I discuss their similarities and differences, examine how they propose to tie it all together, and at the final concluding section, draw conclusions regarding a hybrid approach that may take us one-step ahead toward a general AI system.

2 What is AGI?

Our starting point for formulating an initial concept of AGI is the ideas and operational definitions presented on the Machine Intelligence Research Institute's (MIRI) website.[4] The main idea follows Legg and Hutter (2006), who basically defined intelligence as a measure of the "agent's ability to achieve goals in a wide range of environments". Other AI researchers follow the same line: Goertzel (2006) defines it as the ability to achieve "complex goals in complex environments"; Voss (2005) claims that an AGI system will need "domain-independent skills necessary for acquiring a wide range of domain-specific knowledge—the ability to learn anything", and that such a system should be a "highly adaptive, general-purpose system that can autonomously acquire an extremely wide range of specific knowledge and skills and can improve its own cognitive ability through self-directed learning". Another idea is that AGI systems should have the ability to transfer learning from one domain to other domains.

These statements echo some of the ideas and capabilities mentioned above. We can now try to converge to a working, "good-enough" definition that employs these ideas and others that seem necessary, and which can serve us throughout this paper. One capability, mentioned repeatedly and emphasized by almost every thinker, is learning—autonomous unsupervised learning of any problem-domain, which includes the ability to transfer gained knowledge, i.e., to share knowledge between domains. Instead of systems that should be taught how the world (or the problem-domain) looks like, AGI systems learn independently, *understand* the problem-domain, extract the essential features from it and are able to reuse gained knowledge and re-apply it by tweaking it to fit the new domain—they learn from experience. Learning gives rise to other essential capabilities, such as understanding and reasoning. We expect AGI systems to understand their environment in a way that will enable them to extract from it the necessary features, i.e., those parameters that are most relevant to the problem at hand in the given context, and apply them to the solution. We also expect AGI Systems to be able to use abductive reasoning, and infer the best explanation when only partial information is available. When we refer to abductive reasoning, we shall henceforth employ the prevalent definition of abduction as *theory-formation*: "Given a background theory and an observation to be explained, an abductive inference conjectures one or more best explanations for the observation from the background theory" (McIlraith 1998: 3). Thus, we expect AGI systems to come up with one or more best explanations of observations made in a problem-domain, and to be able to decide which explanation is the most suitable under current conditions. Another important and widely discussed capability is creativity, whose definition commonly involves the concepts of autonomy, intentionality, appraisal and emotion, which are highly controversial in themselves, and even more so when AI systems are concerned. We shall examine and evaluate the essentiality of creativity as far as AGI systems are concerned, and whether such systems can meet the initial conditions, which are commonly considered as pre-requisites for real creativity. In the following sections, we discuss the above set of capabilities in detail.

To sum up the basic idea of a "first-phase" AGI system, let us frame it in a few clear sentences. The rest of the paper is dedicated to explicating this rough definition. Thus, our AGI system is an autonomous system in the sense that it can learn in an unsupervised manner, i.e., without being instructed what kind of model it should follow and what parameters or features it should extract from its surroundings. It understands the world around it in a sense that allows it to realize how to model a new problem, learn from experience in the sense of sharing and transferring the insights learned between different problem-domains, and use abductive reasoning in a way that will enable it to reach decisions and take actions based on uncertain and limited data. To accomplish this and more, our AGI system should be creative in at least a limited sense, which is discussed below.

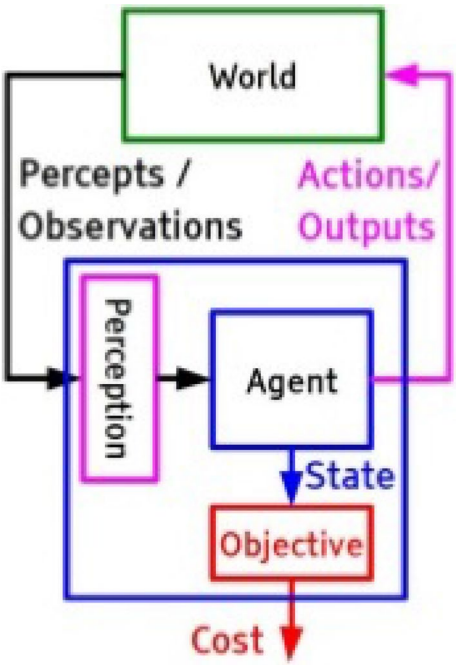
[3] Yann LeCun is a professor of computer science and the Director of AI research in Facebook.

[4] What is AGI? https://intelligence.org/2013/08/11/what-is-agi/ Accessed Nov 2018.

3 Learning, understanding and reasoning

In this section we shall consider a few fundamental capabilities, with which every AGI system should be equipped, namely the ability to learn anything independently, the ability to understand its domain and surroundings in such a way that enables it to extract essential features correctly to model the problem, and the ability to reason based only on uncertain and partial data, i.e., to form hypotheses and explanations, decide which of them optimally suits the situation and act accordingly. Obviously, implementing these abilities is fraught with some philosophical and technological difficulties. In what follows, we shall discuss the principal obstacles.

Let us look at a simple, basic architecture of an AI system, through which we can examine, where these capabilities come into play.[5]



There are three basic modules here: a Perception module, which enables the system to evaluate the state of the world; an Agent module, which generates actions, makes predictions that enable the system to plan and reason, and which has an internal state, of which I shall expand presently; the third module is the Objective, which evaluates the state of the Agent and calculates its motivational level or its drive to act in a certain way. One of the main purposes of the Objective module is to drive the Agent to do what we want it to do. This can be done by hardwiring an immutable sub-module into it, something which is similar in function to instincts, and which will drive the Agent to act in a way that we believe is good. Another sub-module is used to estimate a value function, which basically consists of the human values we want it to follow.[6]

