FROM THE ANTHROPOCENE TO THE MACHINOCENE?

Mario De Caro

Abstract

Is the Anthropocene near the end? Some reasons suggest that it may be so, because of the way machines – and especially intelligent machines – are dramatically changing our lives. Firstly, for the first time in history, new machines may be generating a dramatic increase in unemployment, which can cause severe economic and social problems. Secondly, human error or malice, applied to military or industrial machines can produce terrible consequences for humans and the natural environment. Thirdly, and more importantly, a time may come (the so-called "Singularity") in which artificial intelligence may become uncontrollable and very dangerous for us. Against a common opinion according to which machines cannot do what we do not tell them to do, I will discuss a case in which machines are not just much better than us, but are already creative in ways that we cannot anticipate or even understand.

Keywords: Anthropocene, Machinocene, Singularity, Artificial intelligence, Ethics of technology.

Apocalyptic and dystopian scenarios portraying the overthrow or destruction of humanity have been a pervasive part of our culture, starting at least with Aldous Huxley's *Brave New World* (1932), with many countless short stories and novels that followed in its wake. In the same years, movies started to portray ominous futures in which humanity is deeply at risk or doomed altogether – and this became a cinematic genre that is still very much alive today. Sometimes, in these works, the dooming factors are not realistic (alien species, resurrected dinosaurs, gigantic asteroids), but other times they reflect the most severe worries of ours. The possibility of humanity's decadence or even disappearance does not look too remote anymore. And this happens for several reasons.

In the past decades, humans have begun to feel at risk for the possibility of nuclear wars, possibly triggered either by mistake or by some uncontrollable Dr. Strangelove. Afterwards, other concrete global threats emerged:

chemical and bacteriological weapons, climate change, pandemics that could become uncontrollable. According to some scholars, however, there is a much worse menace, one that we ourselves have created, and that is developing at a whirlwind speed: Artificial Intelligence. Stephen Hawking, along with others, so wrote about this issue:

> Success in creating effective A.I. could be the biggest event in the history of our civilization. Or the worst. We just don't know… Unless we learn how to prepare for, and avoid, the potential risks, A.I. could be the worst event in the history of our civilization (quoted in Kharpal 2017).

Hawking's idea is that technological progress, in addition to enormous benefits for human living conditions, can also bring the seeds of human catastrophe with it. If this is right, the Anthropocene may soon be replaced by the Machinocene. In this article, I will therefore discuss the not-too-remote scenario in which artificial intelligence empowers itself to the point of causing enormous damage to humankind, regardless of its designers' will. Preliminarily, however, I will deal with two others less terrible but more concrete threats related to technological development: the endemic high unemployment that technological advances may generate and the potentially distorted uses of the new technologies.

1. *Unemployment and misuse*

In March 1811, during the industrial revolution, the first Luddite revolt broke out in Nottingham. Organized groups of workers sabotaged the new industrial machines (such as the mechanical chassis and the steam engine), which they saw as harbingers of unemployment and lower wages. However, it was not only the workers who were concerned about technological progress; the economists themselves did not look with particular optimism at the automation of Labor. Thus, David Ricardo, who at first regarded machines as beneficial tools for both industrialists and workers, concluded that they represented a danger to workers' employment. And even Marx and Engels – who had attributed a great emancipatory potential to the machines ("the warfare cannot be abolished without the steam engine," had they written in *The German Ideology*) – argued that in capitalist society the use of machinery very much deteriorated the conditions of the proletariat, both in industry and in agriculture.

More generally, at every major technological breakthrough, there have always been many who have diagnosed severe damage to employment levels.

In reality, however, these diagnoses have always proved overly pessimistic. On the one hand, technological innovation has often improved the living conditions of the workers. On the other hand, as new machines replaced human beings in areas that traditionally were their prerogative, new professions were born, dedicated to the construction, control, and maintenance of those machines. Consequently, despite widespread pessimistic predictions, technological progress did not increase unemployment at all (Visco 2015).

