Will we be masters or slaves of the Information Technology of the future?*

Mario Verdicchio
University of the West of Scotland
School of Media, Culture and Society
mario.verdicchio@uws.ac.uk

Abstract

More and more very sophisticated computers and robots assist us not only in high risk endeavours like space missions but also in our everyday life, and we may wonder where such a technological development will take us in the future. Some researchers seem to try to scare us with dystopian sci-fi scenarios where machines rebel and take over humanity, but the real risks are elsewhere and much more real than we may think.

Keywords

Artificial Intelligence; Future; Robots; Society; Technology

1 Is There a Problem with Artificial Intelligence?

Nowadays there are many expressions of concern, if not alarm, regarding the future developments of Computer Science in general and of Artificial Intelligence (AI) in particular, and their possible consequences for humanity.
An example with much resonance in the media is an open letter, signed not only by AI researchers but also scientists from other fields and world renown entrepreneurs, entitled "Research priorities for robust and beneficial artificial intelligence" and published by the Future of Life Institute, based in Oxford, UK [1].
The main point of the letter is a recommendation to widen the context of AI research to include, in addition to the goal of making it more sophisticated and capable, also that of "maximizing the societal benefit of AI". Such words suggest that we cannot assume AI to be a discipline that is only beneficial to humanity: we must acknowledge the possibility that it may be harmful.
The spectre of an AI that harms humanity is indeed present in many discussions about the future of this discipline. If some scholars see AI as the way for humanity to overcome the natural decay of the body and live forever in digital form in the Internet, or even in an embodied robotic form [2; 3; 4], others foresee a future in which the very existence of the human race will be put at risk by machines that will be both stronger and more intelligent than those who built them [5; 6; 7; 8].
Between the extreme scenarios of the promise of an eternal digital or robotic life and of the extinction of the human species, we find today's AI: self-driving cars [9], package-delivering drones [10], completely automated investment funds [11], just to name a few examples. Each of these research projects, whether completed or still under development, is accompanied by a series of questions, including important ethical issues. Who is responsible when accidents occur with self-driving cars? [12] If drones are capable of both transporting medicines and dropping bombs, are we giving

* This is a translation of the paper "Saremo padroni o schiavi dell'informatica del futuro?" published in *Mondo Digitale*, 72, pp.26, November 2017.

powers of life and death to entities without morals? [13; 14] Will humans be no longer able to follow a financial market evolving at the speed of electronic computations? [15]

These doubts arise from a simple basic idea: AI research aims to create artefacts to which we delegate activities traditionally performed by human beings (Box 1 offers a quick overview of various research topics in AI). Many questions, therefore, revolve around the following: given activity *x*, what are the consequences of delegating the execution of *x* to a machine? We can consider this as one of the fundamental questions of AI, to which many researchers are trying to draw attention.

This question is not new in the history of technology: just think of the Luddites in England at the times of the industrial revolution, and their attempt to oppose the introduction of machines in factories, seen as a serious threat to human jobs. Today, however, the question arises even more, since there are more and more computer and robotic systems that permeate our daily lives, and the range of activities entrusted to them seems to expand without boundaries. Are there, indeed, boundaries? If so, are they intrinsic technological limits or are they introduced by choice of the AI scientists of the future?

In this paper, I propose some general guidelines that could contribute to the discussion on the development of AI and its impact on our lives. Making predictions is never easy, but my intent is to provide a conceptual framework that not only allows us to understand the implausibility of certain scenarios proposed by some scholars, but that also helps us face the future of this technology with more awareness.

Section 2 presents examples of a dystopian future where a very advanced AI ends up causing harm to individuals or even humanity; in Section 3 I will compare these examples with existing technologies, trying to recognize the components of AI systems that could lead to such scenarios; once familiarized with this type of analysis, in Section 4 I will apply it to an existing technology, self-driving cars, whose introduction into society is currently the subject of heated debate; Section 5 will focus on critical issues that AI already poses today; finally, Section 6 will conclude.

2. A dystopian future: Will the machines overtake us?

The following scenarios have been proposed by three different AI scholars, with the aim of making their readers aware of the enormous problems that a very advanced technology could cause if it escaped human control. Please do not be surprised if these scenarios will sound like science fiction to you: it is nevertheless worth to analyse them here, at least for two different reasons. First of all, this is an opportunity to better understand what kind of conceptual slippage many AI scholars fall into. Moreover, and this is one of the real problems of AI today, however absurd these stories may sound, unfortunately they have captured the attention of many, including extremely wealthy and influential entrepreneurs, actively involved in the development of cutting edge AI technology.

