*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002

*Review*

# From Homo Sapiens to Robo Sapiens: The Evolution of Intelligence

**Anat Ringel Raveh and Boaz Tamir**

Faculty of interdisciplinary studies, S.T.S. Program, Bar-Ilan University, Ramat-Gan, Israel

**Abstract:** In this paper we present a review of recent developments in AI towards the possibility of an artificial intelligence equals that of human intelligence. AI technology has always shown a stepwise increase in its capacity and complexity. The last step took place several years ago, with the increased progress in deep neural network technology. Each such step goes hand in hand with our understanding of ourselves and our understanding of human cognition. Indeed, AI was always about the question of understanding human nature. AI percolates into our lives, changing our environment. We believe that the next few steps in AI technology, and in our understanding of human behavior, will bring about much more powerful machines, flexible enough to resemble human behavior. In this context, there are two research fields: Artificial Social Intelligence (ASI) and General Artificial Intelligence (AGI). The authors also allude to one of the main challenges for AI, embodied cognition, and explain how it can viewed as an opportunity for further progress in AI research.

**Keywords:** Artificial Intelligence (AI); artificial general intelligence (AGI); artificial social intelligence (ASI); social sciences; singularity; complexity; embodied cognition; value alignment

---

## 1. Introduction

### *1.1. From intelligence to super-intelligence*

In this paper we present a review of recent developments in AI towards the possibility of an artificial intelligence equals that of human intelligence. So far AI technology has shown a stepwise increase in its capacity and in its complexity (section 2). The last step took place several years ago, due to increased progress in deep neural network technology. Each such step goes hand in hand with our understanding of ourselves and our understanding of human cognition (section 3). Indeed AI was always about the question of understanding human nature.

There is still a long way to go before we can talk about a singularity point. AI is still a weak technology, still too rigid, and too specific to become similar to human intelligence. However, we believe the next few steps in AI technology and in our understanding of human behavior, will bring about much more powerful machines, flexible enough to resemble human behavior. An important major research project in this context is Artificial Social Intelligence (ASI), which we shall shortly describe (section 4). The second project is a new challenge which is known as Artificial General Intelligence (AGI) (section 5). AGI brings about a new approach which is much more flexible and closer to human intelligence. It also suggests a model of consciousness, a new approach to the question of learning, a model of self-referential machines, etc.

One of the biggest challenges for AI is that of embodied cognition. If AI could surpass this hurdle, it will get even closer to true human intelligence. Embodied cognition was first presented as the main reason why AI is impossible. We propose to view embodied cognition as a new challenge to AI and not as an imposition (section 6).

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



We end the discussion by demonstrating a way to overcome our fears of singularity, by the process of value alignment, which is expanded upon in section 7.

### *1.2. The Human-Machine eco-system*

From ancient myths of inanimate objects coming alive to the creation of artificial intelligence, philosophers, scientists, writers, and artists have pondered the very nature and boundaries of humanity. Humans are fascinated by machines that can imitate us but also feel an existential discomfort around them- an uneasiness that stems from their ability to obscure the line between the living and the inanimate.

Claude Levi-Strauss (1969) has examined how the individual process of constructing reality is related to how an entire society develops and maintains its worldview. He argued that the most common way in which both an individual and a community put together a structure of reality is through the use of binary categories. An individual makes sense of the world by organizing things in a series of dual oppositions such as dark/light, living/dead, feminine/masculine, emotion/logic, and so on, which lead to the community's development of more abstract concepts like, chaos/order, natural/unnatural, normal/abnormal, subjectivity/objectivity, and moral/immoral. Such a predetermined schema of reality provides the confidence people need to face the world and explore its boundaries.

As far as our relationship to thinking machines is concerned, it seems that the worldview we have developed for ourselves over time has become pessimistic as the pace of technology development increased. While in the past automata have entertained us mainly because they mimicked human behavior in an inaccurate and ridiculous way that revealed the fact that it was a trick, artificially intelligent machines today can successfully mimic an increasing number of the human's traits, such as natural human language and thought patterns. These traits have always separated us from all the other living creatures; and the fact that this primal distinction between human beings and technology is blurrier than it has ever been, mostly creates fear (Kang 2011).

However, as Strate (2017) explains, "At the very least what ought to be clear is that the physical universe and the biophysical environment are not entirely different and distinct from technology, but are part of a continuum." (p. 70). The view underling this argument is an ecological or systems view which "emphasize the interdependence and interactive relationships that exist, as all forms of life alter their conditions simply by their very presence, by their metabolism, for example, and through their reproduction." (p. 62).

Hence, instead of the dichotomous narrative of us- human beings- vs. them- super-intelligent machines, we should understand both ourselves and AI as parts of a complex and dynamic eco-system. While it might be that we are only a stepping stone on the path of the universe towards an even greater complexity, we have an important and special role; as argued by McLuhan (1964), technologies and media "also depend upon us for their interplay and their evolution" (p. 57), and if we carefully examine their actions, we will see that there is no reason to fear.

## 2. State of the Art: Stepwise incremental AI

In our view, the best way to describe the developmental process of AI so far is as a stepwise incremental progress, and using an analogy from physics, AI "percolates" into our lives. It could indeed make a phase transition into a higher level of complexity above our own, a process we will shortly discuss. But first we want to describe the ongoing process of stepwise incremental AI.

In January 2017 McKinsey (Manyika et al. 2017) published a comprehensive report that maps the progress of artificial intelligence in a variety of areas. The five areas, described herein, are further broken down into other sub-tasks and sub-capabilities:

***Sensory perception***- "This includes visual perception, tactile sensing, and auditory sensing, and involves complex external perception through integrating and analyzing data from various sensors in the physical world." (Ibid, p. 34). In this area the performance level is median compared to humans.

Take machine vision as an example. The creation and assimilation of visual capabilities that surpass human vision in cameras, have been relatively easy. However, the more complex part was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



to add AI technology to the cameras. One such project is Landing.ai (https://www.landing.ai/), formed by Andrew Ng, a globally recognized leader in AI. This startup focuses on solving manufacturing problems such as Quality Control (QC). It offers machine-vision tools to detect microscopic defects that may be found in products such as circuit boards, and which the human eye simply cannot detect.

Another recent and interesting project deals with machine touch. In the paper "Learning Dexterous In-Hand Manipulation" (Andrychowicz et al. 2018), a team of researchers and engineers at OpenAI have demonstrated that in-hand manipulation skills learned with reinforcement learning in a simulator can evolve into a fairly high level of dexterity of the robotic hand. As they explain: "This is possible due to extensive randomizations of the simulator, large-scale distributed training infrastructure, policies with memory, and a choice of sensing modalities which can be modelled in the simulator." (p. 15). The researchers' method "did not rely on any human demonstrations, but many behaviors found in human manipulation emerge naturally, including finger gaiting, multi-finger coordination, and the controlled use of gravity". (p. 1).

***Cognitive capabilities***- "A range of capabilities is included in this category including recognizing known patterns and categories (other than through sensory perception); creating and recognizing novel patterns and categories; logical reasoning and problem solving using contextual information and increasingly complex input variables; optimization and planning to achieve specific objectives given various constraints; creating diverse and novel ideas or a novel combination of ideas; information retrieval, which involves searching and retrieving information from a large range of sources; coordination with multiple agents, which involves interacting with other machines and with humans to coordinate group activity; and output articulation and presentation, which involves delivering outputs other than through natural language. These could be automated production of pictures, diagrams, graphs, or mixed media presentations." (Manyika et al. 2017, p. 34).

By using these capabilities, AI can amplify our own abilities: "Artificial intelligence can boost our analytic and decision-making abilities by providing the right information at the right time. But it can also heighten creativity". (Chui & Francisco 2017, p. 34). Consider for example Autodesk's Dreamcatcher AI which enhances the imagination of designers. As explained in the company's website:

"Dreamcatcher is a generative design system that enables designers to craft a definition of their design problem through goals and constraints. This information is used to synthesize alternative solutions that meet the objectives. Designers are able to explore trade-offs between many alternative approaches and select design solutions for manufacture." (https://autodeskresearch.com/projects/dreamcatcher)

Some of the cognitive capabilities **have achieved human level performance** "such as recognizing simple/complex known patterns and categories other than sensory perception; Search and retrieve information from a large scale of sources- breadth, depth, and degree of integration." (Manyika et al. 2017, p. 35). However, other capabilities **are currently below median performance** such as create and recognize new patterns/categories; solve problems in an organized way using contextual information and increasingly complex input variables other than optimization and planning; create diverse and novel ideas, or novel combinations of ideas. (Ibid).