Today, however, the situation has changed profoundly, and the menace to employment caused by technological progress has become extremely serious. An example can help to understand the problem. In the United States, the most important professional sector is that of motor vehicle drivers. However, according to some reliable estimates, in a few years, with the introduction of automatic driving, five million drivers of motor vehicles will lose their jobs since their vehicles will be replaced by much safer and cheaper driverless ones. More generally, on the one hand, the progressive robotization of many human tasks is making our lives easier; but, on the other hand, it threatens to cause massive unemployment, especially in low-skilled sectors. For the first time in history, rising unemployment of the lower-skilled labor force is a potentially very worrying side effect of technological progress – and this will be one of the main challenges of politics, economics, and law in the coming years.

The solution to this problem has to be, first and foremost, political. The mechanisms of social protection have to be expanded and modified to allow the livelihood of families and entire social groups who may soon find themselves in very precarious economic situations. It is also essential that the governments' attitudes towards vocational education and training become more far-sighted: the young people of today – who will face a complex future in terms of employment – need to be equipped with new skills and greater cultural awareness. Thus it is indispensable to enable all future citizens– and not only the usual small privileged percentage – to understand and master the new technologies, which are going to become more and more pervasive in the decades.

Potential unemployment, however, is only the most obvious problem generated by contemporary technology's progress. To make only a few other obvious examples, one can mention the economic, legal, moral, social, and political challenges connected with the vigorous development of the new forms of artificial intelligence, home automation, and online hyperconnectivity. Of course, the proposed solutions to these challenges advanced by neoluddists, misoneists, conspiracy theorists, and other enemies of technological progress (often inspired by archaizing philosophies) are deeply inadequate and ill-advised. What is the attitude that we should take, then, toward our society's great changes brought in by technological progress? This subject is extremely complex but what is certain is that, symmetrically to the misoneists, the techno-enthusiasts are not able to adequately set the problem. The swirling advance of technology is made of light and shadow – and ignoring one or the other does not help to understand how to manage it.

A different problem that progress brings with it is the morally dubious employment of new technologies. In this sense, we can mention the growing use of algorithms in the legal field. California, for example, has started to use them to decide whether to grant parole to inmates who request it. The results of this new practice, however, are very controversial because they are conditioned, at least in part, by the judges' biases regarding the inmates' socio-economic conditions and ethnic identities. This fact has raised alarm in organizations that care about civil and legal rights, especially of minorities. That said, perhaps something can be said (at least in principle) to defend the application of algorithms in the judicial field. First, the biases that have emerged in the way algorithms decide cases clearly reflect the biases of the data given to the algorithms so that they can make their own decisions; and these data are nothing more than the decisions previously made by human judges. In this perspective, one could speculate that it may be easier to improve algorithms rather than humans in order to make them "race-blind" or "social condition-blind", considering that the latter are notoriously resilient in this respect. But there is more: in addition to racial and socio-cultural bias, a few years ago, a famous study showed that the decisions of human judges may be surprisingly spoiled by non-rational factors that should have no relevance for those decisions. As Gustavo Cevolani and Vincenzo Crupi explained (2018):

> In a well-known 2011 study, the authors examined the decisions of eight Israeli judges who took turns in two courts over a ten-month period. Data were collected on fifty daily sessions, during which the judges had to decide in favor or against the request for parole advanced by the inmates of the penitentiary institutions (in total 1112 decisions were recorded, 64% of which were against the granting of parole). The purpose of the study was to record the percentage of positive decisions (i.e. in favor of the inmate) and its daily trend. In this light, each day was divided into three periods, separated by the two breaks that the judge took to rest and consume a snack or a lunch (the time of the breaks was at the discretion of the judge). The results were striking: the percentage of decisions in favor of parole was regularly around 65% at the beginning of each of the three periods (i.e., at the morning opening of the session; immediately after the first break; and immediately after the second), and then went inexora-

> bly down to almost zero towards the end of the same period (and in any case stayed well below the 20%). In other words, it seems that an inmate has much better hopes of being granted parole if their case is discussed by a "fresh" judge, early in the day or after a break; but their chances shrink drastically as the session progresses and are almost nil ahead of the next break when the judge is supposedly tired, bored and hungry.