Thus, through this simple architecture, we shall try to evaluate how the abilities mentioned above come into play. The main focus has to be on the predictive ability of the Agent. Basically, this ability enables the system to infer the state of the world even when only partial and uncertain information is available, and to be able to infer the future from the past and the present, and the past from the present. Humans and animals learn to do so, from the moment they are born. We learn how the world works by observing it in the first few years of our lives—we learn basic truths about our world, e.g., object permanence, dimensions, gravity, causality and so on. This gives us what AGI systems need more than anything, i.e., common sense. Moreover, this gives us the ability to fill in the blanks, e.g., in our visual field or in conversations, and predict the consequences of our actions. In fact, one can argue that the ability to predict is an important and essential part of being intelligent. The idea that the brain is fundamentally a probabilistic prediction machine is an increasingly influential thesis (often referred to as predictive processing or predictive coding) in cognitive science in recent years. Roughly, the thesis asserts that the brain is continually striving to minimize the mismatch between self-generated predictions of its sensory inputs and the sensory inputs themselves. This process of prediction error minimization (repeated throughout the hierarchical structure of the neocortex) allegedly generates the full range of psychological phenomena that make up the mind.[7] Our Agent needs what is sometimes referred to as a "world simulator", i.e., an internal simulation of the world that it may employ to plan ahead, by examining action proposals on (a simulation of) the world and evaluating their outcomes.[8]

Recent promising developments in the field of unsupervised learning have given rise to Generative Adversarial Networks (GANs), a deep neural network architecture consisting of two networks that compete one other for training. Roughly speaking, the architecture comprises a generator

[5] This is only for illustration purposes. The reader should keep in mind that future AGI architectures might be quite different.

[6] Recent ideas regarding the way to train an agent to follow the right values involve Adversarial Training, i.e., training AI agents by letting them compete against other AI agents.

[7] See Hohwy 2013 and Clark 2016 for overviews.

[8] However, in recent decades, following works such as Brooks 1991, the use of internal representations in artificial systems such as robotic systems has been laid aside. Brooks had noticed that internal representations "get in the way" when building very simple intelligent systems, and therefore urged researchers to "use the world as its own model".

network, whose task is to create or simulate a certain data structure, and a discriminator, whose task is to validate the data structure passed to it by the generator and authenticate it. For example, the generator tries to create a certain image and the discriminator tries to check its authenticity. Thus, the generator is in a feedback loop with the discriminator, and the discriminator is in a feedback loop with the ground truth of the images, which are either fake or are taken from a certain known dataset. Consequently, both nets are trained. Although a GAN may use labeled datasets for training (the dataset of images in the example above), its goal is to model what the data look like and to be able to generate new examples of what it has learned. In other words, unlike supervised learning, where the input data are associated with labels and the goal is for the model to generalize and associate new data with known labels, in this case the data come unlabeled, and the goal is different: to model what the data look like and use them to generate new examples.

The near-future advantages of unsupervised learning are obvious: the ability to learn without needing any examples, i.e., a training data set, or from, at most, a small number of examples and to generalize features that enable the system to apply its experience to different tasks. In the case of AGI systems, this ability allows these systems to learn a new domain or environment based on a very small number of observations and its past experience, without the intervention of humans. For example, the hope is that an AGI system entering a new environment will learn new causal relations based on observation and its past experience with other causal relations (and the generalization of the causal relation features) without any human labeling or intervention.

3.1 Understanding

Understanding goes hand in hand with learning and reasoning. In fact, there is no sharp distinction between successful learning and understanding in a context, wherein a system is supposed to function and act under conditions of uncertainty in an unknown domain. One prominent approach, in the field of AI research, is that definitions of understanding should emphasize utility and refer to something that can be measured. In what follows, we mention a few types of understanding, review a highly AI-relevant view of understanding and a pragmatic approach (Thorisson et al. 2016), which highlights certain criteria that can be measured and tested.

In epistemology, we usually refer to three types of understanding: propositional understanding, understanding-why and objectual understanding. In the context of AGI research, the more interesting types are understanding-why and objectual understanding; the former is implied in sentences that take the form "I understand why X" (for example, "I understand why this and that happened"), whereas the latter is implied in sentences that take the form "I understand X", where X can be thought of as a body of information or a subject matter. Pritchard (2009) referred to these two types as atomistic and holistic understanding, respectively, which can also be thought of as emphasizing the difference between concrete cases of understanding and understanding a structure or a body of information. Grimm (2011) suggests that objectual understanding can be helpfully thought of as akin to a "system or structure [that has] parts or elements that depend upon one another in various ways", and Riggs (2003: 20) agrees, emphasizing the importance of the relations among parts and between the parts and the whole, when trying to understand a subject matter.

An interesting and AI-relevant idea is that of understanding as Representation Manipulability (URM). Roughly, Wilkenfeld (2013) suggests that one understands when one possesses a representation of that which is understood that is sufficiently robust to be manipulable for inferential and practical purposes. In other words, understanding occurs when we have a robust mental representation of the thing to be understood. This robustness is expressed by the ability of the understander to manipulate and tweak this representation to examine inferences and take actions. In the same spirit, Grimm (2011) suggests that manipulating the "system" allows the understander to "see" the way in which "the manipulation influences (or fails to influence) other parts of the system" (Grimm 2011: 11). Thus, according to Grimm, understanding the relationships between relevant parts of a subject matter amounts to manipulating the system by changing parts of it and observing the impact on the overall system. He refers to such ability as Grasping, and suggests that it also allows the understander to anticipate what would happen if things were relevantly different. It allows the agent to make correct inferences about a world in which the relevant differences obtain.

Let us examine the relevancy of URM to AGI architectures of the type mentioned above. First, this account of understanding relies on the premise that "the difference between understanders and non-understanders is that the former, but not the latter, can utilize the understood effectively" (Wilkenfeld 2013: 1004).[9] In other words, when one understands something—an object, a situation etc.—one is able to manipulate the thing understood effectively, e.g., in a way that will enable her to achieve goals. Another basic premise concerns the ability of an agent to represent objects

[9] As Wilkenfeld also mentions, this point is similar to the one made by Woodward (2003) in his manipulationist accounts of causation and explanation, according to which causes can be thought of as devices for manipulating effects, and causal explanations include citing causal variables the alteration of which (in appropriately specified counterfactuals) would have affected the *explanandum*.

mentally, because a mental representation[10] of an object X is a prerequisite for an agent to understand X. According to URM, understanding is the ability to manipulate some mental correlate, i.e., the mental representations, of the understood object such that one would then be able to manipulate the target itself. In a more formal way, a statement that an agent A understands an object O,[11] is true in a certain context, if and only if A has a mental representation R of O that A can modify in certain ways to produce R', which can then be used to manipulate or make inferences pertaining to O.

