 doi:10.20944/preprints201810.0736.v1

From Homo sapiens to Robo sapiens: The evolution of Intelligence

Anat Ringel Raveh and Boaz Tamir

Bar-Ilan University, The STS program, Israel

Abstract

In this paper we present an argument in favor of the possibility of an artificial intelligence above human intelligence. AI technology has shown a stepwise increase in its capacity and complexity. 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, understanding human cognition. Indeed, AI was always about the question of understanding human nature. AI percolates into our lives, step by step, changing our environment. We believe the next few steps in AI technology, and in our understanding of human behavior, will bring about a 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). On the ground of ASI and AGI we present an evolutionary argument which uses research in artificial life simulations, showing an increase in complexity due to an emergent property coming out of lower complexity constituents. The whole process is driven by an evolutionary force. What could such an evolutionary force be? We suggest a social communicative driving force. We end the discussion demonstrating a way to overcome our fears of singularity, harnessing value alignment.

Key words: Artificial Intelligence (AI), artificial general intelligence (AGI), artificial social intelligence (ASI), singularity, complexity, emergence phenomena, evolution, social sciences.

 doi:10.20944/preprints201810.0736.v1

Table of content

1. Introduction
   1.1 The Human-Machine Eco-system
   1.2 From intelligence to super-intelligence
2. State of the Art: stepwise incremental AI
3. AI goes hand in hand with our understanding of ourselves
4. ASI- a new challenge
5. AGI- an overview, is it enough?
6. Super-intelligence as an emergent phenomena, an evolutionary argument
7. The way to overcome our fears: Value alignment
8. Conclusion
   Bibliography

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

1. Introduction

1.1 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. We're fascinated by machines that can imitate humans 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 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 leads 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 increases. While in the past automata have entertained us mainly because they mimicked human behavior in an inaccurate way that revealed the fact that it was a trick, artificially intelligent machines today can successfully mimic more and more of 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 ever has been, mostly creates fear (Kang 2011).

 doi:10.20944/preprints201810.0736.v1

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. Because as McLuhan (1964) argued, 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.

1.2 From intelligence to super-intelligence

In this paper we present an argument in favor of the possibility of an artificial intelligence above human intelligence. Our argument is based on evolutionary ground. So far AI technology has shown a stepwise increase in its capacity, in its complexity. 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, understanding human cognition (section 2). Indeed AI was always about the question of understanding human nature. AI percolate into our lives, step by step, changing our environment (section 3).

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, 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 a much more powerful machines, flexible enough to resemble

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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, much more flexible, closer to human intelligence, it suggests a model of conscious, a new approach to learning to learn, recursive machines, etc.

On the ground of ASI and AGI we suggest an evolutionary argument (section 6). We can look at evolution as a physical process that increases complexity and brings about emergent properties. Evolution produced humans out of a pre-intelligent specie. Our argument harnesses past research in artificial life simulations, showing an increase in complexity due to an emergent property rising out of lower complexity constituents. It is well known in evolutionary theory that the flexibility of the lower complexity constituents could push forward the jump into a higher level of complexity; the flexibility being the constituent's property of having one function at the lower level and possibly some other function at the higher level. Moreover, it is also known that such a process is not necessarily monotonic, and it could indeed be that several properties in the lower level are waved in favor of the higher more complex property. The whole process is driven by an evolutionary force. What could be such an evolutionary force? We suggest a social communicative driving force.

We end the discussion demonstrating a way to overcome our fears of singularity, which is by value alignment (section 7).

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 if to use 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 describe. But first we want to describe the ongoing process of stepwise incremental AI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

In January 2017 McKinsey (Manyika et al. 2017) published a comprehensive report that maps the progress of artificial intelligence in a variety of areas. Those five areas are further broken down into other tasks and capabilities that humans can do (pp. 34-35):

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. In this area the performance level is median compared to humans.

Take Machine vision as an example. Devising cameras with capabilities that surpass the human eye has been the easy part. What AI adds is the increasingly useful ability to make sense of the images. Projects like Landing.ai (https://www.landing.ai/), formed by Andrew Ng, a globally recognized leader in AI, focuses on manufacturing problems such as precise quality analysis. This startup has developed machine-vision tools to find microscopic defects in products such as circuit boards at resolutions well beyond human vision, using a machine-learning algorithm trained on remarkably small volumes of sample images.