In short: it is has been known for a long time that prejudices of various kinds do frequently influence human judges' decisions; but even more worrisome is the new finding that their decisions may be influenced or even determined by purely biological factors such as fatigue, boredom, or appetite. To the advantage of machines and algorithms, it could then be noticed that they do not get tired or hungry: that is, they cannot be conditioned by the primary needs that condition human beings. Who knows if in a not too distant future, algorithms may offer better guarantees than humans in the administration of some branches of justice or (and this is perhaps more plausible) that they will not suitably help human judges, limiting their biorhythmic and appetite conditioning.

Another case of morally controversial applications of the new technologies is the use of artificial intelligence in the military. The effects in this area are now well known, and one of the main ones is the use of drones for scouting hostile territories or carrying out attacks against enemies. A useful parameter for assessing how much things change with the use of new technologies is offered by the engagement rules. According to a traditional rule of engagement of the US Army, for example, officers may not order an attack if its predictable effect is that the losses of the American forces will exceed 25% of the total loss (which means that a necessary condition for ordering an attack is that one can anticipate that the enemy will have triple losses than the US Army). A norm of this kind strongly limits the situations in which one can carry out attacks. However, with the introduction of drones all this has changed because the cost-benefit calculation becomes economical: one has to compare the risk of losing the drone to the damage inflicted on the enemy. This, of course, greatly facilitates the possibility of attacks, even in risky conditions in which one would refrain from using one's troops. In this way, the possibility of having new wars – or making conflicts already underway bloodier – increases noticeably.

Finally, there is the most threatening case, that is, when new technologies are used to support a totalitarian state. Writes John Lanchester (2019):

> Imagine a place in which there is a police station every hundred meters, and tens of thousands of cameras connected to a system of government facial

> recognition; where individuals are obliged to keep in their cars a GPS system operated by the police and to be able to make gasoline only, after having made do a scan face; where, in all the cell has been installed an application that monitors the activity of their holders, and prevent access to "harmful information"; in which the religious activity is monitored; where the state knows if anyone has family and friends abroad and where the government offers free medical visits in order to obtain citizens ' fingerprints, their eye scans and examples of their DNA. There is no need to imagine such a place, because it already exists: this is how the Muslim minority of the Uighurs lives in Xinjiang. Increasingly, in Xinjiang police checks have an algorithmic basis.

However, as Lanchester himself notices, Western democracies are not at all immune from this Orwellian situation. The same data that the Chinese government uses with sharp-eyed ferocity to oppress the Uighurs minority in the Western world are owned by the large corporations that dominate the world of new technologies. In particular, not always the so-called "Big Tech" or the "Big Five" (Google, Amazon, Facebook, Apple, and Microsoft) can stop, and sometimes they do not even try to, the immoral and criminal uses of their platforms – as shown by recent cases, such as that of Cambridge Analytica. In this area, it is essential that democratic governments both place strict limits on the uses that these companies can make of new technologies and big data and try to impose compliance with these limits on autocratic regimes. However, it is doubtful that this will happen easily, because of the vast influence these companies have on policymaking due to their tremendous economic assets and ability to influence the elections (which is one of the most problematic points of the whole issue).

2. *The spectrum of Singularity*

We have considered two threats posed by the rapid development of new technologies: first, the (very concrete) possibility that, in the coming years, unemployment among the less skilled may rise by a great deal; second, the controversial uses of new technologies, both at the public and private levels. Now there is a third challenge to consider.

There are clues that the moment may be near when intelligent and self-conscious artificial creatures will mingle with us with not-so-peaceful intentions. This prospect makes readers and viewers all over the world shudder: and, in this sense, one can mention the dystopian ferocity of HAL 9000, Terminator, *Blade Runner*'s replicants, *The Matrix*'s subjugating A.I., and *Ex Machina*'s delightfully ruthless Ava. However, this is not an issue regarding science-fiction: not a few contemporary scholars envision a scenario in which machines become a real threat to us. This scenario is called "the technological Singularity" or simply "the Singularity", the supposed time of the future when the development of artificial intelligence will become uncontrollable and irreversible – when, in short, A.I. will become intellectually and morally autonomous from its human programmers. Singularity – this is the idea – will cause radical changes: "our civilization" will become "their civilization".