Thus, an AGI system learns the problem-domain (the context in which the object it wishes to understand "resides") through its Agent module and creates its (artificial correlate of a) mental representation. The Agent can review possible actions or examine possible hypotheses (on which we presently expand) and their possible outcomes, by manipulating the artificial representation that it created, examining the outcomes of these manipulations and through all that infer the best course of action under the conditions of the problem-domain. Thus, this view of understanding is related to unsupervised learning (through which the Agent can obtain artificial correlates of mental representations of objects in certain domains), abductive reasoning (through which the Agent can come up with hypotheses it can examine by manipulating its artificial representations) and creativity (which can be a useful tool in manipulating and making significant small alterations to object representations).

Another, more formal, approach to understanding is introduced by Thorisson et al. (2016) as a "pragmatic theory of understanding rooted in an analysis of how predictive controllers compute meaning" (*ibid*., p. 107). The authors are interested in a sort of understanding that allows an agent to act "in a practical and goal-directed way", and that guides its behavior in beneficial ways. Thus, understanding a certain *phenomenon* is defined as the level of accuracy of the *Model*, which the agent holds with respect to the phenomenon. The accuracy of the appropriate model, and, therefore, the agent's level of understanding, is determined by the quality (correctness) of representation of two main factors in the model: The completeness of the set of elements associated with the phenomenon as represented by the model, and the accuracy of the relevant elements. Let us examine this definition in more detail[12]: a *phenomenon* is a process, state of affairs, or occurrence in a certain domain, which is made up of a set of elements that are related (or unrelated) to one other (by causal or part-whole relations, for example). A *Model* is a set of information structures pertaining to a phenomenon. These structures can be used to explain and predict the phenomenon; produce effective plans for achieving goals with respect to the phenomenon, and (re)create the phenomenon. When speaking of the accuracy of the model with respect to the phenomenon, we mean the level of detail by which the information structures of the model describe the elements of the phenomenon and their relations. Hence, we can estimate the level of understanding of a phenomenon by an agent by assessing its capabilities to predict, explain, achieve goals with respect to and (re)create the phenomenon. The latter is considered by the authors as the strongest kind of evidence for understanding, since by (re)creating they mean the ability to produce a model of the phenomenon in sufficient detail to replicate its necessary and sufficient features. This, in turn, attests to the agent's level of understanding according to the two factors mentioned previously, i.e., the completeness of the set of elements associated with the phenomenon and their accuracy. Both factors are required for (re)producing the phenomenon in question.

3.2 Abductive reasoning

Thus, AGI systems are required to learn and understand to accomplish tasks. To do that, they need to put their understanding to use by predicting and taking actions. This requires the agent to reason, often under conditions of uncertainty and with partial information, much like humans do in many situations of their daily life. This, in turn, requires the most common type of reasoning—abductive reasoning. Let us first understand what exactly abductive reasoning is: an abductive inference is often referred to as an inference to the best explanation.[13] That is, given a set of observations O, an agent forms possible and plausible explanations $E_1 \ldots E_i$ and infers E_j, which best explains O. Obviously, one has to be able to determine what a plausible explanation is, what constitutes a best explanation and, moreover, one has to determine whether E_j is true, approximately true, or merely probable. Another problem for an agent that reasons

[10] Wilkenfeld tries to remain as neutral as possible regarding the question of what mental representations are, but commits himself to the assertion that they are, minimally, "computational structures with content that are susceptible to mental transformations" and to the assertion that this is "consistent with classic computationalism" (*ibid*.) Hence, we can assume that at least some versions of the Computational Theory of Mind comply with Wikenfield's minimal description of mental representations.

[11] Object O is any object of understanding and it can include theories in physics, a certain proof in mathematics or logic, a person (as in, "I understand my friend"), a story or an event, an action, or a phrase in a language, to give some examples.

[12] Italicized concepts are rigorously defined in Thorisson et al. 2016, §3 and I did not see any point in reiterating these definitions here.

[13] In this paper, I employ the terms "abduction" and "abductive reasoning" in their more modern sense of justifying hypotheses. In this sense, abductive reasoning is often associated with "Inference to the Best Explanation" (in contrast to the historical sense, according to which "it refers to the place of explanatory reasoning in generating hypotheses" [Douven 2017: 1]).

abductively is that E_j may just be the best explanation from among a set of bad explanations.[14] In other words, $E_1 \ldots E_i$ may consist of not-so-good explanations of O, of which E_j is the best, while a different set of explanations of O may contain several explanations that explain O better than E_j.

McIlraith (1998) specifies three characterizations of abductive reasoning predominant in the AI literature: the logic-based account, the set-covering account and the probabilistic account, of which the logic-based account is the most prevalent. Without getting too much into concrete logical definitions,[15] we take the liberty to describe roughly the prevalent definition of the logic-based account, i.e., an abductive framework based on a background theory. McIlraith (1998, §2) defines a generic abductive framework as consisting of a background theory T and a set of literals W of an assumed language L, from which explanations are drawn. Given this framework and an observation O, E is an abductive explanation of O (from W) *iff* $T \cup E$ entails O and $T \cup E$ is satisfiable, i.e., there is at least one assignment to each variable that makes the proposition evaluated to be true. To come up with the best abductive explanation, we need supplementary definitions of simplicity and minimality[16] (of literals of the language) for example, but these vary and they are language dependent.

These three components, i.e., learning, understanding and reasoning, are tightly connected. The ability to learn and understand is a prerequisite for the other types of cognitive abilities required from our AGI agent, namely abductive reasoning, forming hypotheses, and taking decisions based on the best hypothesis. For example, our agent must be able to learn a new problem-domain, identify a certain phenomenon as belonging to an abstract relation it already understands, e.g., causal relation, and manipulate the environment to take action based on what it decided to be the best explanation of the problem at hand. Much like humans, our agent must be able to associate a single new observation with an already-known relation (as we do, when we observe an occurrence and classify it as an instance of the cause-effect relation) and act accordingly.