***Natural language processing***- "This consists of two distinct parts: natural language generation, which is the ability to deliver spoken messages, including with nuanced human interaction and gestures, and natural language understanding, which is the comprehension of language and nuanced linguistic communication in all its rich complexity." (Ibid, p. 34). As for Natural language generation, although there is progress in this area (such as Google duplex), the levels of performance according to the report are at best median. When it comes to Natural Language understanding, there is a long way ahead of us.

Yet, an example for an effective implementation of these capabilities (and more) is Aida (http://aidatech.io/), a virtual assistant that is being used by SEB, a major Swedish bank. Aida interacts with masses of customers through natural-language conversations, and therefore has access to vast amounts of data. "This way she can answer many frequently asked questions, such as how to open

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



an account or make cross-border payments. She can also ask callers follow-up questions to solve their problems, and she's able to analyze a caller's tone of voice and use that information to provide better service later." (Wilson and Daugherty 2018, para. 16).

***Physical capabilities***- This includes gross motor skills, navigation (these two have reached human level performance), fine motor skills and mobility (these are more difficult and hence the performance levels are currently still median and below). "These capabilities could be implemented by robots or other machines manipulating objects with dexterity and sensitivity, moving objects with multidimensional motor skills, autonomously navigating in various environments and moving within and across various environments and terrain." ((Manyika et al. 2017, p. 35).

While AIs like Cortana are essentially digital entities, there are other applications where "intelligence is embodied in a robot that **augments** a human worker. With their sophisticated sensors, motors, and actuators, AI-enabled machines can now recognize people and objects and work safely alongside humans in factories, warehouses, and laboratories." (Wilson and Daugherty 2018, para. 16, our emphasis).

"Cobots" are probably the best example here. Collaborative robots, as Gonzalez (2018) explains, "excel because they can function in areas of work previously occupied only by their human counterparts. They are designed with inherent safety features like force feedback and collision detection, making them safe to work right next to human operators." (para. 2).

Based on a white paper that Universal Robots- one of the leading companies in the robot market- has published (2018), Gonzalez lists the seven most common applications for Cobots. One of them, for example, is "pick and place": "A pick and place task is any in which a workpiece is picked up and placed in a different location. This could mean a packaging function or a sort function from a tray or conveyor; the later [sic] often requires advanced vision systems." (para. 3).

***Social and emotional capabilities***- "This consists of three types of capability: social and emotional **sensing**, which involves identifying a person's social and emotional state; social and emotional **reasoning**, which entails accurately drawing conclusions based on a person's social and emotional state, and determining an appropriate response; and social and emotional **action**, which is the production of an appropriate social or emotional response, both in words and through body language." (Manyika et al. 2017, p. 34).

Let us Consider Mattersight as an example. The company provides a highly sophisticated data analysis system that tracks customer' responses on the telephone. The software analyzes varied communicative micro-features such as, tone, volume, word choice, pauses, and so on. Then, in a matter of a few seconds, A.I algorithms interpret these features, compare them to the company's databases, and come up with a personality profile for each customer. Based on this profile, the customer will be referred to the most appropriate service agent for him. (Stevens 2013).

To sum-up the report, Manyika et al. (2017) notes that from a mechanical point of view, they are fairly certain that perfection can be achieved. Because, already today, through deep reinforcement learning for example, robots can untie shoelaces and remove a nail from the back of a hammer. However, from the cognitive point of view, although the robot's "intelligence", has progressed, this is still where the greatest technical challenge lie:

"While machines can be trained to perform a range of cognitive tasks, they remain limited. They are not yet good at putting knowledge into context, let alone improvising. They have little of the common sense that is the essence of human experience and emotion. They struggle to operate without a pre-defined methodology. They are far more literal than people, and poor at picking up social or emotional cues. They generally cannot detect whether a customer is upset at a hospital bill or a death in the family, and for now, they cannot answer "What do you think about the people in this photograph?" or other open-ended questions. They can tell jokes without really understanding them. They don't feel humiliation, fear, pride, anger, or happiness. They also struggle with disambiguation, unsure whether a mention of the word "mercury" refers to a planet, a metal, or the winged god of Roman mythology. Moreover, while machines can replicate individual performance capabilities such

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



as fine motor skills or navigation, much work remains to be done integrating these different capabilities into holistic solutions where everything works together seamlessly." (Ibid, pp. 26-7).

## 3. A.I. goes hand in hand with our understanding of ourselves

Singularity is based on several assumptions: first, that there is a clear notion of what is human intelligence; and second, that AI can decrease the gap between human intelligence and machine intelligence. However, both of these assumptions are not clear yet. What is becoming more and more apparent is that AI goes hand in hand with our understanding of our own human intelligence and behavior.

"Intelligence" is a complex and multifaceted phenomenon that has for years interested researchers from countless fields of study. Among others, intelligence is studied from psychological, biological, economical, statistical, engineering, and neurological perspectives. New insights emerge over time from the various disciplines, many of which are adopted into the science of AI and contribute to its development and progress. The most striking example is the special and fruitful interrelationship between artificial intelligence and cognitive science.

Cognitive science and artificial intelligence arose at about the same time, in the late 1950s, and grew out of two main developments: "(1) the invention of computers and the attempts soon thereafter to design programs that could do the kinds of tasks that humans do, and (2) the development of information-processing psychology, later called cognitive psychology, which attempted to specify the internal processing involved in perception, memory, and thought. Cognitive science was a synthesis of the two, concerned both with the details of human cognitive processing and with the computational modeling of those processes." (Collins and Smith 1988, p. 1).

What AI does best is **analyze, categorize, and find the relationships** between large amounts of data, quickly and very effectively, coming up with highly accurate predictions. These capabilities, as Collins and Smith explained in 1988 , were the outcome of three foci that have turned out to be three major bases for progress in AI: **formalisms**- such as mean-ends analysis, which are standard methods for representing and implementing cognitive processes; **tools or languages** for building intelligent programs such as John McCarthy's LISP (1978); and **programs**- beginning with the Dendral project (Lindsay, Buchanan, Feigenbaum and Lederberg 1980), the first expert system to allow the formation of a scientific hypothesis.

"By contrast, psychology historically has made progress mainly by accumulating empirical phenomena and data, with far less emphasis on theorizing of the sort found in artificial intelligence. More specifically, psychological theories have tended to be constructed just to explain data in some experimental paradigm, and have tended to be lacking a well-founded mechanism, probably a relic of the behavioristic or stimulus-response approach that dominated psychology from the 1920s through the 1950s." (Collins and Smith 1988, p. 2).

At first, AI was mistakenly identified with the mechanical psychological viewpoint of behaviorism. The physicalism of stimulus and response looked similar to the action of computers, a reduction of man into its gears (Boden [1977] 1978). In psychology, a more 'humanistic' view of the science was demanded. It was agreed by all 'humanistic' psychologists that a good theory should be irreducible, its terms cannot be reduced to simple physical constituents, and that terms such as 'intention' should have a major part in the theory. Moreover, any action should have some meaning to the actor, and the meaning should be subjective. The 'humanistic' approach to psychology was a scientific revolution against positivistic psychology (in the Kuhnian sense, Ibid, p. 396). It turned out that AI came to be very similar to the 'humanistic' viewpoint. Both AI and cognitive science were beginning to ask similar questions and to use many similar terms.

What was needed for a science of cognition was a much richer notion of knowledge representation and process mechanisms; and that is what artificial intelligence has provided. Cognitive psychologists gained a rich set of formalisms to use in characterizing human cognition. (Collins and Smith 1988, p. 2). Some of the early and most important formalisms were *means-ends analysis* (Newell and Simon 1963), *Discrimination nets* (Feigenbaum 1963), *Semantic networks* (Quillian 1968), *Frames and scripts* (Minsky 1975; Schank and Abelson 1977), *Production systems* (Newell and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



Simon 1972), *Semantic primitives* (Schank 1972; Norman and Rumelhart 1975), *Incremental qualitative analysis* (de Kleer 1979). Through the years a wide range of formalisms were developed for analyzing human cognition, and many of them are still in use today.

