HANDLE: http://hdl.handle.net/10216/110854

Beneficial AI: the next battlefield

Eugénio Oliveira

Faculty of Engineering, University of Porto and LIACC- Artificial Intelligence and Compute Science Lab; Porto, 4200-465, Portugal
eco@fe.up.pt

Letter from Academia

When planting our human print in a new technology-driven world we should ask, remembering Neil Armstrong in 1969, "after many small steps for AI researchers, will it result in a giant leap in the unknown for mankind?" An "Artificial Intelligence-first" world is being preached all over the media by many responsible players in economic and scientific communities. This letter states our belief in AI potentialities, including its major and decisive role in computer science and engineering, while warning against the current hyping of its near future. Although quite excited by several recent interesting revelations about the future of AI, we here argue in favor of a more cautious interpretation of the current and future AI-based systems potential outreach. We also include some personal perspectives on simple remedies to preventing recognized possible dangers. We advocate a set of practices and principles that may prevent the development of AI-based systems prone to be misused. Accountable "Data curators", appropriate Software Engineering specification methods, the inclusion, when needed, of the "human in the loop", software agents with emotion-like states might be important factors leading to more secure AI-based systems. Moreover, to inseminate ART in Artificial Intelligence, ART standing for Accountability, Responsibility and Transparency, becomes also mandatory for trustworthy AI-based systems. This letter is an abbreviation of a more substantial article to be published in IJCA journal.

Keywords. Artificial Intelligence, Beneficial AI

1 AI has an history

Scientific paradigm changes and relevant outcomes of civilization derive from intelligence at work. How can we improve and enlarge those benefits, through artificial intelligence (AI) based systems, without being fully replaced both on the job market and, most important, as final decision-makers?

AI has evolved, during the last five decades, starting with a very classical approach grounded on mathematics and psychology then followed by more romantic times in which almost everything was said to be possible for a computer to solve.

Intelligence, although somewhat difficult to formally define, can be recognized as having many facets including problem solving, learning, recognizing and classifying patterns, building analogies, surviving by adaptation, language understanding, creativity and many others.

It has been proved that it is not too difficult to build systems and algorithms that incorporate some kind of intelligence (although far from encompassing all the possible facets) whenever it is realized that "Most, if not all known facets of intelligence can be formulated as goal driven or, more generally, as maximizing some utility function" (Hutter, 2005).

After a more pragmatic attitude that lead AI researchers to develop Knowledge Based Systems in which transparency and explain-ability were mandatory for the sake of real applications, the new trend became a call back to the fundamentals in which learning, adaptation, cooperation and autonomy became corner stones of more sophisticated intelligent systems.

It was not very long ago that a rupture in the traditional step by step AI systems development happened and, together with euphoria, new warnings reached the scientific community about the future potential dangers of possible misuse of AI algorithms and systems.

This rupture happened since "Big Data" started to become available everywhere, "The Internet of Things" started to grow (the so-called "outernet") and new algorithms like those related with "Deep Learning" concept, lead to striking applications with huge economic and social impact. Reflecting upon such an impact is no longer a kind of an unnecessary distraction, "like worrying about overpopulation of Mars" in the words of Andrew Ng, quoted in (Das, 2017).

The class of algorithms usually referred as Deep Learning mostly rely on the artificial neural networks (connectionist) paradigm. Connectionist-based methods approach has the big advantage of avoiding the knowledge acquisition bottleneck since the proposed model is directly built from observations with very little human intervention. The disadvantage that comes together is that those systems mostly result in a kind of black boxes.

There are however a plethora of different approaches that AI researchers have followed in the past which are responsible for relevant AI based systems application. In "The Master Algorithm: The ultimate Learning Machine that will remake our world" (Domingos, 2015), the author identifies five AI and machine learning "tribes" that currently exist: the symbolists, connectionists, evolutionaries, Bayesians and "analogizers".

Although applying very different paradigms and coming from different schools, the objective is always the same: To develop machine intelligence.

AI has been repeatedly over-hyped in the past, even by some of the founders. As a consequence, so called "AI winters(s)" hit the field decreasing the potential impact of AI realizations. Nevertheless, for more than ten years ago, well known researchers like R. Brooks, a critic of GOFAI (Good Old Fashion AI) have opposed the idea that AI failed and warned that AI would be around us every day (reproduced in (Wess, 2014)). And he is indeed right!