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 achieve an unprecedented level of dexterity on a physical five-fingered 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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.

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. 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.

Natural language processing- This consists of two distinct parts: one is natural language generation, which is the ability to deliver spoken messages, including with nuanced

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

human interaction and gestures. The second is natural language understanding, which is the comprehension of language and nuanced linguistic communication in all its rich complexity. Although there is progress in this area (such as Google duplex), the levels of performance according to the report are at best median (Natural language generation). 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 millions of customers through natural-language conversations, and hence has access to vast stores of data. This way she can answer many frequently asked questions, such as how to open 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).

Physical capabilities- This includes gross motor skills, navigation (these two have reached human level performance), fine motor skills and mobility (those are harder problems 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.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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- have 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 often requires advanced vision systems." (para. 3).

Social and emotional capabilities- This consists of three types of capabilities: 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.

Let us Consider Mattersight as an example. The company provides an extremely sophisticated data analysis system that listens to the way customers respond on the telephone. The software analyzes communication patterns, grammar, word choice, tone, volume, pauses, and other communication metrics. Mathematical algorithms then interpret vocal features, compare them to their databases, and arrive at a personality profile for each customer, who is then matched with a service agent with whom the customer is most compatible. (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. But, although the robot's "intelligence", has progressed, this is still where the most formidable technical challenges are met:

"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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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 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 phenomenon whose varied aspects have attracted the attention of different fields of study, including psychology, neuroscience, biology, economics, engineering, statistics, and linguistics. Naturally, the field of AI has benefited from the progress made by all of these allied fields. 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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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 for scientific hypothesis formation.

"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

 doi:10.20944/preprints201810.0736.v1

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 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 to the ground 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,

 doi:10.20944/preprints201810.0736.v1

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).

A question we should ask: Does the use of the same terms and the same language in AI and cognitive sciences is 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.

To answer these questions, we refer to the viewpoint of Dennett (2017). Let's define the notion of 'meaning'; to put things very simplistic, we will say that an action of a computer agent has a 'meaning' (for the agent) if the action is changing 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 invest 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 it 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

 doi:10.20944/preprints201810.0736.v1

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- which have been inspired by the way that neurons connect in the brain and 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 analyze, deep neural networks have become increasingly powerful, useful and ubiquitous. (Winerman 2018).

The theory of conscious was recently investigated by AI researchers. It was suggested by Dennett (2017) that conscious is an emergent property of many small processes, or agents, each struggle for its homogeneity. Conscious is not a stage with spotlights in which all subconscious processes are the audience. Conscious 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 LIDA (Fagihihi and Franklin 2012) model. These two are examples of a strong reciprocal interaction between AI and cognitive science.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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 a relatively new interdisciplinary science, which was originally introduced years ago by Brent and others (Bainbridge et al. 1994), however only now is gaining power. 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

 doi:10.20944/preprints201810.0736.v1

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 the human individual is a member of a social organism, and his acts must be viewed in the context of social acts that involve other individuals. The social act is therefore viewed as a dynamic and complex system within which the individual is situated, and it is within this situation that individual acts have meaning.

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 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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

and in a way that can account for the possibility of change in both person and society." (p. 495).

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).

On the basis of this perception, a new central aspect becomes the focus of inquiry and understanding of artificial intelligence as well as ourselves as social beings. The main concern of this new interdisciplinary field of science is the formalization of delicate social interactions, using it in AI to implement social awareness (a type of common sense understanding) and behavior into robots. These ASI systems will have to continuously review and evolve their interaction strategies during ongoing interactions which accrue in different situations and contexts.

For that to happened, 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 operates. In order to formulate the principles that were discovered, it is also important to conduct large data-driven studies that aim at validating those principles and deciphering new behavioral traits. Such studies are already happening, due to the large scale availability of socially grounded user data from social media, and due to advances in machine learning and other data-analysis techniques. One such project is "Mark my words!" (Danescu-Niculescu-Mizil, Gamon and Dumais 2011) which demonstrates the psycholinguistic theory of communication accommodation; the fact that participants in conversations tend to converge to one another's communicative behavior. The researches have shown that the hypothesis of linguistic style accommodation can be confirmed in a real life, large scale dataset of Twitter conversations. A probabilistic framework was developed, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

allowed the researchers to measure accommodation and to distinguish effects of style accommodation from those of homophily and topic-accommodation.