Moreover, artificial intelligence has become a kind of theoretical psychology. Researchers who sought to develop a psychological theory could become artificial intelligence researchers without making their marks as experimentalists. Thus, as in physics, two branches of psychology were formed- experimental and theoretical- and cognitive science has become the interface where theorists and experimentalists sort things out. (Collins and Smith 1988).

Boden ([1977] 1978) suggests that AI can be used as a test-lab for cognitive science. It raises and exposes psychological questions that were deeply implicit. It suggests new terms, ideas and questions that were otherwise hidden. In that sense we dare say that computation is playing the role of language for cognitive science. Similar to the role of mathematics in physics, computation has become a language for constructing theories of the mind. Computation is a formal language that imposes a set of constraints on the kind of theories that can be constructed. But unlike mathematics, it has several advantages for constructing psychological theories: while mathematical models are often static, computational models are inherently process-oriented; while mathematical models, particularly in psychology, are content-independent, computational models can be content-dependent; and while computational models are inherently goal-oriented, mathematics is not. (Collins and Smith 1988).

Questions that we should ask includes; is the use of the same terms and the same language in AI and cognitive sciences only an analogy? Could it imply something deeper? Can we insert true 'intention' and true 'meaning' into computer agents? How can we define such terms in AI? In fact, this is the main question of strong AI. This would bring AI and cognitive science much closer.

In an attempt to answer these questions, we refer to the viewpoint of Dennett (2017). Let's define the notion of 'meaning'; to put things very simplistically, we will say that an action of a computer agent has a 'meaning' (for the agent) if the action is changes some part of its environment and the agent can sense that change. For example, if the agent is a ribosome, then the transcription of an RNA into a series of amino-acids, later to become a protein, has a meaning since the protein has some function in changing the agent's environment. The action of the ribosome has a 'meaning' in the cytoplasm environment. Similarly, we can embed a 'meaning' in computer agents. It was suggested by Dennett that we human can insert a derived 'intention' in computers, and computers can derive a lower type of 'intention' in other computers. This was also brought up years ago by Minsky (2006), using a different language.

It was suggested by Boden ([1977] 1978) that we can bridge the gap between 'humanistic' approach to cognitive science (in the sense discussed above) and physical mechanism. The way to do so is by introducing an inner representation of the self into the computer. Intentionality and meaning could be aimed (given a context) into this inner representation; the reduction or mechanism of the intentionality will be enabled by the design or architecture of the inner representation. Hence, in order to describe what is going on in the computer, the language of intentionality will be the most appropriate, in the same sense that we talk about our dog's intentions when we wish to describe or explain its behavior, without the need for a behavioristic language, or other physical terms. It will not be 'natural` or efficient to describe the action of the computer in the language of the state of its switches, we will say that this particular action was 'intended for' to comply with the 'state of mind' that the computer had. This sounds a somewhat pretentious goal, however it is based on the assumption that any future advancement in AI must stand on a basic cognitive architecture, much more basic and deeper than what we have today.

Most of the recent progress in AI have been driven by deep neural networks and these are related to the "connectionist" view of human intelligence. Connectionist theories essentially perceive learning—human and artificial— as rooted in interconnected networks of simple units, either real neurons or artificial ones, which detect patterns in large amounts of data. Thus, some in the machine learning field are looking to psychological research on human learning and cognition to help take AI to that next level. Although the concept of neural networks has existed since the 1940s, only today, due to an enormous increase in computing power and the amount and type of data available to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



analyze, deep neural networks have become increasingly powerful, useful and ubiquitous. (Winerman 2018).

The theory of consciousness was recently investigated by AI researchers. It was suggested by Dennett (2017) that consciousness is an emergent property of many small processes, or agents, each struggle for its homogeneity. Consciousness is not a stage with spotlights in which all subconscious processes are the audience. Consciousness is a dynamical arena where many agents appear and soon disappear. It resembles an evolutionary process occurring in a very short timescale (Calvin [1996] 2001). On this very basis a few AI models were suggested, the Copycat model (Hofstadter 1995) and its more advanced Learning Intelligent Distribution Agent (LIDA) (Fagihihi and Franklin 2012) model. These two are examples of a strong reciprocal interaction between AI and cognitive science.

Similar reciprocal relationships are now beginning to form between social sciences and artificial intelligence to become the field of artificial social intelligence (ASI). ASI is an interdisciplinary science, which was introduced years ago by Brent and others (Bainbridge et al. 1994), and is only now becoming prevalent. ASI is a new challenge for social science and a new arena for the science of AI. It deals with the formalization of delicate social interactions, using it in AI to implement social behavior into robots. The prospects for social scientists were suggested years ago by Anderson (1989):

”It is time for sociology to break its intellectual isolation and participate in the cognitivist rethinking of human action, and to avail itself of theoretical ideas, techniques and tools that have been developed in AI and cognitive science “ (p. 20).

“My argument is that sociologists have a great deal to learn from these disciplines, and that the adoption of concepts, methods and tools from them would change sociologists’ working habits […]” (p. 215).

## 4. ASI, a new challenge

While artificial cognitive intelligence has become a well-established and significant field of research, and has been heavily invested by both cognitive and artificial intelligence researchers, artificial social intelligence is in its early stages and has great potential for the advancement of smart machines in a new and essential way.

While cognitive artificial intelligence scientists “essentially view the mind as something associated with a single organism, a single computational system, social psychologists have long recognized that this is just an approximation. In reality the mind is social, it exists, not in isolated individuals, but in individuals embedded in social and cultural systems.” (Pennachin and Goertzel 2007, p. 24).

It is well established now that there are sets of brain regions that are dedicated to social cognition. It was first shown on primates (Brothers 1990) and later on humans (Adolphs 2003). As Frith (2007) explains: “The function of the social brain is to enable us to make predictions during social interactions.” (p. 67). The social brain includes a variety of mechanisms, such as the amygdala which is activated in case of fear. It is also connected with the mechanism of prejudice, stereotyping, associating values with stimuli. It concerns both people- individual and group- and objects. Another such mechanism is the medial prefrontal cortex, which is connected with the understanding of the other’s behavior in terms of its mental state, with long term dispositions and attitudes, and with self-perception about long term attitudes.

From the social point of view, Mead (1934), in his book, *Mind, Self and Society*, defines the “social organism” as “a social group of individual organisms” (p. 130), or in modern language, as an emergent phenomenon. This means that each individual, as an organism in itself, is also a part of a larger system, the social organism. Hence, each individual’s act must be understood within the context of some social act that involve other individuals. The social act is therefore viewed as a dynamic and complex system within which the individual is situated. As such, the social `organism’ actually defines the individual acts, that is, within it these acts become meaningful.

In his book *Artificial Experts* (1990), Collins argues similarly **that intelligence cannot be defined without considering social interactions.** This is because “[…] the locus of knowledge appears to be not the individual but the social group; what we are as individuals is but a symptom of the groups

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



in which the irreducible quantum of knowledge is located. Contrary to the usual reductionist model of the social sciences, it is the individual who is made of social groups." (p. 6).

Our intelligence, as Yudkowsky (2007) clarifies, "includes the ability to model social realities consisting of other humans, and the ability to predict and manipulate the internal reality of the mind." (p. 389). Another way to put it is through Mead's concept of the 'Generalized other' (1934). As Dodds, Lawrence & Valsiner (1997) explain, "**to take the role of the other involves the importation of the social into the personal**, and this activity is crucial for the development of self-consciousness and the ability to operate in the social world. It describes how perspectives, attitudes and roles of a group are incorporated into the individual's own thinking in a way that is distinct from the transmission of social rules, and in a way that can account for the possibility of change in both person and society." (p. 495, our emphasis).

Hence, as Collins (1990) argues, "The organism into which the intelligent computer supposed to fit is not a human being but a much larger organism; a social group. The intelligent computer is meant to counterfeit the performance of a whole human being within a social group, not a human being's brain. **An artificial intelligence is a '*social prosthesis*'.**" (p. 14, our emphasis).

All the above suggests the emergence of a new interdisciplinary discipline. The main concern of this new field of science is the formalization of delicate social modules, using them in AI to implement social awareness (perhaps a type of social common sense understanding) and social behavior into robots. Because of the dynamic nature of social interactions, these ASI systems face difficult challenges, some of which are not even predictable. In order to address these challenges, ASI systems will have to be dynamic by continuously reviewing and evolving their interaction strategies in order to adapt to new social situations. Moreover, it is essential to examine and assess these strategies in as many contexts as possible, in which ongoing, continuous interactions are taking place.