This letter has in mind to challenge readers in the sense that, together with the recognition of the very relevant achievements AI has already reached, we should reflect upon the current excitement of its potentiality and future social impact. And doing so, it also warns against a new age of intensive overselling that raises huge expectations on AI-based systems without discussing their inherent dangers.

2 Can Artificial General Intelligence be dangerous?

Corroborating Brooks statement, AI-based systems have been around and useful in many different relevant, although narrow, domains. For example, they have been used to make specific medical diagnoses, to allow companies to build up consumer profiles, for satellites to be intelligently controlled, for search engines to do page ranking, for computers to intelligently filter spam. Recommender System such as those used by Amazon and Netflix are welcome. And we feel proud of amazing accomplishments of AI programs like those used by Deep Blue and AlphaGo, at least for the prestige. And who will deny the real importance of using "mentalistic" Agent architectures to represent investors in the stock exchange or, even better, to automatically recognize when skin steins are carcinogenic?

Moreover, Machine Learning (ML) algorithms are working together with a multitude of other algorithms in order to get solutions to complex situations. Siemens Healthineers and IBM Watson Health are tackling population health together. Through the combination of the clinical expertise of Siemens and cognitive computing sophistication of IBM Watson Health, it is already possible to make critical healthcare data meaningful.

Those kind of synergies are also responsible for the impact of AI systems in many domains, and sometimes also more controversial as it is the case of the self-driving car or those NSA algorithms that may decide if you are a potential terrorist or not...

Because most AI-based systems, in some way, reason and interact, we are often tempted to compare them to humans. We sometimes forget there are limitations that still make a great difference.

While humans are good at parallel processing (patterns recognition) and slower at sequential processing (classical reasoning), computers have only recently mastered the former in narrow fields and have always been superfast in the latter. We also can say that "`Just as submarines do not swim, machines solve problems and accomplish tasks in their own way." (Gerbert, Justus & Hecker 2017)

Moreover, according to some scientists and opinion-makers, we could expect that Super-intelligence or General Intelligence, would give artificial systems the property of consciousness, making the boundary between humans and machines, in many decisive aspects, fuzzy.

Artificial General Intelligence can be seen as an intermediate stage between what we have now, a kind of Artificial Specialized Intelligence that is very performant in restricted domains, and a conceivable future Super-intelligence that might endow artificial systems with the capability to exceed human performance in many, if not all, the relevant domains, possibly including leadership.

Some authors (Oliveira, 2017) are now putting the following question: Is the human brain the only system that can host a mind? If digital minds come into existence, and the referred author states that it is difficult to argue that they will not, we have to face all the legal and ethical implications of such a possibility.

It is argued that current hardware development rate, regarding miniaturization and integration, makes us believe that in a few years it will be possible to replicate the number of synapses happening at the brain level. I believe that reasoning patterns of high level of abstraction as well as structured knowledge are not always directly emerging exclusively from those simple operations. It is however worthwhile to prepare ourselves for this future possibility. Legislation and ethical principles should guide a harmonious development of either some kind of "digital minds" or even hybrid minds.

It is not yet the case that we foresee the possibility of humans becoming obsolete in too many situations, but it is the right time to clearly state that real Beneficial AI must be developed in such a way that humans and machines cooperate to solve complex problems together and, in doing so, possibly learn from each other.

More than having intelligent entities, robots, systems, computers, machines, programs, replacing humans everywhere, we need to develop processes, methods and regulations leading to a harmonious coexistence of both for humankind beneficial.

This ultimate goal justifies that we must pay attention to present signs that point to possible dangers in some future research directions of AI, leveraged by a plethora of books and scholarly opinions over-hyping the current and future role of AI.

Although I must express a few warns, I still am a real enthusiastic of the scientific development of the Artificial Intelligence field and stand for a firm position defending the crucial importance of the field.

Security and privacy, data integrity, distributed and parallel computation, software engineering development methods and many other computer science topics should have in mind the needs of intelligence-based systems.

Although this can be prone to controversy, Computer Science and Informatics should thus be seen as contributing to the broader field of Artificial Intelligence. An Artificial Intelligence confined by ethical principles for research, development and deployment.

3 AI realizations and "The Master Algorithm" Claims

"The Master Algorithm" (Domingos, 2015) is a remarkable book that makes us exercise our critical opinion without denying both the beauty and the dangers of its main message. "*Our goal is to figure out the simplest program we can write such that it will continue to write itself by reading data, without limit, until it knows everything there is to know.*"

To be able "*to know everything*", or to make people think that "it knows everything", could be in itself potentially dangerous, but things still change for the worse when the same author also claims that "*Machine learning is remaking science, technology, business, politics, and war ...*", (Domingos, 2015) showing the relevance of it.