Once the 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, self-driving cars, etc. One such system is the *virtual receptionist* developed by Dr. Dan Bohus from Microsoft Research Redmond, 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).

ASI systems has no clear definition of goals, there is no specific task the machine is oriented towards. In a sense the machine's behavior is our goal. Hence goals may not be defined in advance and might evolve dynamically. Therefore, it is extremely difficult to directly measure and evaluate 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?

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).

A marked distinction exists between AGI and specialized "narrow AI" research. The latter is aimed at creating programs carrying out specific tasks like playing chess, diagnosing diseases, driving cars and so forth. Most contemporary AI work falls into this category. But, despite their great importance and contribution, narrow AIs core problem is that they

 doi:10.20944/preprints201810.0736.v1

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). This cognitive ability must operate in real time, in the real world, and with limited knowledge and time.

“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 able to carry out a variety of different tasks in a variety of different contexts, generalizing knowledge from one context to another, and building up a context and task independent pragmatic understanding of itself and the world. 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).

 doi:10.20944/preprints201810.0736.v1

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 (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 doesn't comprise one single invention or design feature, but rather it emerges from the synergetic integration of a number of essential fundamental 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 view on cognition, as Wang and Goertzel (Ibid) explains, "An AGI project often starts with a

blueprint of 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 many heterogeneous parts and subcomponents. Typically, many of these architectures are built to control artificial agents, which run both in virtual worlds and physical robots. (Wang and Wang 2018).

An important kind of CA is the symbolic systems. This type of agents maintains a consistent knowledge base by representing the environment as symbols. 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 is Doug Lenat's CYC project (1995). This was an attempt to create true AI by encoding all human common sense knowledge in first order predicate logic. Alan Newell's SOAR project (1987) was an attempt to build "Unified Theories of Cognition", based on logic-style knowledge representation, mental activity as problem-solving carried out by an assemblage of heuristics, etc. However, the system was not constructed to have a real autonomy or self-understanding.

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 (like GPS) 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

One such integrative approach described by Pennachin and Goertzel (2007), was given the name 'Novamente'. This approach involves taking elements from various approaches and creating a combined, synergistic system. However, as the two explains: "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 confronts the problem of "creating a whole mind" in a direct way that has not been done before. "The fundamental principles underlying the system design derive from a novel complex-systems-based theory of mind called the "psynet model," which was developed by

 doi:10.20944/preprints201810.0736.v1

Goertzel in a series of cross disciplinary research treatises published during 1993-2001 (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-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) which illustrates how cognitive science principles can be applied towards the hard problems of AI (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 conscious as a working space for many agents. The agents are competing one another and those that dominate the workspace are identified as those that constitute our attention. The process is dynamic, information flows in from the environment, and action is decided by a set of heuristics, which are themselves dynamic.

 doi:10.20944/preprints201810.0736.v1

The LIDA architecture is partly symbolic and partly connectionist, and the mechanisms used in implementing the several modules have been inspired by a number of different 'new AI' techniques. The architecture is partly 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).

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. And as the authors 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." (Ibid, p. 6).

A third project worth mentioning is Schmidhuber's Gödel Machines (2006). Schmidhuber describe this 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).

Some of the researchers in the field believe that an almost sure way to create AGI would be to emulate the human brain, down to the atomic level, in a digital simulation. At present, however, it is no more than a futuristic speculation; for we don't understand enough about the brain to make a detailed simulation of a functioning brain. Proceeding one step up the ladder of abstraction, another way to create AGI is to emulate the human mind, as studied by cognitive psychologists. 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 on 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

 doi:10.20944/preprints201810.0736.v1

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).

In other words, it is not enough to have models that are able to model the world in a single context. 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 remember that languages are not static — they evolve over time with new concepts.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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 are other intelligences that exist.

As we have argued above (see also section 6 here) , we believe that the next step needed to make human intelligence and machine intelligence come closer together, is to focus on the social aspect of human intelligence and on the ways to implement social behavior in machines.