For making ASI come true, there are some fundamental steps which needs to be solved (Microsoft Research India workshop on ASI 2007). Firstly, there is a need to discover the principles of socio-culture interactions in which the ASI system could have a role. In order to formulate those principles there is considerable importance for conducting large data-driven studies aimed at validating these principles, as well as identifying and characterizing new behavioral traits. Such studies are already being conducted, using the enormous amounts of socially grounded user data generated and highly available from social media; as well as the significant advancements in machine learning and the wide variety of data-analysis techniques. One such project is "Mark my words!" (Danescu-Niculescu-Mizil, Gamon and Dumais 2011). This project demonstrates the psycholinguistic theory of communication accommodation according to which participants in conversations tend to adapt to the communicative behavior patterns of those with whom they converse. The researches have shown "that the hypothesis of linguistic style accommodation can be confirmed in a real life, large scale dataset of Twitter conversations." (p. 754). A probabilistic framework was developed, which allowed the researchers to measure "accommodation and, importantly, to distinguish effects of style accommodation from those of homophily and topic-accommodation." (Ibid).

Once the relevant socio-cultural principles have been extracted and defined, the next step will be to understand how they can be assimilated into ASI systems such as chatbots, recommender systems, autonomous cars, etc. One such system is the *virtual receptionist*, "which keeps track of users attention and engagement through visual cues (such as gaze tracking, head orientation etc.) to initiate the interaction at the most appropriate moment (Bohus, Andrist & Jalobeanu 2017). Further, it can also make use of hesitation (e.g., "hmmm… uhhh") to attract the attention of the user, buy time for processing or even to indicate uncertainty in the response (Bohus and Horvits 2014)." (Microsoft Research India workshop on ASI, 2007, para. 6).

ASI systems have no clear definition of goals, there is no specific task the machine is oriented towards. In a sense, the machine's social behavior is the goal. In other words, it is impossible to defined clear goals in advance, and these may even emerge dynamically. This means that measurement and evaluation methods are very difficult to apply to a socio-cultural intelligence of such a system. This is one of the biggest challenges the ASI field has to deal with.

## 5. AGI, an overview, is it enough?

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



An important concept to dwell on is that of artificial general intelligence (AGI). AGI constitute a new step towards strong AI. General intelligence is not a fully well-defined term, but it has a qualitative meaning: "What is meant by AGI is, loosely speaking, AI systems that possess a reasonable degree of self-understanding and autonomous self-control, and have the ability to solve a variety of complex problems in a variety of contexts, and to learn to solve new problems that they didn't know about at the time of their creation." (Pennachin and Goertzel 2007, p. VI).

There is a clear distinction between AGI and narrow AI research. The latter is aimed at creating programs that specialize in performing specific tasks, such as ordering online shopping, playing GO, diagnosing diseases or driving a car. But, despite their great importance and popularity, narrow AIs core problem is that "they are inherently narrow (narrow by design) and fixed. Whatever capabilities they have, are pretty much frozen in time. It is true that narrow AI can be designed to allow for some limited learning or adaptation once deployed, but this is actually quite rare. Typically, in order to change or expand functionality requires either additional programming, or retraining (and testing) with a new dataset." (Voss 2017, para. 4-5).

Intelligence, in general, "implies an ability to acquire and apply knowledge, and to reason and think, in a variety of domains" (Goertzel and Pennachin 2007, p. 15). In other words, intelligence in its essence has a large and dynamical spectrum.

"Narrow AI systems cannot adapt dynamically to novel situations — be it new perceptual cues or situations; or new words, phrases, products, business rules, goals, responses, requirements, etc. However, in the real world things change all the time, and intelligence is by definition the ability to effectively deal with change." (Voss 2017, para. 6).

Artificial general intelligence requires the above characteristics. It must be capable of performing various tasks in different contexts, making generalizations and tapping from existing knowledge in a given context to another. Hence, as Voss (2017) explains, "it must embody at least the following essential abilities:

1. To autonomously and interactively acquire new knowledge and skills, in real time. This includes one-shot learning — i.e. learning something new from a single example.
2. To truly understand language, have meaningful conversation, and be able to reason contextually, logically and abstractly. Moreover, it must be able to explain its conclusions.
3. To remember recent events and interactions (short-term memory), and to understand the context and purpose of actions, including those of other actors (theory of mind).
4. To proactively use existing knowledge and skills to accelerate learning (transfer learning).
5. To generalize existing knowledge by forming abstractions and ontologies (knowledge hierarchies).
6. To dynamically manage multiple, potentially conflicting goals and priorities, and to select the appropriate input stimuli and to focus on relevant tasks (focus and selection).
7. To recognize and appropriately respond to human emotions (have EQ, emotional intelligence), as well as to take its own cognitive states — such as surprise, uncertainty or confusion — into account (introspection).
8. Crucially, to be able to do all of the above with limited knowledge, computational power, and time. For example, when confronted with a new situation in the real world, one cannot afford to wait to re-train a massive neural network over several days on a specialized supercomputer." (para. 12).

In conclusion, general intelligence is a complex phenomenon that emerges from the integration of several essential components. "On the structural side, the system must integrate sense inputs, memory, and actuators, while on the functional side various learning, recognition, recall and action capabilities must operate seamlessly on a wide range of static and dynamic patterns. In addition, these cognitive abilities must be conceptual and contextual – they must be able to generalize knowledge, and interpret it against different backgrounds." (Voss 2007, p. 147).

From the point of view of strategy and methodology AGI sometimes uses a top down approach on cognition, as Wang and Goertzel (Ibid) explains, "An AGI project often starts with a blueprint of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



a whole system, attempting to capture intelligence as a whole. Such a blueprint is often called an "architecture"." (p. 5).

Cognitive architecture (CA) research "models the main factors participated in our thinking and decision and concentrates on the relationships among them. In computer science, CA mostly refers to the computational model simulating human's cognitive and behavioral characteristics. Despite a category of loose definition, CAs usually deal with relatively large software systems that have numerous heterogeneous parts and subcomponents. Typically, many of these architectures are built to control artificial agents, which run both in virtual worlds and physical robots." (Ye, Wang and Wang 2018, p.1).

Symbolic systems are one important type of cognitive architecture. "This type of agents maintains a consistent knowledge base by representing the environment as symbols." (Ibid, p. 2). Some of the most ambitious AGI-oriented projects in the history of the field were in the symbolic-AI paradigm. One such famous project is the General Problem Solver (Newell and Simon 1961), which used heuristic search (means-ends analysis) to solve problems. Another famous effort was the CYC project (Lenat 1995). The project's aim was to create human-like AI by collecting and encoding all human common sense knowledge in first order logic. Alan Newell's SOAR project (1987) was an attempt to create unified cognition theories, based on "logic-style knowledge representation, mental activity as problem-solving carried out by an assemblage of heuristics, etc." (Pennachin and Goertzel 2007, p. 3). However, the system was not constructed to be fully autonomous or to have self-understanding. (Ibid).

These and other early attempts failed to reach their original goals, and in the view of most AI researchers, failed to make dramatic conceptual or practical progress toward their goals. Some (GPS for example) failed because of exponential growth in computational complexity. However, more contemporary AGI studies and projects offer new approaches, combining the previous knowledge- both theories and research methods- accumulated in the field.

One such integrative scheme described by Pennachin and Goertzel (2007), was given the name 'Novamente'. This scheme involves taking elements from various approaches and creating an integrated and interactive system. However, as the two explain: "This makes sense if you believe that the different AI approaches each capture some aspect of the mind uniquely well. But the integration can be done in many different ways. It is not workable to simply create a modular system with modules embodying different AI paradigms: the different approaches are too different in too many ways. Instead one must create a unified knowledge representation and dynamics framework, and figure out how to manifest the core ideas of the various AI paradigms within the universal framework." (p. 5).