James Barratt (2015) describes A.I. as our "latest invention," an invention that will cause the end of the human era, and a few years earlier, Ray Kurtzweil (2005), a theorist of the Singularity, announced that this catastrophe will occur around 2045. Nick Bostrom – an Oxford philosopher who is the most famous Nostradamus of the Singularity – wrote that, in our interactions with artificial intelligence, we are "like small children playing with a bomb" and that it is indispensable to place limits and time constraints to technological growth. According to Bostrom, the threat of machines to the survival of the human race is more significant than that represented by climate change. Our urgent goal, in his opinion, should be that of maximizing "the probability of an 'OK outcome' where an OK outcome is any outcome that avoids existential catastrophe" (Adams 2016).

In this perspective, it becomes essential to carefully control the developments of A.I., limiting its threatening potential. Bostrom thinks about putting legal constraints on A.I. development, but this raises two problems. Firstly, there is always the possibility that certain countries and individuals may escape these rules. This, however, is a problem of police control, and we are not particularly interested in it here. The second problem is more interesting for us: what kind of legislative action should we take to develop artificial intelligence while depowering its danger?

Famously, Isaac Asimov gave us some preliminary indications when he tried to think about the limits to be placed on the machines of the future so that they would not turn against their human builders. In this light, Asimov formulated his famous "Three Laws of robotics", which are still mentioned in the philosophical discussions on this topic:

> FIRST LAW. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
>
> SECOND LAW. *A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
>
> THIRD LAW. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

Later, Asimov realized that one can imagine cases in which, for the sake of humanity, a robot *should* harm specific human beings (and in extreme cases, even kill them). Imagine the case of a terrorist who is about to commit a terrible massacre: if A.I. artifacts can stop that terrorist, they must do so even if this would imply the violation of the first law of robotics. For this reason, Asimov introduced another law, more fundamental than the others, the "Zeroth Law":

> ZEROTH LAW. *A robot may not harm humanity, or, by inaction, allow humanity to come to harm.*
>
> Having introduced this new law, Asimov had to reformulate the other three:
>
> FIRST LAW*. *A robot may not injure a human being or, through inaction, allow a human being to come to harm, provided that this does not contravene the Zeroth law.*
>
> SECOND LAW*. *A robot must obey the orders given it by human beings, provided that such orders do not contravene the Zeroth law and the First law.*
>
> THIRD LAW*. *A robot must protect its own existence, provided that this self-defense does not conflict with the Zero law, the First Law and the Second Law.*

Asimov's laws are aimed at programmers, so that they do not design machines able to violate them. However, if the problem were just that, the machines' threat would not be very different from that presented by weapons of mass destruction, about which the international bodies legislate and the individual nations sign bilateral treaties to prevent distorted uses by human beings. Nevertheless, technological progress also poses other threats. The first is that, simply, programmers may be wrong in designing A.I. machines such that those machines may cause unintended harm to humanity. This threat is analogous to that represented by accidents in nuclear power plants (such as Chernobyl or Fukushima): in both cases, technology may cause destruction because of human ineptitude, carelessness, and lack of oversight. However, the real nightmare is another one. Let's think of the anxiety caused by Hal 9000, Terminator, & Co, that is, the fear that machines reach the ability to program themselves and turn against humans. They may then try to subjugate them or, in the most catastrophic scenario, even to exterminate them.

In this pessimistic scenario, machines are conceived of as intentional agents that can intentionally turn against the humans who built them. However, some experts do not believe that in the near future we will be able to build machines endowed with free will, intentionality, and conscience – that is, machines that one should consider as full-fledged agents.