Although this last claim may be accepted as partially true, it also reveals a well–known tendency to oversell a specific research topic, trying to ignore that, often, Machine Learning (ML) algorithms work together with a multitude of other different tools in order to get things done.

Artificial Intelligence should be neither glorified nor blamed in isolation for the important outcomes to appear soon.

It is true that ML algorithms look like artifacts that produce new artifacts. In some way, a "Master Algorithm" would be a powerful and absolute General-purpose learner, a kind of "Holy Grail" which, in reality, I believe will be very difficult to find.

If it exists, the Master Algorithm, seen as a combination of current ML algorithms working over big data, "*can derive all knowledge in the world - past, present, and future - from data*". Inventing it would be one of the greatest advances in the history of science. It would be, as the author names it, the "*ultimate learning machine*", (Domingos, 2015).

However, it definitively seems to me that, up to now, those algorithms work over data that, although collected in large amounts, have a relatively simple or already known structure. You do not need much extra knowledge to build up a theory that explains those extracted patterns. This is not the case whenever big data has to be first recognized and then extracted from many image-based sources (video, pictures, MRI- Magnetic Resonance Images) in which recognizing what is relevant in data also becomes a crucial issue. Apriori knowledge to guide the system focus of attention on different dynamic and noisy situations becomes of utmost importance for collecting and interpreting data.

Without our explicit consent, there are also large data brokers that collect, analyze and sell to others all the harvested details about consumers' online activities for marketing purposes.

It may even be the case that, who knows, whenever you decide to act differently from what was expected, when you are upset with your past choices and decide to do it radically differently, it may happen that you will become suspect to someone or some organization, seen as a disruptive person, half a way to become a potential terrorist...

Are current AI algorithms ready to derive all possible and needed knowledge from any kind of data sets? Of course not. You may supply hundreds of thousands of medical cases about, let us say, different cancer types, but if you miss a few tenths of cases regarding very specific situations, they will always remain invisible to the inferred algorithms.

Sundar Pichai, chief executive of Google and an AI enthusiast assures that "*Google is going to be AI first*". Very recently he even stated that "*In an AI-first world, we are rethinking all our products*," (see The New York Times, May 18, 2017).

Although he is confident that AI will make available a general tool designed for general purposes in general contexts, he also adds, and I fully agree, that "*for the moment, at least, the greatest danger is that the information we're feeding them* [AI-enhanced assistants] *is biased in the first place*" (Lewis-Kraus, 2016).

Reliable Data Curators become then necessary to guarantee that the recorded past is not adulterated and remains trustworthy.

Chaining and mixing existent different machine learning principles, may not be enough to solve the overall learning problem. Even if we accept the inherent power of data, it might take more than collected observations to directly induce natural selection "*as Darwin did*" (Domingos, 2015).

Is it just a matter of observing data? I do not believe it is only that.

There are specific abilities that some minds (and brains also) have developed, and others did not, to extract from, as well as apply to, the same data, in some identified contexts, more sophisticated knowledge than other minds. And, perhaps, there are many different capabilities that need to be developed in the future that, even the most gifted minds and brains cannot yet imagine.

We should also be cautious about the scope of AI and ML. In the same book it is stated that "*The Master Algorithm would provide a unifying view of all of science and potentially lead to a new theory of everything.*" (Domingos, 2015).

I recall that a Theory of Everything is sought because quantum physics only deals with the very small, Einstein's general relativity theory deals with the very big and we are looking for a unique theory that works everywhere.

However, physicists do not think that the Theory of Everything will come out of a kind of combination of the previous two theories mentioned before. They are still looking for something radically new. The same will happen, in my humble opinion, with the so-called "Master Algorithm" and it is an over simplification to believe that it (like a kind of "master key") will come precisely out of the ML algorithms that we already know now.

I am not as radical as those who state that "*big data is not the new oil; it's the new snake oil*" (worth of mouth). But, nevertheless, I would be more cautious in targeting the possible goals of current ML algorithms working over big data as the "*ultimate learning machine*".