In their paper, "Novamente: an integrative architecture for Artificial Intelligence" (2004), Goertzel et al. suggest such an integrative AI software system. The Novamente design incorporates evolutionary programming, symbolic logic, agent systems, and probabilistic reasoning. The authors clarify that "in principle, integrative AI could be conducted in two ways: Loose integration, in which different narrow AI techniques reside in separate software processes or software modules, and exchange the results of their analysis with each other. Tight integration, in which multiple narrow AI processes interact in real-time on the same evolving integrative data store, and dynamically affect one another's parameters and control schemata. Novamente is based on a distributed software architecture, in which a distributed processing framework called DINI (Distributed Integrative Intelligence) is used to bind together databases, information-gathering processes, user interfaces, and "analytical clusters" consisting of tightly-integrated AI processes." (p. 2).

Novamente is extremely innovative in its overall architecture, which seeks to deal with the difficulty of creating a "whole brain" in a completely new and direct way. The basic principles on which the design of the system is founded are derived from the "psynet model"- an innovative complex-systems theory of mind- which was developed by Goertzel (1993a; 1993b; 1994; 1997; 2001). "What the psynet model has led us to is not a conventional AI program, nor a conventional multi-agent-system framework. Rather, we are talking about an autonomous, self-organizing, self-evolving AGI system, with its own understanding of the world, and the ability to relate to humans on a "mind-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



to-mind" rather than a "software-program-to-mind" level." (Pennachin and Goertzel 2007, pp. 64-65).

Another interesting project is the Learning Intelligent Distribution Agent (LIDA) (Ramamurthy, Baars, D'Mello and Franklin 2006). The LIDA architecture is presented as a working model of cognition, a Cognitive Architecture, which was designed to be consistent with what is known from cognitive sciences and neuroscience. Ramamurthy et al. argue "that such working models are broad in scope and could address real world problems in comparison to experimentally based models which focus on specific pieces of cognition. […] A LIDA based cognitive robot or software agent will be capable of multiple learning mechanisms. With artificial feelings and emotions as primary motivators and learning facilitators, such systems will 'live' through a developmental period during which they will learn in multiple ways to act in an effective, human-like manner in complex, dynamic, and unpredictable environments." (P. 1).

In a nutshell, LIDA is a modified version of the old COPYCAT architecture suggested years ago by Hofstadter (1995). It is based on the attempt to understand consciousness as a working space for many agents. The agents compete one another and those that dominate the workspace are identified as the ones that constitute our awareness. The process is dynamic, information flows in from the environment, and action is decided by a set of heuristics, which are themselves dynamic.

The LIDA architecture is partly symbolic and partly connectionist; part of the architecture "is composed of entities at a relatively high level of abstraction, such as behaviors, message-type nodes, emotions, etc., and partly of low-level codelets (small pieces of code). LIDA's primary mechanisms are perception, episodic memory, procedural memory, and action selection." (Ramamurthy, Baars, D'Mello and Franklin 2006, p. 1).

With the design of three continually active incremental learning mechanisms- perceptual learning, episodic learning and procedural learning- the researchers have laid the foundation for a working model of cognition that produces a cognitive architecture capable of human like learning. As the authors (Ibid) explain:

"The architecture can be applied to control autonomous software agents as well as autonomous robots "living" and acting in a reasonably complex environment. The perceptual learning mechanism allows each agent controlled by the LIDA architecture to be suitably equipped so as to construct its own ontology and representation of its world, be it artificial or real. And then, an agent controlled by the LIDA architecture can also learn from its experiences, via the episodic learning mechanism. Finally, with procedural learning, the agent is capable of learning new ways to accomplish new tasks by creating new actions and action sequences. With feelings and emotions serving as primary motivators and learning facilitators, every action, exogenous and endogenous taken by an agent controlled with the LIDA architecture is self-motivated." (p. 6).

A third project worth mentioning is Schmidhuber's Gödel Machines (2006). Schmidhuber describe these machines as "the first class of mathematically rigorous, general, fully self-referential, self-improving, optimally efficient problem solvers. Inspired by Kurt Gödel's celebrated self-referential formulas (1931), such a problem solver rewrites any part of its own code as soon as it has found a proof that the rewrite is *useful,* where the problem-dependent utility function and the hardware and the entire initial code are described by axioms encoded in an initial proof searcher which is also part of the initial code. The searcher systematically and in an asymptotically optimally efficient way tests computable *proof techniques* (programs whose outputs are proofs) until it finds a provably useful, computable self-rewrite." (p.1).

In other words, the Gödel machines "are universal problem solving systems that interact with some (partially observable) environment and can in principle modify themselves without essential limits apart from the limits of computability. Their initial algorithm is not hardwired; it can completely rewrite itself, but only if a proof searcher embedded within the initial algorithm can first prove that the rewrite is useful, given a formalized utility function reflecting computation time and expected future success (e.g., rewards)." (p. 2).

A completely different approach to AGI suggests imitating the complex architecture of the human brain and creating its exact digital simulation. However, this method is questionable since the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



brain has not been fully deciphered yet. Another, more abstract way to create AGI is to follow cognitive psychology research and to emulate the human mind. A third way is to create AGI by emulating properties of both aspects- brain and mind. But, as Wang (2012) stresses, the main issue is not "whether to learn from the human brain/mind (the answer is always "yes", since it is the best-known form of intelligence), or whether to idealize and simplify the knowledge obtained from the human brain/mind (the answer is also always "yes", since a computer cannot become identical to the brain in all aspects), but on *where* to focus and *how much* to abstract and generalize." (pp. 212-213).

One of the unsolved problems of AGI research is our lack of understanding of the definition of "Generalization", but what Perez (2018) suggests "is that our measure of intelligence be tied to our measure of social interaction." (para. 7). Perez calls his new definition for generalization "Conversational Cognition" and as he explains:

"An ecological approach to cognition is based on an autonomous system that learns by interacting with its environment. Generalization in this regard is related to how effectively automation is able to **anticipate** contextual changes in an environment and perform the required context switches to ensure high predictability. The focus is not just in recognizing chunks of ideas, but also being able to recognize the relationship of these chunks with other chunks. There is an added emphasis on recognizing and predicting the opportunities of change in context." (para. 11).

The most sophisticated form of generalization that exists demands the need to perform conversations. Moreover, Perez (Ibid) clarifies that this conversation "is not confined only to an inanimate environment with deterministic behavior. […] we need to explore conversation for computation, autonomy and social dimensions. […] The social environment will likely be the most sophisticated system in that it may demand understanding the nuisances of human behavior. This may include complex behavior such as deception, sarcasm and negotiation." (Para. 13-14).

Another critical aspect of social survival is the requirement for cooperative behavior. But as Perez (Ibid) argues, effective prediction of an environment is an insufficient skill to achieve cooperative behavior. The development of language is a fundamental skill, and conversations are the highest reflection of intelligence. "They require the cognitive capabilities of memory, conflict detection and resolution, analogy, generalization and innovation." (para. 15). But at the same time it is important to keep in mind that languages are not static — they evolve over time with new concepts.

Moreover, Perez (Ibid) clarifies that "effective conversation requires not only understanding an external party but also the communication of an automaton's inner model. In other words, this conversation requires the appropriate contextualized communication that anticipates the cognitive capabilities of other conversing entities. Good conversation requires good listening skills as well as the ability to assess the current knowledge of a participant and performing the necessary adjustment to convey information that a participant can relate to." (para. 16). For Perez, the ability to effectively perform a conversation with the environment is the essence of AGI. Interestingly enough, what most AGI research avoids is the reality that an environment is intrinsically social- i.e. that there exist other intelligences.

As we have argued above, we believe that the next step to take to make human and machine intelligence come closer together, is to focus on the social aspect of human intelligence and on the ways to integrate social behavior in machines.

## 6. Embodiment

One of the biggest challenges for AI is the challenge of embodied cognition. If AI could surpass this hurdle it will be very close to true human intelligence. Embodied cognition was first presented as the main reason why AI is impossible. We propose to view embodied cognition as a step towards better AI and not as an imposition. Let us make a small detour to the history and philosophy of computation.

Dreyfus (1979) claimed that true AI is impossible since it implicitly assumes that human intelligence is symbolic in its essence. Some AI researchers are attempting to build a context-free machine that manipulates symbols, assuming the human mind works similarly. Dreyfus claimed that the symbolic conjecture is a fault, basing his arguments primarily on philosophical grounds. AI assumes that we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



have a type of `knowledge representation' in our brain, a representation of the world, this idea is based on Descartes' theory, and has a long tradition. Moreover, Descartes claimed that there is a duality, a separation between our body and our mind, therefore the mind cannot be embodied. So far, claimed Dreyfus (Ibid), all AI research is based on these assumptions that we have a model of the world in our mind and that the mind is separated from the body.