In this regard, it is interesting to consider some potentially disturbing aspects regarding A.I. machines built in recent years. We have known for a long time that machines can offer much better performance than humans in several areas (think of expert systems). Besides, for several decades, we have also known that, based on the programs with which they are built, machines can improve their performances in dealing with experience. Today, however, we have reached another stage of this process: a stage that, to be pessimistic, could also outline a terrible threat in the not-so-distant future. Now, some machines that are able to improve themselves by giving themselves the rules to do so – rules that we are not able to understand fully. These machines can progress creatively in directions that may be completely unpredictable for us.

An example will clarify this point. Let's consider the history of computer chess, which traditionally has been seen as the litmus test of A.I. advances. If we now see that such history has been successful, it is interesting to remember that, for several decades, computers were not very good at playing chess against humans. In this regard, in the famous *Gödel, Escher, Bach: An Eternal Golden Braid* (1980, 152), Douglas Hofstadter wrote:

> In the early days of computer chess, people used to estimate that it would be ten years until a computer (or program) was world champion. But after ten years had passed, it seemed that the day a computer would become world champion was still more than ten years away.

However, as is well known, things had a sudden turn in 1996, when the computer Deep Blue defeated the world champion Garry Kasparov – arguably the best chess player in history –, in a six-games match (the final result was 3½ to 2½). Ever since, computers have become increasingly better than humans in playing chess, and now the dominance of machines has become almost embarrassing. During the 2018 world championship, played in 2018 by Magnus Carlsen and Fabiano Caruana, the grandmasters who commented on the games used computer programs – especially Stockfish, which then was the world champion chess computer – to judge how good the moves played by the contenders were and which player had, after each move, a strategic and tactic advantage over the other. The chess computers used by the commentators on that occasion, however, were programmed in the traditional way. Programmers, helped by the best chess players, had programmed them with hundreds of notions of human strategy and tactics and a gigantic amount of games played in the past. On this basis, the computers' spectacular computational force did the job.

After the 2018 World Championship, however, something shocking happened: Stockfish was challenged, and gutted, by a new computer, AlphaZero, which had been built based on entirely different principles. The numbers of the match between the two machines are impressive: in a first series of 100 matches, AlphaZero won 28 times and tied 72 times, without any loss. In a second series of 1000 matches, AlphaZero won 155 times, tied 839, and lost only 6 times (0.6%). The dominance of AlphaZero, therefore, was indisputable. The most interesting thing, however, is to understand how this happened. While Stockfish, the defeated computer, analyzed 60 million positions per second, AlphaZero analyzed only 60,000 positions. In short: AlphaZero analyzed a mere thousandth of the positions analyzed by Stockfish; but, despite having only a fraction of the computational strength of its opponent, AlphaZero triumphed. Where was its incredible strength, then?

AlphaZero's programmers, headed by David Silver, explained in two articles published in the most prestigious scientific journals (*Nature* and *Science*) the force of this unbelievable machine. The fundamental point was that they taught AlphaZero only the most basic chess rules, without inputs regarding tactics and strategy or any previously played games (as it instead happened with all previous chess-machines). Rather, the builders made AlphaZero play millions of games against itself: from these games, depending on the outcomes, AlphaZero deduced its own tactical-strategic principles, partly unknown to us, to be followed in each particular case. In a word, this machine learned to play chess on its own, by trial and error, and so it became by far the strongest player of all time.

When the best human chess players analyzed AlphaZero's games, they discovered brilliant moves, sometimes even incomprehensible to them – moves that challenged the fundamental principles on which humans and other computers have always set their way of playing (principles such as those relating to the relative importance of the pieces or the relevance of the pawn structure). In short: AlphaZero is not only practically unbeatable, but human beings cannot even quite understand how it thinks! Moreover, the surprises are not just those. AlphaZero also tore away the champions and the best computers that play go and shogi (Japanese chess), which computationally are games much more complex than chess. Also, in these cases, AlphaZero was given only the basic rules: for the rest, it learned everything himself. As Garry Kasparov wrote:

> Chess has been used as a Rosetta Stone of both human and machine cognition for over a century. AlphaZero renews the remarkable connection between an ancient board game and cutting-edge science by doing something extraordinary (quoted in Silver *et al.* 2018).