4 Creativity

Creativity is a fundamental feature of human intelligence. Everyone has it to a certain degree, and it is essential for problem solving, both in everyday life and in the context of innovative scientific-artistic activity. Creativity is also closely related to many other cognitive capacities, such as the association of ideas, memory, perception, analogical thinking and reflection as well as to motivation, emotion, cultural context and personality. This multitude of connections is one reason that creativity is such a challenging and elusive concept to define. There are two main reasons for researching and implementing creativity in AI systems: one is to understand creativity in human beings and the other is to construct creative AI-systems, which will be able to cope with problems of different domains in a similar way to human beings. In this section, we focus on the latter—a non-exhaustive definition of creativity is provided, followed by a discussion of the question of machine creativity, i.e., is it possible for an artificial system to have "real" creativity?

One way to define creativity is to see it as the ability to generate creative ideas or artefacts, where a creative idea is one that is novel, surprising, and valuable.[17] These terms have several interpretations, which we elaborate presently. Regarding the concept of novelty, an idea can be new to its originator or to all human thought (as far as it can be known); the former is a form of psychological creativity and is termed by Boden P-Creativity, while the latter is historically creative, or H-Creativity. For our purposes here, P-creativity is the more interesting type of creativity, since it concerns the psychological mechanisms that underlie originality and the question of whether it may be instantiated in AGI systems. Compared to novelty, the criterion of value is much more complicated. Must an idea have value to be considered creative? If so, how can we evaluate its value? These difficulties can take at least two forms: the idea can be extremely specific to a domain, culture or group, in which case it can be considered valuable only by a very limited number of people; another difficulty is related to the period during which an idea or an artifact may be considered valuable—are long lasting second-rate ideas more valuable than ephemeral, trendy, first-rate ideas? From the last statements, it becomes clear that value judgments are relative and, therefore, difficult to agree on. What about the criterion of surprise? There are three types of surprise that correspond to three types of creativity mentioned by Boden. First, the unexpected, statistically unusual type of surprise, which we have always known to be possible, e.g., when a friend wins the lottery. Second, the type of surprise we experience when we come across something that we did not expect and have never even considered, but once experienced, seems to fit a certain familiar pattern, e.g., a new artwork by a known artist. The third type of surprise is the amazement we feel when presented with (what we have believed up to this moment to

[14] See also van Fraassen 1980: 143, who termed this problem "the best of a bad lot".

[15] See McIlraith 1998, §2 and especially §3.

[16] Other definitions of best explanation can include additional criteria such as priority rankings or probabilities.

[17] See Boden 1998; 2004; 2014.

be) an impossible idea, e.g., the stochastic, indeterministic nature of the quantum domain.

As mentioned above, these three types of surprise correspond to three types of creativity suggested by Boden: combinational, exploratory and transformational creativity. The first involves novel combinations of familiar ideas, e.g., metaphors, collages in art, the use of analogies in science, etc. In the latter case, the combination can have an explanatory value, as when a physicist compares the atom to the solar system. The other two types of creativity are different and closely linked, for they involve the alteration of conceptual spaces, i.e., structured styles of thought, which are already accepted within a certain group. In other words, the conceptual space includes the disciplined ways of thinking rooted in one's culture or social group. In exploratory creativity, the originator explores the familiar terrain of his or her conceptual space and comes up with something novel, limited by the constraints of his existing structures of thought. In transformational creativity, a true alteration of one's conceptual space occurs—an impossibility from the point of view of the structured mind, which occurs when one removes or alters a previously inherent limitation or adds new ones. This enables the formation of new structures that could not have arisen before, which in turn enables the generation of previously impossible ideas. It should be noted that some writers consider Boden's work to be non-exhaustive, e.g., Novitz (1999) argues that at least one important type of creativity has eluded Boden's taxonomy, namely "invention from scratch".[18] In addition, Novitz claims that within her classification, Boden leaves no room for correct measurement of degrees of creativity.

We can now move to examine the question of machine creativity. Must "real" creativity involve autonomy, intentionality and emotion? One common objection to machine creativity is that real creativity requires autonomy and programmed machines cannot be autonomous. On this view, combinational and exploratory creativity can be simulated by programmed systems, but to allow transformational creativity is to say that programmed systems can go beyond their instructions. This kind of objection seems to be outdated. If we achieve unsupervised learning and understanding, we also achieve the ability to learn new things and new rules, to drop refuted hypotheses and adopt newly confirmed ones as well as to learn and understand new patterns, revealed as the system explores uncharted domains.

When considering autonomy, intentionality and emotion as prerequisites for real creativity, we can tentatively maintain that emotions (and consciousness) are not necessary for creativity,[19] and that intentionality is an essential property of ideas and artefacts, which were conceived or made for a certain purpose.[20] Thus, we are left with the question of autonomy, i.e., with the question of whether AGI systems can be autonomous and consequently creative. Thus, it can be argued that creativity is dependent on the other capabilities we previously mentioned, i.e., learning, understanding and reasoning. First, we should clarify the question. The notion of autonomy here is directly associated with spontaneous action and freedom of choice, and to the question whether these can be applied to artificial systems. Obviously, programmed systems, by definition, are limited to a certain extent by instructions of a human programmer. Narrow AI is also limited by its problem domain and the fact that its learning is supervised. In these types of systems, it seems that it is justified to claim that only combinational and exploratory creativity are possible. Genuine, radical transformations can arise when unsupervised learning and interaction with the environment occur, followed by understanding of the sort discussed above. Autonomous action can arise when learning and understanding of the environment occur without limitations and the system is able to reason and modify the environment to achieve its goals. How are these goals determined? Ultimate goals are likely to be dictated by the system's human creators (at least in the first phase, when AGI systems will still be constructed by humans), but this fact should not influence the system's transformational creativity—it can be manifested when trying to achieve intermediate or instrumental goals. In this sense, people are limited in practically the same way: our goals, purposes and wishes are dictated by external factors as well—society, education, etc.