Could these assumptions be wrong? Could it be that some part of our intelligence in embedded in our body? We interact with the world with our body, we perceive with our body, we sense with our body. Could it be that symbolic intelligence is not enough?

For Dreyfus (Ibid), embodiment is enrooted in the deep philosophical grounds of existentialism. Existentialism discusses the notions of involvement and detachment. Most of the time, humans are involved in the world, they interact, they solve practical problems, they are involved in everyday coping, finding their way about in the world. However, when things become difficult, the individual retracts into detachment. For most of the things you do, there is no need for any type of awareness; while climbing stairs you do not think about the next step. If you do you will probably fall. Only when the stairs are too steep, you might consider your next step, and then you retract to a state of detachment. For Heidegger (1962) there is the World where we live, where everything has `meaning', and there is the Universe where detachment and science lives. Science is the outcome of our involvement in the world, and not the other way around. Science cannot explain the `meaning' of things. Existentialism is therefore the opposite of Descartes' dualism.

Part of our intelligence is therefore somewhere in our bodily interactions with the world. In addition to our senses of sight, smell, hearing etc., we have `senses' of time, space, surroundings, etc. The discovery of neural place cells (O'Keefe and Dostrovsky 1971; O'Keefe 1979; Moser, Kropff and Moser 2008) emphasizes the embodiment of our sense of space. A good example to illustrate embodiment is the proven connection between movement and intelligence in baby development. Free movement, such as rolling, crawling, sitting, walking, jumping, etc., is associated with the development of the frontal cortex, the area where higher-order thinking occurs (Hannaford [1995] 2005).

We can coin the above set of `senses' as `environmental intelligence'. The question is how much of our intelligence is grounded in our body, and how much is `context free' in our mind? If we had no body, could we think the same way we think? Could we think at all? Would we have anything to think about? What is the connection between our `environmental intelligence' and our `context free symbolic manipulation intelligence'?

Dreyfus (1979) thought that neural network computations are indeed a step in the right direction. It is an attempt to formalize our perceptions in terms of virtual neurons which have some parallel in our brain and body.

There were computer scientists that took Dreyfus' stand seriously. Brooks (1990) came up with the idea that there is no need for knowledge representation at all: "The key observation is that the world is its own best model" (p. 6).

Brooks' robots were able to wander around, to avoid obstacles and to perform several basic tasks. A more modern version would be an intelligent swarm, where a set of simple agents interact and can bring about some emergent property (Kennedy 2006).

Brooks and Dennett cooperated in a project involving a humanoid named COG (Dennett 1995; Brooks 1997), where they tried to implement the above ideas by letting the COG computer (with a torso, camera eyes, one arm, and three fingers) interact with its environment, trying to learn new things as if it is a newborn. Brooks used a `subsumption architecture' where several simple modules were competing for dominance. A few years later, the science of Embodied Cognition was born (Lakoff and Johnson 1999; 2003) and reached similar conclusions from a different point of view, the point of view of cognitive psychology.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



The science of embodied cognition has several basic assumptions. First, humans have a set of atomic and primitive cognitive abilities that are embodied. These abilities build our perceptions of the world. Lakoff and Johnson (1999) used the term `image schema'. Such a schema is a small process, which therefore could also be dynamic. Furthermore, any higher feature, any concept that we form, is the result of aggregations of the above primitives. Finally, we think by using Metaphors and Frames and these are formed using associations of schemas.

As for Frames, many words have a large and natural context and cannot be understood without their context, for example prisoner, nurse, doctors, etc. These are the Frames. Frames were suggested in social science by Goffman (1974), and they were also referred to in the context of AI by Minsky (1974). Minsky was also interested in issues such as: the symbolic aspect of vision, the relation of his theory of Frames to Piaget's theory of development, language understanding, scenarios etc. Dreyfus (1979), on the other hand, stressed the fact that real frames are infinite in nature and could not be truly described in AI. This was coined the 'Frame Problem'.

Metaphors are formed using Hebbian learning (Hebb 1949). In early childhood we learn to associate between several concepts, such as `warm' and `up', or `cold' and `down'. The more these associations of schemas are presented to us in childhood the stronger we relate the schemas. Later we use more elaborated metaphors to reason. We solve real situations by using homeomorphisms into a world of metaphors, solving the imaginary situation first. This is `thinking by metaphors' (Lakoff and Johnson 2003).

To prove all of the above, embodied cognition scientists search for clues in language where they look for invariants. The existence of such invariants can imply that something deep, common to all languages, underlies. In many examples a word has several meanings, one is environmental and embodied, the other much more abstract. For example `to grasp' is first of all `to catch', however it also has the meaning of `to understand'. We `see' things in the sense of understanding, we talk about a `warm` or `cold' person, etc. Old proverbs are a good source for such examples.

Lakoff and Johnson's (Ibid) argument is that the more abstract meaning is the secondary one, it is the derived meaning. We derive such meanings from the simpler ones, the embodied and primitive meanings. It is a bottom up process.

Artificial Social Intelligence is also concerned with the environment of the intelligent agent, in particular its social environment. However, there is a difference between the theory of ASI and Embodied Cognition. ASI is rooted in social science, and the notion of a `social brain'. Embodied cognition is rooted in the intersection of cognitive sciences and linguistics, it is also based on philosophical grounds. Embodied cognition focuses on universal principles of understandings, or in other words, ***the*** cognitive architecture.

Can we follow the way from the embodied primitives upwards to the abstract concepts? Can we use AI to boost such research? This was attempted by several researchers; Regier (1995) used the theory of deep neural networks (and recurrent neural networks) to identify the primitives of cognition used in relation to our sense of space. In deep neural networks we take very simple and primitive features and build upon the next layer of more complex features. This is a multi-step process. This bottom up process happens until a global pattern can be recognized and it resembles the process that was defined by Lakoff and Johnson (1999) for schemas. Regier (1995) was struggling with old methods of back propagation to tune his network. Today we can advance Regier's idea by using new techniques in Deep Neural Networks.

Another way in which we can implement embodiment cognition is by formalizing the idea of metaphors. To be able to use metaphors we need to enable the computer the capability to simulate a situation in which the machine itself resides. This was already done in the context of value alignment. Blum and Liu's (2014) defined a `consequence machine' that could simulate a situation, but could also observe itself in that simulation. The machine then had to decide on a `moral' dilemma.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



Embodied cognition is a new hurdle to overcome, it is the missing bridge between robotics and AI. It should not be thought of as an imposition on AI but as a new challenge.

## 7. The way to overcome our fears: Value alignment

In an article called "How Do We Align Artificial Intelligence with Human Values?" (2018), Conn explains that "highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation". (para. 5).

One of the main challenges in aligning AI with values is to understand (and to agree upon) what exactly these values are. There are many factors that must be taken into account which depend mainly on context - cultural, social, socioeconomic and more. It is also important to remember that humanity often does not agree on common values, and even when it does, social values tend to change over time.

Eliezer Yudkowsky offered the first attempt at explaining AI alignment in his seminal work on the topic, "Creating Friendly AI" (2001), and he followed this up with a more nuanced description of alignment in "Coherent Extrapolated Volition" (2004). Nearly a decade later Stuart Russell began talking about the value alignment problem, giving AI alignment its name and motivating a broader interest in AI safety. Since then numerous researchers and organizations have worked on AI alignment to give a better understanding of the problem.

As Tegmark (2017) explains: "aligning machine goals with our own involves three unsolved problems: making machines learn them, adopt them and retain them. AI can be created to have virtually any goal, but almost any sufficiently ambitious goal can lead to subgoals of self-preservation, resource acquisition and curiosity to understand the world better- the former two may potentially lead a superintelligence AI to cause problems for humans, and the latter may prevent it from retaining the goals we give it." (p. 389).

How to implement Value Alignment? Wilson and Daugherty (2018) describe three critical roles that we, humans, need to perform:

***Training***: Developing 'personalities' for AI requires considerable training by diverse experts. For example, in order to create Cortana's personality, Microsoft's AI assistant, several human trainers such as a play writer, novelist and poet, spent hours in helping developers create a personality that is confident, helpful and not too 'bossy'. Apple's Siri is another example. Much time and effort was spent to create Siri with a hint of sassiness, as expected from an Apple product.

