AI and Us: Existential Risk or Transformational Tool?

Neantro Saavedra-Rivano
Associate Fellow, World Academy of Art and Sciene
Professor Emeritus, University of Tsukuba (Japan)
Project Coordinator, ABIPTI (Brazil)
Brasilia, Brazil
neantro@sk.tsukuba.ac.jp

Abstract— **This paper analyzes the short-term and longer-term impacts of AI. While the short-term impact is deemed to be mostly positive, the longer-term impacts are considered to be disastrous under a variety of scenarios, including the adoption of man-machine symbiosis tools. The paper offers suggestions as to policy measures that could correct this disastrous outlook.**

Keywords— Artificial Intelligence, existential risk, man-machine symbiosis, global crises

I. Introduction

This paper looks at the challenges posed by the recent important advances in the field of Artificial Intelligence. We look at the impact of AI in the short term and the longer term, the dividing point between these two periods being the moment when AI entities would acquire self-consciousness and be able to reason according to their own views of the world. There is of course plenty of discussion on whether that state of "sentience" is possible at all for machines and also on when that would happen.

The next section discusses the impact of AI in the near future while sections 3 and 4 present two scenarios for longer-term impact that are perceived as most likely. The first scenario appears as extremely negative while the second scenario would save the world by incorporating AI into humans (the so-called man-machine symbiosis). Section 5 adds a missing component into the discussion, namely the recognition that our world is heterogeneous and that, as a consequence, adaptation in these scenarios would not be uniform. It is thus seen that the second scenario has a fundamental flaw and that the world would not become a better place after all. At last, the final section offers some policy guidelines that, if adopted, have the potential to correct for the heterogeneity of the world and bring back the positive features of the second scenario.

II. The Impact of AI in the Near Future

The point in time defining the border between our present and the "near future" is deliberately ambiguous but we will understand it as being defined by that moment when Artificial Intelligence (AI) capabilities will basically match those of humans. There is of course no consensus as to when that will happen and, undoubtedly, some capabilities will be matched (and surpassed) earlier than others. The work by Grace et al (2017) points towards a horizon of at least 40 years. As with any disruptive set of technologies, we can distinguish between positive and negative impacts of Artificial Intelligence.

A. Positive impacts of Artificial Intelligence

We can already appreciate some of the many contributions that AI and related technologies are making to our welfare. Safety of transportation will be enhanced by the adoption of autonomous vehicles and AI-assisted driving. The accuracy of medical diagnosis performed using AI systems will vastly outperform that of even the most accomplished physicians. Human security will be heightened thanks to the processing and analysis of data from a variety of sources and sensors. All of those improvements will combine to make a reality the notion of a "smart city" and this will bring huge positive effects to the welfare of citizens. These positive effects are related to some characteristics of AI systems that set them apart from human-driven systems. In the first place, the elimination of any bias potentially affecting the quality of the actions of the system. Thus, an artificial agent dealing with consumers will not care about their gender, race or other characteristics which might induce human agents to, consciously or not, discriminate among them. In second place, the removal of human error in the actions of AI systems (although, of course, human error might have contaminated elements of the systems through their design or deployment). AI systems can also assume tasks which, by their particular nature, are unsuitable for human participation (deep underwater or space operations, just to provide some obvious examples).

B. Negative impacts of Artificial Intelligence

It may sound like an old-fashioned complaint, but it is an undisputable fact that the progressive substitution of humans by AI systems impoverishes social interaction and, together with so many other trends of modern life, contributes towards the creation of a less humane and less kind society. In a similar vein, we may note the loss of the human touch and sensitivity from certain actions where humans (say, a family doctor, or your friendly local driver) will be substituted by AI systems. But perhaps the most severe complaint about undesired effects of AI, is the potential loss of jobs that is expected to take place