5 How does it all fit together?

In the previous sections we discussed a basic set of abilities that seem to be necessary, although perhaps not sufficient, for the creation of an initial-phase AGI system. In this section, we shall examine two approaches that integrate these capabilities into a whole system architecture. One approach is referred to as the cognitive-architecture approach. It emphasizes the tight integration of the capabilities discussed above and other cognitive mechanisms. The other approach, the brain-emulation approach, stresses the need to learn from the human brain, and it suggests that we do so by emulation. This feature is common to both approaches—they both

[18] An instance of creative thinking that does not involve an exploration or transformation of an existing conceptual space, but rather develops a new one, or creates it from scratch.

[19] This is due to the fact that there are examples of creativity without consciousness or emotions being exhibited. See Boden 2014, §4.

[20] Therefore, creativity defined by reference to ideas and artefacts must be intentional. See Boden 2014, §4, for an exception to this rule regarding artefacts such as artworks and poems.

contain turn to the human brain in search of insights, solutions and ideas.[21]

Langley (2012) characterizes the cognitive-architecture paradigm by introducing several general features and assumptions that researchers should focus on, to make progress toward understanding intelligence. The first is high-level cognition, which goes beyond the "ability to recognize concepts, perceive objects, or execute complex motor skills" (*ibid*.: 4), and refers to complex, multi-step reasoning, understanding (mainly, of natural language), planning, and even reasoning about one's own reasoning. The second is structured representations, i.e., structures that represent the complex relations that arise during reasoning, understanding and other higher-level cognitive processes. The third is an emphasis on system-level accounts of intelligence, i.e., accounts of the relations between the different components and capabilities discussed in previous sections, and the way these relations may give rise to high-level mental abilities. The fourth feature involves the use of heuristics. Analogously to their use by the human mind, heuristics allow the user to tackle difficult problems without looking for "guarantees", by finding acceptable rather than optimal solutions. The fifth concerns our study of human cognition and our ability to implement this knowledge when constructing intelligent systems. Langley believes that there is much to learn and re-engineer into artificial systems from the cognitive abilities that humans exhibit so effortlessly, despite their limited computational power and working memories. The final feature concerns the exploratory research of constructing new systems that can exhibit intelligent capabilities that we identify in humans. This research yields many insights[22]: the importance of seeking unified architectures for cognition and not focusing on individual capabilities and focusing on finding a small set of very general mechanisms, through the interactions of which the diversity of intelligent behavior arises.

Let us examine how these features map to the capabilities we discussed in the previous sections. The first, i.e., high-level cognition, is self-evident, for high-level cognition includes understanding and reasoning. The second, i.e., structured representations, is tightly linked to understanding, especially to such views as URM—the ability to represent mentally the environment or part of it and to be able to manipulate it to achieve goals. The third deals with the relations between these capabilities and is the core of cognitive architecture, i.e., that which gives rise to whole-system properties, which we try to achieve when we emulate the human brain. The fourth—the use of heuristics—is tightly linked to abductive reasoning. Finally, putting to use the fifth and sixth points will lead us to implementing the kind of features we discussed in the previous sections—the ability to learn from experience, understand and reason from uncertain, partial data.

To exemplify the role of the above-mentioned features and emphasize the architectural and systematic characteristics of the cognitive-architecture approach, let us consider the sigma cognitive architecture (Rosenblum et al. 2016). Sigma is an integrated computational model of intelligent behavior, which eventually aims to achieve the original grand goal of AI/AGI development, i.e., a working implementation of a full cognitive system, and to ultimately support the real-time needs of intelligent artificial agents, e.g., robots. The development of Sigma is guided and motivated by four goals: (1) grand unification, which aims to go beyond the cognitive capabilities required for human-level intelligent behavior, and include key non-cognitive aspects such as perception, motor control, and affect; (2) generic cognition, spanning both natural and artificial cognition at a suitable level of abstraction; (3) functional elegance, i.e., a broad enough cognitive (and non-cognitive) functionality, which will ultimately suffice for human-level intelligence; and (4) sufficient efficiency, i.e., enabling real-time performance for applications of intelligent artificial agents/robots, and for large-scale experiments in modeling human cognition.

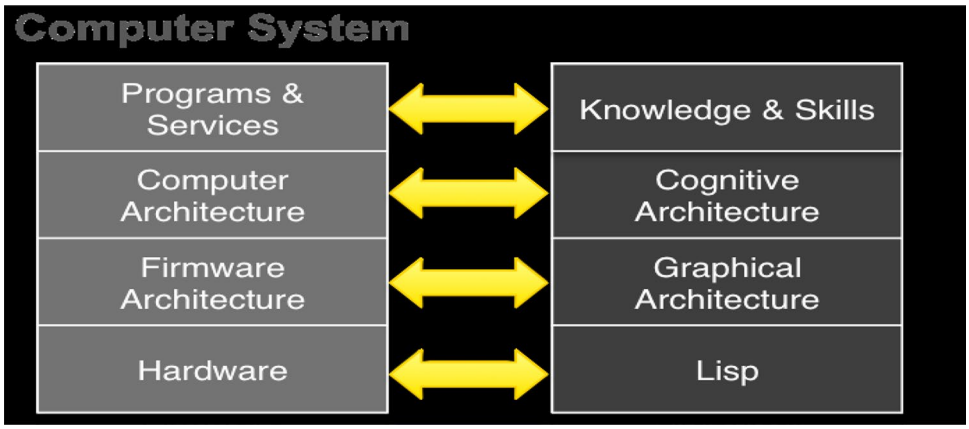
Sigma is built on lessons learned from over three decades of independent work in cognitive architectures and graphical models,[23] among which are the need for a long-term memory, a working memory, and perceptual buffers; the

[21] At this point, we should also mention the 4e (Embodied, Embedded, Extended, Enactive) cognition methodology, which is prevalent in robotics as it focuses on issues of embodiment of cognition. In general, proponents of this methodology claim that many (if not all) cognitive phenomena are in some sense dependent on the morphological, biological and physiological details of the agent's body, its environment and its interaction with the environment. Thus, they claim that cognition involves extracarnial processes. This claim can take a strong and a weak form—the former suggests that cognitive processes are essentially based on extracarnial ones and the later suggests that they are only causally dependent on them. In this paper, I assume that the main requirements of a suitable weak interpretation of this methodology can be implemented within the two approaches—both can accommodate a sort of causal interaction with the environment. A discussion of the strong interpretation of 4e cognition as a separate approach is beyond the scope of this paper.