Creating AI with more complex and subtle human traits is sought after by new startups for AI assistants. Koko, a startup born out of the MIT Media Lab, has created an AI assistant that can display sympathy. For example, if a person is having a bad day, it will not just say 'I'm sorry to hear that', but will ask for more information and perhaps provide advice like 'tension could be harnessed into action and change'. (Hardesty 2015).

***Explaining***: As AI develops, it reaches results through processes that are unclear to users at times, a sort of internal 'black box'. Therefore, they require expert, industry specific 'explainers' for us to understand how AI reached a certain conclusion. This is especially critical in evidence-based industries such as medicine and law. A medical practitioner must receive an explanation of why an AI assistant gave a certain recommendation, what is the internal 'reasoning' that led to a decision. In a similar way, law enforcement investigating an autonomous vehicle accident, need experts to explain the AI's reasoning behind decisions that led to an accident. (Wilson and Daugherty 2018).

***Sustaining***: AI also requires sustainers. Sustainers oversee and work on making sure AI is functioning as intended, in a safe and responsible manner. For example, a sustainer would make sure an autonomous car recognizes all human diversity and takes action not to risk or harm any human being. Other sustainers may be in charge of making sure AI is functioning within the desired ethical norms. For example, when analyzing big data to enhance user monetization, a sustainer would oversee that the process is using general statistical data and not specific and personal data (which may generate negative sentiment by users) to deduce its conclusions and actions. (Ibid).

The unique roles of human values presented here, have been linked to the workplace environment, but they are undoubtedly relevant to all spheres of life. As we claimed earlier, we are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



at the beginning of a new developmental stage in AI, the one of artificial social intelligence. Within this realm, new questions concerning human values may arise.

## 8. Summary

In his book *Technopoly* (1992) Postman writes that "[…] once a technology is admitted, it plays out its hand; it does what it is designed to do. Our task is to understand what that design is- that is to say, when we admit a new technology to the culture, we must do so with our eyes wide open." (p. 7). For "technological change is neither additive nor subtractive. It is ecological. I mean 'ecological' in the same sense that the word is used by environmental scientists. One significant change generates total change." (p. 18).

AI is woven into our lives, changing our environment. The short history of AI has shown that developments in AI go hand in hand with our understanding of ourselves. Although there is still a long way to go before we can talk about a singularity point, it is almost clear that the next few steps in AI technology (e.g. ASI and AGI) will bring about a much more powerful machines, flexible enough to resemble human behavior.

A major part of human intelligence is social, we interact with others, we compete, we cooperate, we imitate, etc. A symbolic `context free' intelligence cannot be complete without this social constituent. ASI is therefore necessary for building `true' human intelligence.

One of the most complex challenges that AI faces is the issue of embodiment. At first, one should recognize that part of our intelligence is indeed embodied. We are physically situated in the world; we move in space, perceive, feel and communicate through our bodies. We argue that embodied cognition should be dealt with as a new challenge to AI and not as an imposition.

## References

1. Adolphs, Ralph. 2003. Cognitive neuroscience: Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience* 4, no. 3.
2. Aida- Artificial intelligence driven analytics Home Page. Available Online: http://aidatech.io/
3. Anderson, Bo. 1989. On artificial intelligence and theory construction in sociology. *Journal of Mathematical Sociology* 14, no. 2-3. pp. 209-216.
4. Andrychowicz, Marcin, Baker, Bowen, Chociej, Maciek, Jozefowicz, Rafal, McGrew, Bob, Pachocki, Jakub, Petron, Arthur, Plappert, Mattias, Powell, Glenn, Ray, Alex, Schneider, Jonas, Sidor, Szymon, Tobin, Josh, Welinder, Peter, Weng, Lilian, Zaremba, Wojciech. 2018. Learning Dexterous In-Hand Manipulation. *OpenAI*. Available Online: https://arxiv.org/pdf/1808.00177.pdf
5. AutoDesk's Dreamcatcher Home Page. Available Online: https://autodeskresearch.com/projects/dreamcatcher
6. Bainbridge, William Sims, Edward E. Brent, Kathleen M. Carley, David R. Heise, Michael W. Macy, Barry Markovsky, and John Skvoretz. 1994. Artificial social intelligence. *Annual Review of Sociology* 20, no. 1, pp. 407-436.
7. Bedau, Mark, and Paul Humphreys, eds. 2008. *Emergence: Contemporary readings in philosophy and science*. Cambridge, MA: MIT press.
8. Boden, Margaret A. 1978. *Artificial Intelligence and the Natural Man*. London: The MIT Press. First published 1977.
9. Bohus, Dan, and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 2-9. ACM.
10. Bohus, Dan, Sean Andrist, and Mihai Jalobeanu. 2017. Rapid development of multimodal interactive systems: a demonstration of platform for situated intelligence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 493-494. ACM.
11. Brooks, Rodney A. 1990. Elephants don't play chess. *Robotics and autonomous systems* 6, no. 1-2, pp. 3-15.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002



12. Brooks, Rodney A. 1997. The cog project. *Journal of the Robotics Society of Japan* 15, no. 7, pp. 968-970.
13. Brothers, Leslie. 2002. The social brain: a project for integrating primate behavior and neurophysiology in a new domain. *Foundations in social neuroscience,* pp. 367-385.
14. Calvin, William H. [1996] 2001. *How brains think*. New York: Basic Books.
15. Cann, Ariel. February 3, 2018. How Do We Align Artificial Intelligence with Human Values? *Future of life institute*. Available Online: https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/
16. Collins, Allan, and Edward E. Smith, eds. 2013. *Readings in cognitive science: A perspective from psychology and artificial intelligence*. Elsevier.
17. Collins, Harry. M. 1990. *Artificial experts: Social knowledge and intelligent machines (inside technology)*. Cambridge, MA: The MIT Press.
18. Danescu-Niculescu-Mizil, Cristian, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pp. 745-754. ACM.
19. de Kleer, Johan. 1990. The origin and resolution of ambiguities in causal arguments. In *Readings in qualitative reasoning about physical systems*, pp. 624-630.
20. Dennett, Daniel C. 1995. Cog: steps towards consciousness in robots. *Conscious experience*, pp. 471-487.
21. Dodds, Agnes E., Jeanette A. Lawrence, and Jaan Valsiner. 1997. The Personal and the Social: Mead's Theory of the Generalized Other. *Theory & Psychology* 7, no. 4, pp. 483-503.
22. Dreyfus, Hubert L. 1979. *What computers can't do: The limits of artificial intelligence*. Vol. 1972. New York: Harper & Row.
23. Dunbar, Robin IM. 1998. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* 6, no. 5, pp. 178-190.
24. Faghihi, Usef, and Stan Franklin. 2012. The LIDA model as a foundational architecture for AGI." In *Theoretical Foundations of Artificial General Intelligence*, pp. 103-121. Paris: Atlantis Press.
25. Feigenbaum, Edward .A. 1963. The simulation of verbal learning behavior. In Edward .A. Feigenbaum & Julian Feldman (Eds.), *Computers and Thought*. pp. 297-309. New York: McGraw-Hill.
26. Frith, Chris D. 2007. The social brain. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362, no. 1480. pp. 671-678.
27. Goertzel, Ben. 1993a. *The Structure of Intelligence*. New York: Springer-Verlag.
28. Goertzel, Ben. 1993b. *The evolving mind*. New York: Gordon and Breach.
29. Goertzel, Ben. 1994. *Chaotic Logic.* New York: Plenum Press.
30. Goertzel, Ben. 1997. *From Complexity to Creativity*. New York: Plenum Press.
31. Goertzel, Ben. 2001. *Creating Internet Intelligence.* New York: Plenum Press.
32. Goertzel, Ben. 2007. *Artificial general intelligence*. Edited by Cassio Pennachin. Vol. 2. New York: Springer.
33. Goertzel, Ben, Cassio Pennachin, Andre Senna, Thiago Maia, and Guilherme Lamacie. 2004. Novamente: an integrative architecture for Artificial General Intelligence. In *Proceedings of AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, Washington DC*.
34. Goertzel, Ben, Cassio Pennachin, Nil Geissweiller, Moshe Looks, Andre Senna, Welter Silva, Ari Heljakka, and Carlos Lopes. 2008. An integrative methodology for teaching embodied non-linguistic agents, applied to virtual animals in second life. *Frontiers in Artificial Intelligence and Applications* 171.
35. Goffman, Erving. 1974. *Frame analysis: An essay on the organization of experience*. Harvard University Press
36. Gonzalez, Carlos. Jun 18, 2018. Seven Common Applications for Cobots. *MachineDesign*. Available Online: http://www.machinedesign.com/motion-control/7-common-applications-cobots