6. Superintelligence as an emergent phenomenon, an evolutionary argument

Having achieved all the above, it could indeed be that at some point in the future, a new emergent artificial intelligence will pop up. This is an argument based on evolutionary theory. If the elements of the system are complex enough then it could be that some emergent new system will pop up. In Ofria and Wilke's paper (2004) a set of artificial organisms were discussed. Each had its own genome of instructions. All organisms where going through an evolutionary process. Each such genome was mutated by deletions, insertions of new actions etc. The organisms had to solve several tasks. Organisms that solved a task were rewarded by increasing their rate of reproduction. The set of tasks had different level of complexity, solutions of harder problems were better rewarded exponentially. The authors showed that the organisms' complexity grew while trying to solve harder problems. Several interesting phenomena were observed:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

a) The phylogenetic level at which the hardest problem was solved appeared above and only after the capacity to solve simpler problems was acquired.
b) No clear path was established between the capacity to solve simple problems and the capacity to solve harder problems. The higher level of complexity was emergent above the lower level. For the organism to solve a problem of higher complexity it had to solve/compute a set of simple problems/functions. Different organisms used a different set of basic functions to reach the same complex higher level.
c) In the way to achieve a higher level of complexity sometimes one of the lower level functions was waved.
d) The length of the genomes was increasing. The ancestor organisms that could only reproduce had the shortest genome, while the final trait of organisms that could solve the hardest problem had the longest genome.

When implementing all the above, consider the following gedankenexperiment. Assume a set of intelligent agents, each with its high level of intelligence. We could let the whole system go through an evolutionary process. The agents could interact. Suppose the agents are rewarded for a task which is known to be hard for each one of them. We could test the possibility of evolving an emergent protocol that could solve the task by distributed computation. This emergent intelligence would be one level higher than the level of intelligence of each of the agents. If each of the agents' intelligence is close to human intelligence, then this emergent new capability could be above human capacity. Note that we start with a high level of intelligence from which we can jump a step above human intelligence. There is no way to jump a few steps altogether, Ofria and Wilke (2004) have shown that trying to reward the highest level of complexity without rewarding lower levels did not produce traits with the ability to solve the hardest problem.

The system we suggest here is going through group evolution. We reward all agents if some group task is solved. Therefore, the driving force is meaningful for the whole group.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

We also let the agent communicate to produce a protocol that best solves the task. This is motivated by the theory of the 'social brain`.

It was suggested by Dunbar (1998) that the size of the human brain is a function of the social interactions a human needs to make in order to survive. From this we can assume that the size of the brain- a way to measure intelligence- is a function of the size of its social group. If we follow this argument it is reasonable that we can use the method of Ofria and Wilke (2004) while letting the system solve complex **social** problems, distributed network problems, etc. This could be the way to pull up the system into a new stronger form of intelligence.

In a paper named "The social and cultural roots of whale and dolphin brain" (2017), Fox, Muthukrishna and Shultz evaluate the extent to which cetacean brains are social. As they explain: "Encephalization, or brain expansion, underpins humans' sophisticated social cognition, including language, joint attention, shared goals, teaching, consensus decision-making and empathy. These abilities promote and stabilize cooperative social interactions, and have allowed us to create a 'cognitive' or 'cultural' niche and colonize almost every terrestrial ecosystem." (p. 1).

The researchers' conclusion was that the "cetacean social and brain evolution represent a rare parallel to those in humans and other primates. We suggest that brain evolution in these orders has been driven largely by the challenges of managing and coordinating an information-rich social world. Although these challenges may increase with group size, it is not group size itself that imposes the challenges. In both primates and marine mammals, structured social organization is associated with higher levels of cooperation and a greater breadth of social behaviors. Thus, we propose reframing the evolutionary pressures that have led to encephalization and behavioral sophistication to focus on the challenges of coordination, cooperation, and 'cultural' or behavioral richness." (p. 4).

Note that above, in our gedankenexperiment, there is no way to explain the emergent new intelligence out of the simpler ones. This is the essence of the notion of emergent

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

property. By explaining we usually mean a recursive set of arguments, and if there was an explanation for such a property it could not have been 'emergent`. Therefore, our new intelligence is an emergent higher level of complexity above human intelligence (Bedau and Humphreys 2008), a property we cannot explain its appearance.