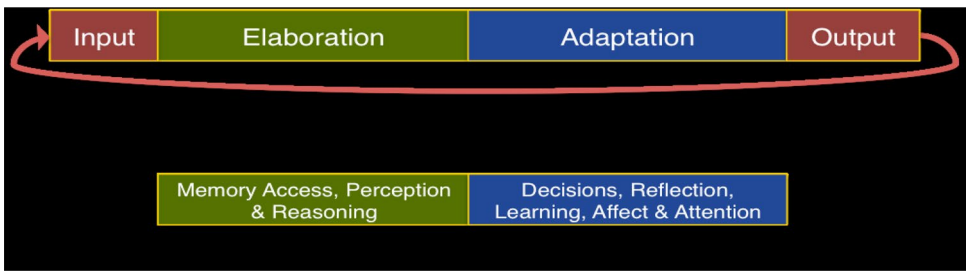
[22] See Paul Rosenbloom's interview on lessons learned from Soar to Sigma: Paul Rosenbloom on Cognitive Architectures. https://intelligence.org/2013/09/25/paul-rosenbloom-interview/. Accessed Nov. 2018.

[23] See especially Rosenblum et al. 2016, Sect. 5.1 for a technical introduction of graphical models. In general, these models concern efficient computation with complex multivariate functions. This includes decomposing these functions into products of simpler functions, mapping the resulting decompositions onto graphs, and computing over these graphs via message passing or sampling algorithms. Graphical models provide working approaches to both symbol and signal processing, and to both logical and probabilistic reasoning.

importance of multiple forms of long-term memory; the centrality of the ~ 50 ms cognitive cycle; and the criticality of combining symbolic and statistical information. At a high level, Sigma as a cognitive system is best understood in terms of a sequence of layers that are analogous to those found in computer systems:



At the core of the cognitive architecture is a cognitive cycle that is intended to correspond to the ~ 50 ms cycle in humans:



In the first major phase, i.e., the elaboration phase, knowledge about the current situation is processed and conclusions are drawn from it. The role of this phase is to elaborate what is understood about the current situation. At this phase, the system does not make choices or learn, but tries to determine, recalculate and update the content of its memory. The second major phase is responsible for making choices and persistent changes to memories based on the understanding achieved in the elaboration phase while also engaging in meta-architectural processing. These may include short-term modifications to working memory as well as alterations to the system's affective and attentional states. It also includes long-term modifications to long-term memory, i.e., learning. This phase is inherently about making changes. This includes decisions about the operators to be applied to make progress in the current situation, changes to working memory and different aspects of reflection, learning, affect (i.e., emotion), and attention. The last paragraphs were an attempt to examine how cognitive architectures can be a suitable platform for the implementation of (some of) the features discussed above. In the concluding section, we shall further expound on this.

The second general approach we shall discuss supports the artificial emulation of the biological brain. Proponents of this approach include large enterprises such as the human brain project (HBP), as well as small companies such as Jeff Hawkins' Numenta. The HBP takes two paths toward artificially imitating the brain: one is Brain Simulation and the other is Neuromorphic Computing. The brain simulation platform (BSP) aims to replicate the brain and its workings on a computer. To do so, they have to characterize the biophysical and computational properties of human neurons, a very challenging task for two main reasons: the complexity of the brain, its components and their relations and the ethical limitations involved in acquiring data directly from human brains. As regards Neuromorphic computing, this approach implements aspects of biological neural networks as analogue or digital copies in electronic circuits. Its goal is twofold: offering a tool for neuroscience to understand the

dynamic processes of learning and development in the brain and applying brain-inspired processes to generic cognitive computing.

Numenta is a company that focuses on reverse-engineering the neocortex, according to the ideas of its founder Jeff Hawkins.[24] Numenta researches the way the human brain works—more specifically, how the neocortex works and how brain cells work together to create perceptions, predictions and behavior. The project's aim is to build machine intelligence based on the principles of neocortical function, which the company believes are necessary and sufficient for any system that exhibits general-purpose intelligence, biological or artificial. Why the neocortex? Because it is a sophisticated learning system that learns how to model the world from the data streaming through our senses, and it is this fact that causes Hawkins and others to believe that by emulating the neocortex, we will be able to build machine intelligence endowed with the abilities of learning, understanding, reasoning and creativity. A recent example of the research in this field is the "Thousand Brains Model of Intelligence",[25] a cognitive model proposing that cortical recognition and modeling capabilities are much more powerful than previously assumed. According to Hawkins, this research avenue will become indispensable to the future development of AI, facilitating the solutions of such problems as generalization and flexibility in applying the learning of one domain to other domains.

One of the major difficulties that this approach is facing is related to the thesis of Connectionism[26] and the long-discredited assumption that the neuron is simple enough to be artificially emulated. A recent criticism can be found in Sardi et al. (2017), who claim that the long-lasting computational scheme of the neuron as a single electrical, excitable, threshold unit is misguided, and that the neuron is in fact a much more complex unit. According to the "old" scheme, neurons sum the incoming electrical signals via their dendritic trees and generate a spike to their axon if the membrane potential reaches a certain threshold. Therefore, the waveform of the spikes, e.g., rise time, peak values, depolarization period and decay time unto a resting potential, is consistently reproducible by the neuron, but varies among neurons. Thus, this relatively simple computation scheme of biological neurons consists of a single, centralized, excitable mechanism that linearly sums its entire signal input. The authors question this common scheme, and suggest, based on experiments, that a neuron functions as an anisotropic, i.e., directionally dependent, threshold unit. Every neuron contains several independent excitable sites, each functioning as an independent threshold unit that sums up the incoming signals from a given limited spatial direction, most probably by one or more dendrites. These independent threshold units are not identical and are characterized by different spike waveforms and different summation specifications. This, in turn, suggests that the neuron[27] is a more complex and structured computational element than assumed heretofore, and the implications for its computational capabilities will surely have interesting consequences for any artificial system with emulated neurons as its basic component. The authors note, however, that it is not suggested that each neuron is composed of several separate and detached neuron-like units that sum up signals coming from certain dendrites, for the case may well be that the dynamics of the threshold units will prove (experimentally) to be coupled, since they share the same axon and may also share a refractory period.