In the abstract of an article published in *Science*, Silver *et al.* (2018) so wrote about the triumph of their machine against the world champion of Go:

> The game of chess is the most studied field in the history of artificial intelligence. The best programs are based on a combination of research strategies, domain-specific adaptations and craft evaluation functions, refined by human experts over several decades. AlphaGo Zero has recently achieved superhuman performance in the game of Go through the reinforcement obtained by playing alone. In this article, we generalize this approach into a single AlphaZero algorithm, which can achieve superhuman performance in many intellectually challenging games. Beginning to play randomly and without having any prior knowledge of those games, if not their basic rules, AlphaZero defeated the world champion programs in chess, in shogi (Japanese chess) and in Go.

As said, in order to improve its play, AlphaZero only played against itself. The amount of training that the system requires depends on the each game's complexity, but it was extremely fast in all cases: for chess it took 9 hours, for shogi 12 hours and for Go 13 days. AlphaZero chooses its moves by using a Monte Carlo tree search, a heuristic that only analyzes the most promising moves, expanding the search-tree by considering random sampling of the search space.

In chess, in particular, there are 10^{47} possible positions – an astronomical number. That said, while other chess programs attempt to compute as many positions as possible, using their brute force computational force, AlphaZero self-taught using a Monte Carlo tree search (MCTS).- This heuristic analyzes only the most promising moves, which are a small fraction of the positions analyzed by a conventional computer. More precisely, AlphaZero's search is limited to analyzing random examples of the research space and assessing whether they lead to positive consequences. In some ways, then, AlphaZero resembles quantum computers more than traditional ones.

According to many experts, AlphaZero shows that it is creative in choosing the moves and strategies it plays. In this regard, so writes the chess Great Master Matthew Sadler:

> [In chess] traditional search engines are exceptionally strong at making few obvious mistakes, but they can go astray when faced with positions that do not have concrete and calculable solutions. It is precisely in those positions, where "intuition", "foreboding" and "intuition" are needed, that AlphaZero gives the best of himself (quoted in Silver *et al.* 2018; see also Sadler & Regan, 2019).

Chess, go, and shogi are only board games, somebody could say: one cannot infer much from those cases to much more complex ones. Still, besides the fact that Silver is now trying to apply AlphaZero to medicine, the experience of his creation suggests that we are approaching the moment in which machines may become much better than us in performing complex tasks without the need for us to help them understand how to perform those tasks. They will be able to do everything themselves. It seems fair to wonder, then, whether we humans will remain able to prevent (possibly using laws inspired by Asimov's) the possibility that this surprising new ability of machines completely escapes our control, as Bostrom and other futurologists fear. The answer to this question is not yet known. Let's hope it will be positive.

References

Adams, T.
2016 "Artificial intelligence: 'We're like children playing with a bomb'", *The Observer*, https://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine.

Barrat, J.
2015 *Our Final Invention: Artificial Intelligence and the End of the Human Era*, Thomas Dunne Books, New York.

Cevolani, G., Crupi, V.
2017 "Come ragionano i giudici: razionalità, euristiche e illusioni, cognitive", *Criminalia. Annuario di scienze criminalistiche*, ETS, Pisa, pp. 181-208.

Hofstadter, D.
1980 *Gödel, Escher, Bach: An Eternal Golden Braid*, Basic Books 1979, Vintage Books Edition.

Kharpal, A.
2017 "Stephen Hawking says A.I. could be 'worst event in the history of our civilization'", https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html?&qsearchterm=

Kurtzweil, R.
2005 *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Lanchester, J.

2019 “Document Number Nine”, London Review of Books, 41 (19), https://www.lrb.co.uk/the-paper/v41/n19/john-lanchester/document-number-nine

Sadler, M., Regan, N.

2019 *Game Changer AlphaZero's Groundbreaking Chess Strategies and the Promise of A.I.,* New in Chess, Alkmaar.

Silver M. *et al.*

2018 “AlphaZero: Shedding new light on chess, shogi, and Go”, https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Visco, I.

2015 *Perché i tempi stanno cambiando*, Il Mulino, Bologna.