The author is grateful to the Research Support Foundation of the Federal District (Fundação de Apoio à Pesquisa do Distrito Federal – FAPDF) for financial support

as a consequence of the introduction of AI-enhanced systems. As with many aspects of this complex subject, there is no consensus about the anticipated magnitude of job losses due to the technological advances brought by Information and Communication Technologies (ICT), Artificial Intelligence in particular. Authoritative studies from the World Bank, World Economic Forum and others (see, for instance, World Bank (2016), Brynjolfsson & McAffee (2014), World Economic Forum (2016)) indicate that at least 50% of extant jobs might be lost in the foreseeable future. In some developing regions of the world these losses would be much bigger. There is of course the possibility that new occupations would be created that would partly make up for those surrendered to the machines. At any rate, the social cost of adaptation to new occupations would be severe.

C. *On balance*

If the positive and negative impacts of AI in the "near future" are put together, most analysts would conclude that the net effect is positive. That is certainly the position taken by the large companies that are active in the development of AI technologies and applications. Perhaps the largest reservation on whether the net effective is positive comes from the uncertainty surrounding the impact on employment of the new occupations that are expected to appear in this near future.

III. The Long Term I: AI as a Risk

The "long term" is, of course, what comes after the "near future" so that both terms share the same charge of ambiguity. We will present here a first scenario about the long term and will start by discussing the differences between humans and AI in regard to rationality. In a sense, it is easier to accept rationality for machines than it is for humans. This is because, in the case of machines, rational behavior follows consistently a set of rules which, at least in current machines, have been set in advance. The rules constitute a logical system which may or not be the logic of predicates commonly used in deductive reasoning. There are many other logical systems, such as modal logics and non-monotonic logics which are suited for particular contexts. Rationality for humans is much more complex. To begin with, human behavior is not only determined by a set of rules but also by emotion and instinct. The survival instinct, for instance, may easily induce a change of the set of rules that was apparently guiding observed behavior. A sudden shift from one set of rules to another does not necessarily mean irrational behavior but may instead correspond to adaptability to a rapidly evolving context. Going back to machines or robots, they can be easily equipped with behavioral logics which do not correspond to the conventional predicate logic and, in principle, there seems to be no obstacle with providing them with dynamic features allowing for shifts in behavior depending on the context. But we are nowhere near to endowing machines and robots with emotions and it is an open subject whether that is possible at all.

All of this is related to the concept of "sentience" or self-awareness. It is generally accepted that animals (at least higher species) are sentient and indeed this is one of the bases for the arguments in favor of animal rights. Sentience for artificial intelligence (AI) entities is another matter and the possibility of sentience for them is a hot topic of discussion among specialists in the field and analysts at large. Some, like Ray Kurzweil (2005) and Nick Bostrom (2014) are fully persuaded that machines will become sentient in the foreseeable future. Kurzweil is perhaps the most vocal and confident proponent of this idea and offers a 30 year horizon (that would be 2035, counting from the publication of his book) for the advent of what he calls "the singularity", that is, the point where machines will equal and surpass all human capabilities. Others are less sure whether sentience is possible or even meaningful for machines. The idea that we humans might be surpassed by machines is certainly disturbing and it is only natural that the debate is charged with emotion.

The fact is that most specialists in the burgeoning field of Artificial Intelligence accept the possibility that AI entities will eventually acquire some degree of self-awareness and that raises a host of questions about how our interaction will be with them. The best situation would be one where machines, despite their self-awareness, "know their place" and obey to our commands without questioning our motives. On the other extreme, a worst-case scenario would be one where sentient machines understand the world and their position within it in a way that is unfavorable to us, realize their superiority, and decide either to use us towards their (unfathomable) ends or simply dispose of us. For sure, there will a wide variety of AI systems, dealing with human health, safety, mobility and even war or defense. Within that multiplicity of AI systems, a variety of forms of self-awareness will emerge, and it is conceivable and even likely that at least some of them will be hostile to us having control over them. Even supposedly benevolent systems, such as those dealing with human health, might have their own ideas about the termination decisions concerning severely sick humans.