4 "Artificial General Intelligence" and consciousness

Learning is becoming the hard kernel of AI, enabling more sophisticated and general-purpose AI-based systems capabilities. Artificial General Intelligence can be seen as fostering the property of consciousness. This property can also be translated as self-awareness or even capability of feeling (sentience).

John Searle, in his book "Minds, Brains, and Programs" (Searle, 1980), clearly states that. "A program cannot give a computer a mind, understanding or consciousness regardless its intelligence."

The main argument he used, the well-known Chinese room, seems more like a paradox which, like the Zeno paradox, contradicts observed events. This is the opinion of Jean E. Tardy, who in the book "Meca sapien blueprint" (Tardy, 2015) argues that Machine consciousness is feasible and can be an emergent property. Is it not the case that a movie is made of a large amount of static frames?

"Consciousness is equal to that specific capability also called sentience [capable of feeling] and self-awareness" (Tardy, 2015). As a definition it does not help much. Is awareness the acknowledgment of Self? How to define the Self?

Even if we admit, and I could, that it might be possible that some simple type of "consciousness" will emerge from very complex interactions of more primitive forms of intelligence included in AI-based Systems, we cannot assure that such a complexity will be reached with current "in silico" hardware systems.

Moreover, the possibility either to download a mind or to make it evolve from a simpler digital mind, and, here, I agree with the ideas expressed in "The Digital Mind" (Oliveira, 2017), would need an non existing reverse engineering capability of the brain or, for the latter alternative a kind of real body, plenty of sophisticated sensors, which is not yet available today.

However, to replicate “in silico” what exists “in vivo” in the biological brain seems to be, for now, out of our grasp as far as we can preview based on scientific grounds.

5 Mind the dangers

It is obvious that there are potential applications in which data gathering, data mining and Machine Learning algorithms outcome become not at all crystal clear and may lead to conclusions backing some kind of artificially justified dominance in many different aspects.

Taking the AI researchers’ role, we should be mainly concerned with establishing a set of practices and principles that may prevent the development of AI-based programs and systems prone to be misused for the bad of humanity. And the first major concern is privacy.

Many data mining algorithms rely in analyzing sensitive personal data including individual identification, photos, genetic and medical records or even brain signals.

We must enforce and support all the efforts trying to ensure that individual privacy will always be guaranteed and are not just feeding someone else’s commercial interests.

Are we over-reacting? Should we really be afraid of some potential future AI-based systems? Haven’t we always known how to deal with similar possible threats? Naïve answers to these threats can be: “remove the plug”, use a “kill switch”, use a “cage” (virtual machine), but current learning algorithms and data dispersion in the cloud make this kind of possibilities innocuous.

We have then to recognize that the problem is real and we, as researchers and developers, we need to take actions to reinforce AI-based systems security well beyond simplistic solutions. Individual privacy should not be for sale, specially by others.

6 The Human in The Loop

Developing Autonomous Software Agents taught me that it is always mandatory to include the human in the control loop. We have done that in different contexts like Airlines Operations Control (A. J. M. Castro, Rocha, and Oliveira 2014) and, also, to manage critical damages when ships are under severe conditions. One can never forget the intrinsic responsibilities assigned to humans (here, commanders and officers in the first place) in charge.

To make this possible in a transparent way, developers need to take human-machine interactions into consideration from the initial design steps. Therefore, appropriate systems specification methods, of AOSE- Agent Oriented Software Engineering kind, (Zambonelli, Jennings & Wooldridge, 2003) (Castro & Oliveira, 2008), become crucial in guaranteeing that we can trust the system.

Despite a good specifications practice, is it a definitive answer to AI and ML potential dangers to include the human in the loop? It might not be. We should not forget that Drones can fly autonomously and despite being monitored by humans, we should not be sure of the drone’s goodness in many different situations...

7 Is Rationality mandatory?

The recent western economic crisis made many economists to believe that it is wrong to build strategies upon computer-based models in which agents are believed to always act rationally. Real intelligent agents, in order to be included in economic models, should be aware of more emotion-based decision-making capabilities that go beyond strict economic rationality represented by what the 2017 Nobel prize in Economics calls "Econs" (Thaler, 2016).

In a different scientific domain, back in 1997, I published a short paper about "Robots as responsible Agents" (Oliveira, 1997). My naïve approach, twenty years ago, was that the then novel cognitive software agents architecture based on "mentalistic" concepts like "Beliefs", "Desires" and "Intentions" (BDI) could bring a positive influence in the designing of more self-aware robots controlled by those BDI software agents.