We conclude this section with a remark a la Dennett (2017). A **non-intelligent** mechanism of evolution has produced us, human, with our intelligence, there is no reason why such a mechanism could not produce other and more powerful intelligences. Moreover, most of our intelligent capacities have no evolutionary meaning, the invention of high and pure mathematics has no evolutionary benefit. In other words, pure mechanistic evolutionary process can indeed produce a high order of intelligence way above its immediate use and meaning.

The rather swift jump evolution has made from non-intelligent creatures to humans could be explained by the flexibility of the intermediate constituents. These intermediate constituents can function in several different ways according to the level of complexity, i.e. the phylogenetic level in which they act (see also Calvin [1996] 2001). Calvin gives as an example the 'swim bladder' in fish, turning eventually into a lung (Ibid Cha.6). The evolution process could therefore be non-monotonic as indeed observed in Ofria and Wilke (2004). The flexibility of the agent's role in its environment is best dealt in the context of the theory of Complex Adaptive Systems (for a discussion of complex adaptive system in AI see Yang 2008).

8. The way to overcome our fears: Value alignment

In an article called "How Do We Align Artificial Intelligence with Human Values?" (2018), Cann provides a definition for value alignment: "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). The challenge that arises in light of this definition 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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

socioeconomic and more. It is also important to remember that humanity often does not agree on common values, and even if 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 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 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 won't

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2018 doi:10.20944/preprints201810.0736.v1

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 types of humans and takes action not to risk or harm them. 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 deduct its conclusions and actions. (Ibid).

The unique roles of humans 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 at the beginning of a new developmental stage in AI, the one of artificial social science. Within this realm, new questions arise and new opportunities are revealed - not only for artificial intelligence but also for social sciences.

Conclusion

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

 doi:10.20944/preprints201810.0736.v1

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 and 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.

To demonstrate singularity we suggested an argument which is based on the behavior of a set of complex adaptive agents under some evolutionary force. . Research made in artificial life simulations have shown an increase in complexity due to an emergent property coming out of lower complexity constituents. We argued that if the elements of the AI system are complex enough, and the evolutionary driving force is of social type, then it could indeed be that such a process could yield some emergent new super-intelligent system.

Bibliography

Adolphs, Ralph. 2003. Cognitive neuroscience: Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience* 4, no. 3.

Aida- Artificial intelligence driven analytics Home Page. Available Online: http://aidatech.io/

Anderson, Bo. 1989. On artificial intelligence and theory construction in sociology. *Journal of Mathematical Sociology* 14, no. 2-3. pp. 209-216.

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

AutoDesk's Dreamcatcher Home Page. Available Online: https://autodeskresearch.com/projects/dreamcatcher

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.

Bedau, Mark, and Paul Humphreys, eds. 2008. *Emergence: Contemporary readings in philosophy and science*. Cambridge, MA: MIT press.

Boden, Margaret A. 1978. *Artificial Intelligence and the Natural Man*. London: The MIT Press. First published 1977.

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.

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.

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.

Calvin, William H. [1996] 2001. *How brains think*. New York: Basic Books.

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/

Collins, Allan, and Edward E. Smith, eds. 2013. *Readings in cognitive science: A perspective from psychology and artificial intelligence*. Elsevier.

Collins, Harry. M.1990. *Artificial experts: Social knowledge and intelligent machines (inside technology)*. Cambridge, MA: The MIT Press.

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.

de Kleer, Johan. 1990. The origin and resolution of ambiguities in causal arguments. In *Readings in qualitative reasoning about physical systems*, pp. 624-630.

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.

Dunbar, Robin IM. 1998. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* 6, no. 5, pp. 178-190.

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.

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.

Fox, Kieran CR, Michael Muthukrishna, and Susanne Shultz. 2017. The social and cultural roots of whale and dolphin brains. *Nature ecology & evolution* 1, no. 11.

Frith, Chris D. 2007. The social brain. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362, no. 1480. pp. 671-678.

Goertzel, Ben. 1993a. *The Structure of Intelligence*. New York: Springer-Verlag.

Goertzel, Ben. 1993b. *The evolving mind*. New York: Gordon and Breach.

Goertzel, Ben. 1994. *Chaotic Logic.* New York: Plenum Press.

Goertzel, Ben. 1997. *From Complexity to Creativity.* New York: Plenum Press.

Goertzel, Ben. 2001. *Creating Internet Intelligence.* New York: Plenum Press.