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



37. Hannaford, Carla. [1995] 2005. *Smart moves: Why learning is not all in your head*. Salt Lake City, Utah: Great River Books.
38. Hardesty, Larry. March 30, 2015. Crowdsourced tool for depression. *MIT News*. Available Online: http://news.mit.edu/2015/crowdsourced-depression-tool-0330
39. Hebb, Donald O. 1949. The organization of behavior: A Neuropsychological Theory. New York: JOHN WILEY & SONS, Inc.
40. Heidegger, Martin. 1962. Being and time. 1927. Trans. John Macquarrie and Edward Robinson. New York: Harper.
41. Hofstadter, Douglas R. 1995. *Fluid concepts and creative analogies*. New York: Basic Books.
42. Kang, Minsoo. 2011. *Sublime dreams of living machines*. Harvard University Press.
43. Kennedy, James. 2006. Swarm intelligence. In *Handbook of nature-inspired and innovative computing*, pp. 187-219. Springer, Boston, MA.
44. Laird, John E., Allen Newell, and Paul S. Rosenbloom. 1987. Soar: An architecture for general intelligence. *Artificial intelligence*33, no. 1. pp. 1-64.
45. Lakoff, George, and Mark Johnson. 1999. *Philosophy in the Flesh*. Vol. 4. New York: Basic books.
46. Lakoff, George, and Mark Johnson. 2003. *Metaphors we live by*. London: The University of Chicago Press.
47. Landing.ai Home Page. Available Online: https://www.landing.ai/
48. Lenat, Douglas B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38, no. 11. pp. 33-38.
49. Lévi-Strauss, Claude, and John Weightman. 1969. *The raw and the cooked: Introduction to a science of mythology*. Vol. 1. New York: Harper & Row.
50. Lindsay, Robert K., Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. 1980. Applications of artificial intelligence for organic chemistry: the DENDRAL project. *New York*.
51. Manyika, James, Chui, Michael, Miremadi, Mehai, Bughin, Jacques, George, Katy, Willmott, Paul, and Dewhurst, Martin. January 2017. *The future that works: Automation, employment and productivity*. Mckinsey and Company Global Institute. Available Online: https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works_Full-report.ashx
52. McCarthy, John. 1978. History of LISP. In *History of programming languages I*, pp. 173-185. ACM.
53. McLuhan, Marshal. 1964. *Understanding Media: Extensions of Man*. New York: McGraw Hill.
54. Mead, George Herbert. 1934. *Mind, self and society*. Vol. 111. Chicago: University of Chicago Press.
55. Microsoft Research India workshop on ASI, 2007. Available Online:
56. https://www.microsoft.com/en-us/research/event/microsoft-research-india-summer-school-artificial-social-intelligence/
57. Minsky, Marvin. 1975. A framework for representing knowledge. In Patrick H. Winston (Ed.), *The Psychology of Computer Vision*. pp. 211-277. New York: McGraw-Hill.
58. Minsky, Marvin. 2006. The emotion machine. *New York: Pantheon* 56.
59. Moser, Edvard I., Emilio Kropff, and May-Britt Moser. 2008. Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience* 31, pp. 69-89.
60. Newell, Allen, and Herbert Alexander Simon. 1961. *GPS, a program that simulates human thought*. No. P-2257. RAND CORP SANTA MONICA CALIF.
61. Newell, Allen, and Herbert Alexander Simon. 1963. GPS a program that simulates human thoughts. In EA Feigenbaum and J. Feldman eds., *Computer and Thoughts*. McGraw-Hill, New-York.
62. Newell, Allen, and Herbert Alexander Simon. 1972. *Human problem solving*. Vol. 104, no. 9. Englewood Cliffs, NJ: Prentice-Hall.
63. Norman, Donald A., Rumelhart, David E. & the LNR Research Group. 1975. *Explorations in Cognition*. San Francisco: W.H. Freeman.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at Information 2019, 10, 2; doi:10.3390/info10010002




64. O'Keefe John and Jonathan Dostrovsky. 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res*. 34, pp. 171–75.
65. O'Keefe, John. 1979. A review of the hippocampal place cells. *Progress in neurobiology* 13, no. 4, pp. 419-439.
66. Pennachin, Cassio, and Ben Goertzel. 2007. Contemporary approaches to artificial general intelligence. In *Artificial general intelligence*, pp. 1-30. Springer, Berlin, Heidelberg.
67. Perez, Carlos E. Feb 9, 2018. Conversational Cognition: A New Measure for Artificial General Intelligence, *Medium*. Available Online: https://medium.com/intuitionmachine/conversational-cognition-a-new-approach-to-agi-95486ffe581f
68. Postman, Neil. 1992. *Technopoly: the surrender of culture to technology*. New York: Vintage Books.
69. Quillian, M. Ross. 1968. Semantic memory. In Marvin Minsky (Ed.), *Semantic Information Processing*. pp. 216-270. Cambridge, MA: MIT Press.
70. Ramamurthy, Uma, Bernard J. Baars, D'Mello SK, and Stan Franklin. 2006. LIDA: A working model of cognition. In Danilo Fum, Fabio Del Missier, & Andrea Stocco (Eds.), Proceedings of the 7th international conference on cognitive modeling. pp. 244–249. Trieste: Edizioni Goliardiche.
71. Regier, Terry. 1995. A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)* 6, no. 1, pp. 63-88.
72. Schank, Roger C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology* 3, no. 4. pp. 552-631.
73. Schank, Roger C. & Abelson, Robert P. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.
74. Schmidhuber, Jürgen. (2007). Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pp. 199-226. Springer, Berlin, Heidelberg.
75. Stevens, Greg. Oct 17, 2013. AI is over: this is artificial empathy. *Kernel Magazine*. Available Online: https://kernelmag.dailydot.com/features/report/5910/ai-is-so-over-this-is-artificial-empathy/#
76. Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg et al. 2016. Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*.
77. Strate, Lance. 2017. *Media Ecology: An Approach to Understanding the Human Condition*. Peter Lang Publishing, Incorporated.
78. Tegmark, Max. 2017. *Life 3.0: Being human in the age of artificial intelligence*. Knopf.
79. Universal Robots. 2018. White paper: An Introduction to Common Collaborative Robot Applications. Available Online: https://cdn2.hubspot.net/hubfs/2631781/HQ%20Content%20and%20Enablers/HQ%20Enablers/White%20papers/Common%20Cobot%20Applications.pdf
80. Voss, Peter. 2007. Essentials of general intelligence: The direct path to artificial general intelligence. In *Artificial general intelligence*, pp. 131-157. Berlin, Heidelberg: Springer.
81. Voss, Peter. Oct 4, 2017. From narrow to general AI. *Medium*. Available Online: https://medium.com/intuitionmachine/from-narrow-to-general-ai-e21b568155b9
82. Wilson, H. James, and Daugherty, Paul R. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, July-August Issue. Available Online: https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces
83. Winerman, Lea. April 2018. Making a thinking machine. *American Psychological Association* Vol. 49, No 4. Available Online: https://www.apa.org/monitor/2018/04/cover-thinking-machine.aspx
84. Winfield, Alan FT, Christian Blum, and Wenguo Liu. 2014. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards autonomous robotic systems*, pp. 85-96. Springer, Cham.
85. Ye, Peijun, Tao Wang, and Fei-Yue Wang. 2018. A Survey of Cognitive Architectures in the Past 20 Years. *IEEE transactions on cybernetics* 99. pp. 1-11.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201810.0736.v2

Peer-reviewed version available at *Information* **2019**, *10*, 2; doi:10.3390/info10010002



86. Yudkowsky, Eliezer. 2001. Creating friendly AI 1.0: The analysis and design of benevolent goal architectures. *The Singularity Institute for Artificial Intelligence, San Francisco, USA*.
87. Yudkowsky, Eliezer. 2004. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*. *San Francisco, USA*.
88. Yudkowsky, Eliezer. 2007. Levels of organization in general intelligence. In *Artificial general intelligence*, pp. 389-501. Berlin, Heidelberg: Springer.