These prospects have led many famous scientists and public figures to state their concerns about the potentially dire consequences that further discoveries in Artificial Intelligence might have for mankind. The title of the book by James Barrat (2013), "Our Final Invention", makes clear the danger we humans, as a species, are facing. The famous physicist Stephen Hawking warned about the risks posed by super-intelligent robots on numerous occasions, as when he declared to the BBC that "the development of full artificial intelligence could spell the end of the human race" (Hawking, 2014). Another prominent figure who expressed worries about these developments is the scientist and entrepreneur Elon Musk, who is actually taking a more active role following up on his concerns.

IV. Longer term II: AI as a Transformational Opportunity

An intriguing possibility arises when trying to find ways out of the existential dilemma posed by apparently unstoppable advances in Artificial Intelligence. And that is the rather obvious point: if confronting AI systems appears as an unwinnable fight, why not join them, become one of them? Man-machine symbiosis was a concept first proposed by Joseph Licklider at a time when Artificial Intelligence was in its early years and was not yet perceived as a possible threat to mankind. His seminal work on the subject (Licklider, 1960) discussed several of the issues appearing when trying to augment human capabilities through a close interaction with computers. In his words, *"The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."*

The idea was ahead of its time as the technology to realize that vision was not yet at hand. The concept of man-machine symbiosis has resurfaced in recent years as a possible answer to the existential threat posed by the quick advances in Artificial Intelligence. Research in man-computer symbiosis is moving forward at a good pace and a community of researchers in the field has consolidated. One of their vehicles is the Symbiotic Interaction network (https://www.symbiotic-interaction.org/) which organizes international congresses on the field (see, for instance, Ham 2018). In addition, some firms are working to develop technologies to realize this vision. Perhaps the most famous, because of its association with Elon Musk, is Neuralink (https://www.neuralink.com/). Established in 2016, Neuralink is working on the development of ultra-high bandwidth brain-machine interfaces to connect humans and computers. These computing devices would be implanted directly in the brain and the first products would have medical applications for patients with brain damage caused by strokes. Of course, the longer-term vision is to make possible, for any human, the enhancement of a variety of capabilities through these interfaces.

If successful, these technologies would result in the "design" of super-intelligent humans, that is, humans with intellectual capabilities matching those of AI systems. Those people would be able to communicate with machines at their level and, presumably, control them despite them having achieved sentience.

This rosy vision, shared by scientists such as Kurzweil and others, means that mankind would transcend its current situation and move into another stage of evolution.

V. A Worldwide Perspective

At this point it is prudent to take a wider view of the issues at hand and, in particular, not to forget the huge and multi-dimensional disequilibria that characterize our world. Such disequilibria (wealth, knowledge and preparedness) exist not only among nations but within them as well. For instance, if we look at the short-term impacts of AI, they are bound to lead to a heterogeneous pattern of results. Thus, the positive consequences of AI that are expected to take place in the near future will be more significant for developed countries than for developing countries. In addition, studies about the "future of work" are unanimous in concluding that poorer countries will suffer more than wealthier countries from the job displacement produced by AI systems and robots.

Looking now at the longer term, and assuming scenario I (AI as an existential risk), it is conceivable that backward countries and regions within countries would suffer later from the repression or annihilation brought by AI entities. This is explained because those regions would be late in bringing the advances of AI technologies. Even so, the reprieve would not last for long as eventually the machines would expand their reach and attain all regions of the world.

But it is scenario II (transcendence of mankind) which appears as most worrying. The advent of super-intelligent humans would appear first in the more advanced regions of the wealthiest countries. It is of course theoretically conceivable that this new race of super-humans would be kind and gentle and move fast in order to spread the tools of transition to the rest of mankind. The historical record, unfortunately, does not justify this kind of optimism. A more likely, and sinister, possibility is that the world would be split into two classes, the new super-humans and the common people, with the latter being subjugated by the former.