Goertzel, Ben. 2007. *Artificial general intelligence*. Edited by Cassio Pennachin. Vol. 2. New York: Springer.

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.

Gonzalez, Carlos. Jun 18, 2018. Seven Common Applications for Cobots. *MachineDesign*. Available Online: http://www.machinedesign.com/motion-control/7-common-applications-cobots

Hardesty, Larry. March 30, 2015. Crowdsourced tool for depression. *MIT News*. Available Online: http://news.mit.edu/2015/crowdsourced-depression-tool-0330

Hofstadter, Douglas R. 1995. *Fluid concepts and creative analogies*. New York: Basic Books.

Kang, Minsoo. 2011. *Sublime dreams of living machines*. Harvard University Press.

Laird, John E., Allen Newell, and Paul S. Rosenbloom. 1987. Soar: An architecture for general intelligence. *Artificial intelligence*33, no. 1. pp. 1-64.

Lenat, Douglas B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38, no. 11. pp. 33-38.

Lévi-Strauss, Claude, and John Weightman. 1969. *The raw and the cooked: Introduction to a science of mythology*. Vol. 1. New York: Harper & Row.

Landing.ai Home Page. Available Online: https://www.landing.ai/

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*.

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

McCarthy, John. 1978. History of LISP. In *History of programming languages I*, pp. 173-185. ACM.

McLuhan, Marshal. 1964. *Understanding Media: Extensions of Man*. New York: McGraw Hill.

Mead, George Herbert. 1934. *Mind, self and society*. Vol. 111. Chicago: University of Chicago Press.

Microsoft Research India workshop on ASI, 2007. Available Online:

https://www.microsoft.com/en-us/research/event/microsoft-research-india-summer-school-artificial-social-intelligence/

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.

Minsky, Marvin. 2006. The emotion machine. *New York: Pantheon* 56.

Newell, Allen, and Herbert Alexander Simon. 1961. *GPS, a program that simulates human thought*. No. P-2257. RAND CORP SANTA MONICA CALIF.

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.

Newell, Allen, and Herbert Alexander Simon. 1972. *Human problem solving*. Vol. 104, no. 9. Englewood Cliffs, NJ: Prentice-Hall.

Norman, Donald A., Rumelhart, David E. & the LNR Research Group. 1975. *Explorations in Cognition.* San Francisco: W.H. Freeman.

Ofria, Charles, and Claus O. Wilke. 2004. Avida: A software platform for research in computational evolutionary biology. *Artificial life* 10, no. 2. pp. 191-229.

Pennachin, Cassio, and Ben Goertzel. 2007. Contemporary approaches to artificial general intelligence. In *Artificial general intelligence*, pp. 1-30. Springer, Berlin, Heidelberg.

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*.

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

Postman, Neil. 1992. *Technopoly: the surrender of culture to technology*. New York: Vintage Books.

Quillian, M. Ross. 1968. Semantic memory. In Marvin Minsky (Ed.), *Semantic Information Processing*. pp. 216-270. Cambridge, MA: MIT Press.

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.

Schank, Roger C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology* 3, no. 4. pp. 552-631.

Schank, Roger C. & Abelson, Robert P. 1977. *Scripts, Plans, Goals, and Understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidhuber, Jürgen. (2007). Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pp. 199-226. Springer, Berlin, Heidelberg.

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/#

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*.

Strate, Lance. 2017. *Media Ecology: An Approach to Understanding the Human Condition*. Peter Lang Publishing, Incorporated.

Tegmark, Max. 2017. *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

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

Voss, Peter. 2007. Essentials of general intelligence: The direct path to artificial general intelligence. In *Artificial general intelligence*, pp. 131-157. Berlin, Heidelberg: Springer.

Voss, Peter. Oct 4, 2017. From narrow to general AI. *Medium*. Available Online: https://medium.com/intuitionmachine/from-narrow-to-general-ai-e21b568155b9

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

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

Yang, Ang, ed. 2008. *Intelligent complex adaptive systems*. IGI Global.

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.

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*.

Yudkowsky, Eliezer. 2004. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*. *San Francisco, USA*.

Yudkowsky, Eliezer. 2007. Levels of organization in general intelligence. In *Artificial general intelligence*, pp. 389-501. Berlin, Heidelberg: Springer.