This approach suffers from additional substantial difficulties, originating mainly from our lack of knowledge: we know little about neural connectivity—what are the rules neurons follow when they create different types of connections? What is the importance of these types? We lack knowledge of the precise timing of action potentials and of organizational levels with regard to the processing of information by large neuronal groups or circuits. These seem to be substantial obstacles for brain emulation projects.

However, it should also be noted that the brain-emulation approach has recently received apparent support from a study performed by DeepMind researchers (Banino et al. 2018), regarding navigation of artificial agents using grid-like cells.[28] The aim of the study was to examine whether the computational functions of grid cells can be leveraged to develop a deep-reinforcement learning agent with mammalian-like navigational abilities. After training the network with input mirroring the signals available to the mammalian brain, certain units (~25%) within the network developed what the authors refer to as "spatial activity profiles", which resemble the activity and functionality of grid cells. Pending more meticulous examination of the training and the details of these profiles, we can cautiously agree that these findings give a *certain degree of support* to the view that emulating the structure, functionality and principles of the biological brain will give rise to the desired cognitive abilities discussed above.

Having mentioned the difficulties and recent support, let us examine more closely the hidden assumptions behind

[24] See Hawkins 2004.

[25] See Hawkins et al. 2017.

[26] The debate revolving the value of Connectionism as a view that hopes to explain intellectual abilities using artificial neural networks has been going on for several decades now. For a comprehensive overview, see Garson and Buckner 2019.

[27] This study was limited to the examination of pyramidal neurons.

[28] A grid-like cell is a type of neuron in the brains of many species that allows them to understand their position in space.

the claim that emulating the human brain or the neocortex will bring about such higher order cognitive abilities as understanding, abductive reasoning, extended learning capabilities and more. Proponents of brain emulation claim that these abilities will *emerge* when we are able to construct an emulated neocortex. They use the concept of emergence in such manner: "Getting back to AGI… intelligence is an emergent phenomenon. It must emerge from the interactions of non-intelligent components" (Anderson 2011, §3); and: "The cortical algorithm can be deployed in novel ways, with novel senses, in a machined cortical sheet so that genuine, flexible intelligence emerges outside of biological brains." (Hawkins 2004: 38). What do they mean by that?

Emergentism is a view in the philosophy of mind, one of a variety of positions grouped under the more general position of Non-Reductive Materialism (NRM). Proponents of NRM hold that the mental is ontologically part of the material world; yet mental properties are causally efficacious without being reducible to physical properties. More specifically, proponents of Emergentism assert that consciousness, including all higher-order cognitive abilities, emerges from the physical brain. There are two types of emergence—weak and strong, and the difference between them boils down to our ability to infer the emergent property from knowledge of the lower-level components, their structure and their relationships. In the case of weak emergence, we may not be able to predict the emergence of a property, because we lack sufficient knowledge. However, in the case of strong emergence, we cannot infer the emergence of a property in principal, i.e., even when having complete knowledge of all low-level components, their structure, relationships, etc. Consciousness and higher-order mental abilities are considered to be the prevalent examples of strong emergence,[29] and so if proponents of the brain-emulation approach assume the emergence of such higher-order abilities as understanding, reasoning and creativity, they will have to face objections raised by several philosophers[30] who believe that strong emergence does not make sense.

Another issue we should examine is whether the analogical inference that the proponents of Emergentism make is justified. The inference is the following: the human neocortex is built in a certain way and has certain higher-order cognitive properties; therefore, if we emulate the neocortex, i.e., create an artificial entity or system which is very *similar* to the neocortex, the higher-order cognitive properties of the neocortex will emerge in this similar system as well. *Prima facie*, this inference appears justified. Analogical reasoning is based on the concept of similarity[31] and indeed it seems that by creating an artificial neocortex, which resembles the human neocortex, with a similar functionality of its basic components, a similar structure and similar relationships between its components, we can analogously infer the similarity of properties of the whole. However, beyond the question of how we define similarity, we should also mind issues of granularity and material. Granularity refers to the level of accuracy at which the emulated neocortex will be implemented. How detailed will the structure and functionality of the emulated neuron be in comparison to its biological counterpart? If we emulate the neocortex at the level of neuronal groups—a level that may suffice to achieve functionality—we may lose granularity and consequently similarity. This will make our analogical inference weaker, since it was originally based on similarity. The same applies to the structure, connections and relationships of larger neuronal groups. These implementation details matter greatly to the validity of the inference, for it may well be that what we consider as similar enough may not warrant the conclusion of the analogical argument. The other issue is material: the biological matter composing biological neurons could be of great importance to the whole in the sense that consciousness and higher-order mental abilities may all be properties of *biological brains*.[32] Hence, the material factor may also influence similarity and consequently the validity of the analogical inference. Moreover, it may warrant the inference completely false, for if higher-order cognitive abilities are properties of biological brains, then no matter how similar the emulated neocortex is to its biological counterpart, no higher-order properties will emerge. With these points in mind we can now conclude.

6 Concluding remarks

To sum up, both approaches may be dependent on advancements in other areas such as cognitive science, hardware development and perhaps quantum computing. However, we can still reach a few interim insights regarding the near future path AGI development should take: First, let us consider the brain emulation approach. Although it encompasses the dream and promise of building artificial brains, in its current state this approach suffers from multitude of problems—current projects are falling short of their declared goals, and more generally, as discussed above, this approach faces several major technical-scientific problems and is

[29] See, for example, Chalmers 2006.

[30] See, for example, Taylor 2015; Howell 2009; and Kim 1998, and 2006.

[31] In general, it is difficult precisely to define similarity relations, but see Helman 1988, especially the chapters by Stuart Russell ("Analogy by Similarity") and Ilkka Niiniluoto ("Analogy and Similarity in Scientific Reasoning").

[32] As proponents of Biological Naturalism claim. See Searle 2007.

founded on a rather weakly-justified philosophical basis. The cognitive architecture approach on the other hand, is based on decades of continuous research in cognitive and computer science and is concerned with building systems that implement at least some of the features discussed in this paper as required for AGI systems, e.g., understanding, learning and reasoning. To be sure, brain emulation projects can produce invaluable results and knowledge, not only in neuroscience but also in computer science and in the general study of artificial intelligence, but they require neurological knowledge that we currently do not possess.