VI. Some Remarks on Policy

We saw in the previous sections that the consequences of AI in the near future are bound to be more favorable to advanced countries and advanced regions within countries. More precisely, the positive impacts will affect more markedly those regions while negative impacts will be felt more noticeably on less developed countries and regions. Of course, policies can contribute to mitigate those imbalances. Those technologies that are expected to have positive impact (on health, mobility, human security and others) could be more widely shared. The facilitation of technology transfer and conscious efforts at human capacity building would be important tools in this respect. As for the negative impact, arising mostly from the loss of jobs through substitution of humans by machines, those same tools can contribute to controlling this impact.

When looking at the longer-term impact of AI, at least at first sight, scenario II (the transcendence of mankind) looks better than scenario I (AI as an existential risk). However, after closer examination, it is clear that both of them are disastrous. One of them would result in the annihilation of mankind at the hands of machines of our creation; the other would witness the emergence of a "brave new world" where an elite of super humans would control and dominate an underclass of normal people. It is quite clear that the world faces a global crisis with potentially catastrophic consequences (see Saavedra-Rivano, 2016, for a general discussion on global crises).

As bleak as this outlook appears to be, there is still time to prepare and avoid an appalling future. First of all, we must recognize that, even if we so wanted, progress cannot be stopped. Secondly, we also need to acknowledge that machines will eventually become smarter than "pure" humans. The development of ever smarter AI systems is an ongoing process that will continue. They already are better than us in many endeavors and those areas keep expanding. What we need are governance and international cooperation initiatives that go in the direction of guiding the development of Artificial Intelligence in such a way that it becomes a force for the good of all of mankind rather than the threat many perceive it to be.

The following general policy guidelines are proposed:

- Education programs throughout the world must incorporate into the general curriculum disciplines that enhance computer and technological literacy of the genetral population
- Develop cooperation programs to foster wide sharing among nations of research on Artificial Intelligence and of its applications
- Promote an immediate intensification of research (theoretical and applied) on man-machine symbiosis with the objective of readying the world population for full incorporation of AI advances. Such research is to be geographically distributed so as to ensure that all nations participate in these activities. The aim of this research activities is to contribute to the preparation of mankind for its transition to a stage

where it can interact with machines and AI entities from a position of advantage

- Establish supervisory structures on AI research in order to prevent the emergence of sentient AI before mankind is ready to make the transition to the stage mentioned before.

Consistent and effective application of these guidelines requires the establishment of international governance structures and, of course, the wide recognition of the urgency and gravity of the current situation.

REFERENCES

[1] J. Barrat, Our Final Invention: Artificial Intelligence and the End of the Human Era. New York, NY: Thomas Dunne Books, 2013

[2] N. Bostrom, Superintelligence: Paths, Dangers, Strategies.Oxford, UK: Oxford University Press, 2014

[3] E. Brynjolfsson and A. McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, New York, NY: W.W. Norton & Company, 2014

[4] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When Will AI Exceed Human Intelligence? Evidence from AI Experts," Future of Humanity Institute, Oxford University, 2017 (available at https://arxiv.org/pdf/1705.08807).

[5] S. Hawking, Interview to the BBC on 12/2/2014 (https://www.bbc.com/news/technology-30290540)

[6] A. Hussain, The Sentient Machine: The Coming of Artificial Intelligence. New York, NY: Scfribner, 2017

[7] R. Kurzweil, "The Singularity is Near: When Humans Transcend Biology. New York, NY: Viking, 2005

[8] J.C.R. Licklider, Man-Computer Symbiosis, IRE Transactions on Human Factors in Electronics, HFE 1, vol. 1, pp. 4-11, March 1960

[9] N. Saavedra-Rivano, "Towards an Understanding of Global Crises," Cadmus Journal, vol. 2, issue 6, pp. 149-157, May 2016

[10] World Bank, World Development Report 2016: Digital Dividends, Washington, D.C, 2016

[11] World Economic Forum, The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution, World Economic Forum, January 2016