It was only about five years later that I realized that one important and decisive component of human-like reasoning is deeply related with emotions and could be helpful for intelligent AI-based systems.

Some, like John Searle (Searle, 2011), arguing, through an article in the Wall Street Journal, against real intelligence of IBM Watson, the program that brilliantly won the "Jeopardy" competition against humans, sarcastically said that the referred sophisticated program did not become happy after winning.

I, nevertheless, believe that it would not be very difficult to program Watson or other AI based system in such a way that, after winning the game, it would reach an "emotional state" similar to happiness. Not regarding the external signs of happiness, which would be too easy to implement, but in which concerns the internal reasoning capability changes, along with its way of acting and memorizing for a certain period of time, until that emotional state gradually declines.

Past experiences, in different scenarios and with different meanings, can be mapped to kind of primitive emotions (fear, anxiety, ...) intensity, through accumulator-like variables.

Including these "emotion-like" states in the reasoning loop, makes it more difficult to take decisions that possibly lead to bad results in terms of causing harm or some kind of pain to the agent or its environment. This implies that artificial and intelligent decision-making may benefit in taking into consideration these more human-like influential factors, like emotion states, in order to become more human friendly and compatible.

8 Ethical issues

I believe we do not want to see the boundaries between the individual self and artificial systems to dissolve. Are we ready to accept what the author of "The Master Algorithm" book said in a TEDX Talk: "the question what means to be human will no longer have an answer. But maybe it never did."? (in "Next 100 years of your life" (Domingos, 2016)).

Are we going to leave AI plus IoT (The Internet of Things), plus ML, to create some kind of future dystopia? Or will we be able to circumscribe the potential dangers and fortunately live with the obvious advantages of this new technology?

It seems that there is now a main concern of AI main players (from researchers to the big high-tech companies) leading to the searching for ethical laws that could prevent situations like those happening during the industrial revolution or even in those decades immediately after the development of nuclear energy.

To make the scenario still more strange, it may also be the case that a super-intelligence might not be perceptible as such. It could even be in the so-called "Technium", a huge network of computers.

That is why so many people are now contributing to the discussion on how to guide AI research development in such a way that, whatever results we will get in the future they will point to a beneficial AI age.

We, thus, stick in line with the 23 Asilomar principles pushing AI research towards the creation of, not undirected intelligence, but beneficial intelligence instead (Conference, 2017).

We are also aware of the efforts made by M. Delvaux, at the European Parliament, about the possibility to give intelligent robots a limited "e-personality", that could be comparable with what already happens with "Corporate personalities", a legal status which enables to sue or to be sued in court.

However, if we have learned something from the past about law, it is that it does not change as fast as technology does. We will have to wait a long time before relevant legal system changes will occur.

We prefer here to emphasize that we should enforce decisive principles to be applied to AI systems, like those brought from good Corporate Governance and that V. Dignum (Dignum, 2017) also advocates: To inseminate ART in Artificial Intelligence. Here, ART standing for Accountability, Responsibility and Transparency.

There is a need to know, in all circumstances, who is to blame whenever an AI based system's misconduct is noticed, the typical example being the situation of a self-driving car accident harming humans.

Hardware builders, software developers, licensor authorities, car owner, or the car itself? In fact all of them should be accountable.

Moreover, AI researchers and developers should take the responsibility to create models and algorithms to enable AI systems to take decisions, in such a way that they can justify them according to rational and logic principles. This is not the case with current deep learning based mechanisms.

It is also evident that, if algorithms are not transparent enough when making relevant decisions on our behalf, we cannot judge where the responsibility lies and how can we argue against the quality of those decisions.

9 Just to conclude

Stuart Russell, the well-known AI scientist drafted and became the first signatory of an open letter calling for researchers to look beyond the goal of merely making artificial intelligence more powerful. "*We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial*" (Russel, 2017).

Although some consider a myth that AI will either turn evil or conscious, we believe it is time to recognize the actual worry that AI is more and more turning competent and, simultaneously, there is a possibility that its goals become misaligned with well-formed human goals.

We remain excited about all the potential benefits of possible Super-intelligent either agents, systems, networks alone, or in cooperation with humans, and their respective relevant impact in the future human society. Meanwhile we believe that current glorification of AI is not proportional to the reality.

That impact may still be decades away.

Nevertheless, the scientific community in general and the AI community in particular, should be proud of launching all the interrogations that have to be made about the potential impact of AI in the future.