Thus, it may seem more reasonable to pursue the development of more innovative and *hybrid* systems. In other words, combine the two above-mentioned approaches in a manner that highlights the strengths of each approach and mutually compensates for their weaknesses. Traditionally (although not without exceptions), the symbolic/rule-based or classical AI paradigm is identified with the cognitive architecture approach, whereas the neural networks paradigm is usually identified with approaches trying to simulate the workings of the human brain. Each of these paradigms has shortcomings: Deep learning (DL) is considered to be data inefficient, hard to generalize and uninterpretable,[33] whereas symbolic AI is limited when it comes to unstructured data, be it text, audio or images. It needs a formally specified set of inference rules to carry out logical-like reasoning, and, therefore, has problems when it comes to understanding the unordered and messy world out there; Combining the two approaches into a hybrid neuro-symbolic system, i.e., the reasoning power of rule-based AI with the learning capabilities of DL networks, may be the first step in the right direction to a general system. In such a hybrid system, symbolic components take advantage of the processing and analysis of unstructured data done by the DL components, while these components benefit from the reasoning power of the rule-based components, which enables them to learn new things with much less data.[34] Thus, I concur with Garnelo and Shanahan (2019: 23) that given such a hybrid system, "the desired properties of data efficiency, powerful generalisation, and human interpretability would likely follow." Moreover, I suggest that this is a first step down the hybrid road, leading to general systems having even more complex capabilities of independent learning and understanding of complex environments, abductive reasoning and creativity.

[33] See Garnelo and Shanahan (2019: 17).

[34] The Neuro-Symbolic Concept Learner is an example for such an hybrid system. See Mao et al. 2019.

References

Anderson M (2011) Reduction considered harmful. https://hplusmagazine.com/2011/03/31/reduction-considered-harmful/. Accessed July 2018

Banino A, Barry C, Uria B, Blundell C, Lillicrap T, Mirowski P, Kumaran D (2018) Vector-based navigation using grid-like representations in artificial agents. Nature 557:429–433

Boden MA (1998) Creativity and artificial intelligence. Artif Intell 103:347–356

Boden MA (2004) Creativity in a nutshell. In: Boden MA (ed) The creative mind: myths and mechanisms. Routledge, London, pp 1–10

Boden MA (2014) Creativity and artificial intelligence: a contradiction in terms? In: Paul ES, Kaufman SB (eds) The philosophy of creativity: new essays. Oxford University Press, Oxford

Brooks RA (1991) Intelligence without representation. Artif Intell 47:139–159

Chalmers D (2006) Strong and weak emergence. In: Clayton P, Davies P (eds) The re-emergence of emergence. Oxford University Press, Oxford, pp 244–255

Clark A (2016) Surfing uncertainty. Oxford University Press, Oxford

Douven I (2017) Abduction. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/abduction/. Accessed Nov 2018

Garnelo M, Shanahan M (2019) Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 29:17–23

Garson C, Buckner J (2019) Connctionism. Stanford encyclopedia of philosophy. https://plato.stanford.edu/entries/connectionism/. Accessed Dec 2019

Goertzel B (2006) The hidden pattern. Brown Walker Press, New York

Grimm S (2011) Understanding. In: Bernecker S, Pritchard D (eds) The routledge companion to epistemology. Routledge, New York.

Hawkins J (2004) On intelligence. Times Books, New York

Hawkins J, Ahmad S, Cui Y (2017) A theory of how columns in the neocortex enable learning the structure of the world. Front Neural Circuits 11:81

Helman DH (1988) Analogical reasoning. Springer, Dordrecht

Hohwy J (2013) The predictive mind. Oxford University Press, Oxford

Howell R (2009) Emergentism and supervenience physicalism. Australas J Philos 87:83–98

Kim J (1998) Mind in a physical world. MIT Press, Cambridge

Kim J (2006) Essays in metaphysics of mind. Oxford University Press, Oxford

Langley P (2012) The cognitive systems paradigm. Adv Cognit Syst 1:3–13

Legg S, Hutter M (2006) A formal measure of machine intelligence. In: Proc. 15th annual machine learning conference of Belgium and the Netherlands (Benelearn'06), pp. 73–80

Mao J, Gan C, Kohli P, Tenenbaum JB, Wu J (2019) The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. ICLR 2019 https://arxiv.org/abs/1904.12584v1. Accessed Dec 2019

McIlraith SA (1998) Logic-based abductive inference. https://www.cs.utoronto.ca/kr/papers/abduction.pdf. Accessed Nov 2018

Novitz D (1999) Creativity and constraint. Australas J Philos 77(1):67–82

Pritchard D (2009) Knowledge, understanding and epistemic value. R Inst Philos Suppl 64:19–43

Riggs W (2003) Understanding virtue and the virtue of understanding. In: DePaul M, Zagzebski L (eds) Intellectual virtue: perspectives from ethics and epistemology. Oxford University Press, Oxford

Rosenbloom P, Demski A, Ustun V (2016) The sigma cognitive architecture and system: towards functionally elegant grand unification. J Artif Gen Intelli 7(1):1–103

Sardi S, Vardi R, Sheinin A, Goldental A, Kanter I (2017) New types of experiments reveal that a neuron functions as multiple independent threshold units. Sci Rep 7:18036

Searle J (2007) Biological naturalism. In: Velmans M, Schneider S (eds) The blackwell companion to consciousness, Malden, MA. Blackwell Pub, Oxford, pp 325–334

Taylor E (2015) Collapsing emergence. Philos Q 65:732–753

Thórisson KR, Kremelberg D, Steunebrink BR, Nivel E (2016) About understanding. In: International conference on artificial general intelligence, Springer, New York, pp. 106–117

Van Fraassen BC (1980) The scientific image. Oxford University Press, Oxford

Voss P (2005) Essentials of general intelligence: the direct path to AGI. In: Goertzel B, Pennachin C (eds) Artificial general intelligence. Springer-Verlag, Berlin

Wilkenfeld D (2013) Understanding as representation manipulability. Synthese 190(6):997–1016

Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, Oxford

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.