The promoted symposium dedicated to the social and economic impacts of artificial intelligence in the next 10 years (AI Now), by the previous White House Administration, was a very relevant forum for discussing social, inequality, ethics, labor and health domains in which AI is raising pressing questions.

According to Kate Crawford and Meredith Whittaker (Crawford & Whittaker, 2016), an uncomfortable truth has been revealed "*there are no agreed-upon methods to assess the human effects and longitudinal impacts of AI as it is applied across social systems. This knowledge gap is widening as the use of AI is proliferating, which heightens the risk of serious unintended consequences.*"

It is also possible that spontaneous generation of synergistic control systems that will be no longer accessible to human control is nothing but another myth. But we should never forget that any algorithm can be as biased as the data they draw on. As simple as that.

Even if we look at the present, we are not willing to replicate what happened with Microsoft Corporation Chatbot Tay that began to post offensive tweets, forcing Microsoft to shut down the service about 16 hours after its launch. In some specific scenarios, 16 hours could be too late ...

In conclusion, I would like to emphasize this letter main message. It is at least smart to start worrying about how to enforce human beneficial AI by using human intelligence to direct AI research in the benefit of humankind. We hope that, also in the future, ethical concerns will remain behind the law.

10 References

Castro, A. J. M., Rocha, A. P. & Oliveira, E. (2014). A New Approach for Disruption Management in Airline Operations Control. Studies in Computational Intelligence, V. 562. Springer.

Castro, A. & Oliveira, E. (2008). The Rationale Behind the Development of an Airline Operations Control Centre Using Gaia Based Methodology. In Int. J. Agent-

Oriented Software Engineering, Vol.2, N.3, 350–77.

Conference, Beneficial AI. (2017). Asilomar AI Principles. Accessed 30th May 2017. https://futureoflife.org/bai-2017/.

Crawford, K. & Whittaker. M. (2016). Artificial Intelligence Is Hard to See: Why We Urgently Need to Measure Ai's Societal Impacts. Accessed 30th May 2017. https://medium.com/@katecrawford/artificial-intelligence-is-hard-to-see-a71e74f386db.

Das, S. (2017). The Death of True Intelligence? Accessed 31st May 2017. https://www.linkedin.com/pulse/death-true-intelligence-subrata-das/.

Dignum, V. (2017). Robots and Ai Are Going to Take over Our Jobs! Or Work with Us for a Better Future? Accessed 30th May 2017. https://www.linkedin.com/pulse/robots-ai-going-take-over-our-jobs-work-us-better-future-dignum.

Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books.

Domingos, P. (2016). Next 100 Years of Your Life, Tedx Talks La. Accessed 30th May 2017. https://vimeo.com/200120546.

Gerbert, P., Justus, J. & Hecker, M. (2017). Competing in the Age of Artificial Intelligence. Accessed 31st May 2017. https://www.bcgperspectives.com/content/articles/ strategy-technology-digital-competing-age-artificial-intelligence/.

Hutter, M. (2005). Universal Artificial Intelligence. Springer.

Lewis-Kraus, G. (2016). The Great A.I. Awakening. In. The New York Times Magazine, Dec. 14, 2016.

Oliveira, A. (2017). The Digital Mind: How Science Is Redefining Humanity. MIT Press.

Oliveira, E. (1997). Robots as Responsible Agents. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Computational Cybernetics and Simulation, V.3, 2275–9.

Russel, S. (2017). Research Priorities for Robust and Beneficial Artificial Intelligence. Accessed 30th May 2017. Future of Life Institut, https://futureoflife.org/ai-open-letter/.

Searle, J. (1980). Minds, Brains, and Programs. Behavioral; Brain Sciences.

Searle, J. (2011). Watson Doesn't Know It Won on 'Jeopardy!' Accessed 16th August 2017. https://www.wsj.com/articles/SB10001424052748703407304576154313126987674

Tardy, J. (2015). The Meca Sapiens Blueprint: A System Architecture to Build Conscious Machines. Monterege.

Thaler, R. H. (2016). Misbehaving: The Making of Behavioral Economics. W. W. Norton Company.

Wess, S. (2014). AI: Dream or Nightmare. In. TEDx Zurich Talks.

Zambonelli, F., Jennings, N. & Wooldridge, M. (2003). Developing Multi-Agent Systems: The Gaia Methodology. In ACM Transactions on Software Engineering and Methodology, V.12, N.3, 317–70. ACM.