Money and power are outside the scope of this work, but you can imagine how wrong ideas supported by those who are able to influence political decisions can lead to negative consequences for society. I will go back to the real problems of AI in the following sections. Allow me, for the moment, to linger in science fiction.

The chessboard killer

In describing the risks of advanced AI, American physicist Stephen Omohundro presents a scenario in which technology built for a very specific goal ends up harming people in the most disparate ways. Such goal, in Omohundro's example, is to play chess [16]. The scholar imagines an advanced version of IBM's Deep Blue computer system, which in 2005 beat the then world chess champion Garry Kasparov. The difference lies in the fact that this advanced AI is not simply a computer that plays chess, but a robot that does everything to continue playing (and winning). When, after a great number of games, the human player gets tired and wants to turn the robot off, "because nothing in the simple chess utility function gives a negative weight to murder, the seemingly harmless chess robot will become a killer out of the drive for self-protection [16, p. 15]." Moreover, "the chess robot (...) would benefit from additional money for buying chess books (...) it will therefore develop subgoals to acquire more computational power and money. The seemingly harmless chess goal therefore motivates harmful activities such as breaking into computers and robbing banks [ibid. p.16]."

Happiness at all costs

Latvian computer scientist Roman Yampolskiy imagines a super-intelligent machine of the future created with the directive "to make all people happy [17, p.131]." Since the machine is advanced AI technology, it only needs to receive a directive: the rest, that is the way to achieve the objective, will be taken care of by the machine itself. Yampolskiy provides us with a (non-exhaustive) list of how things can go wrong with an escalation that culminates with the extinction of mankind. The super-intelligent machine could make humanity "happy" with a daily dose of ecstasy; it could put a permanent smile on all faces by means of surgery carried out by robots or, to remain in the context of surgery, by means of lobotomies to send the people's minds into a state of happy dementia. The machine could even apply logic in in a pedantic way, and transform the sentence "all people are happy" into the formalised conditional "for every x, if x is a person, then x is happy". The goal for which the machine was built is to make this sentence true. Unfortunately for humanity, the machine could make the conditional trivially true by eliminating all people: since there is no person, it is true that all people are happy.

The dominion of the superintelligent machine

Nick Bostrom, a Swedish professor of philosophy at Oxford and founder of the Future of Life Institute (where the abovementioned open letter initiative kicked off) devotes an entire book to the possibility that machines will take over human beings. In "Superintelligence" [18], Bostrom focuses on a particularly critical moment in the future, when AI will not only improve on current technology, but it will improve its very capability to improve, triggering a cascade effect culminating with the birth of a machine whose intelligence is not even understandable by a human being, a "superintelligence" indeed. Endowed with an immensely vast knowledge, this developing AI "knows" well that if humanity knew about this process they would do everything possible to stop it, for example shutting down all the computers on the planet. For this reason, at least in the beginning, there will be a phase of covert preparation, during which the AI will continue to improve hidden from any potential human witness. The AI will elaborate plans to achieve its long-term goals, and since it is able to improve itself, at every step of this evolution these plans will get better, with more chances of success. When the AI is powerful enough, it will no longer need to

hide, and it will come out and launch an attack on humanity. At this point, according to Bostrom, the AI will be a completely "autonomous" technology, out of the control of human beings, acquiring objectives without any guidance from its original creators, and able to find the resources necessary to pursue them. Faced with such superintelligence, humanity will become totally irrelevant at best, if not enslaved or, in the worst case, it will be annihilated.

Let's go back to reality. First of all, readers must be made aware that not all AI researchers devote their time to apocalyptic scenarios in which human beings are overwhelmed by machines that not even the experts in the field can fathom: these futurologists constitute only a minimal part of the AI community. However, their impact on society is far from minimal: a previous version of Omohundro's article (the one about the killer chess player) appeared in the "Journal of Experimental & Theoretical Artificial Intelligence" in 2014, and is still the most downloaded article in the history of this publication [19]; Bostrom's book was defined by the American computer magnate Bill Gates as one of the two books[†] anyone who really wants to understand AI should read [20]. Another great admirer of Bostrom's work is South African inventor and entrepreneur Elon Musk, according to whom AI is a threat to human existence just like, if not worse than nuclear weapons, and it must be controlled at all costs [21].

There seems to be a lot of confusion, both within the discipline of AI and also on the outside, in the broader context of human society, with its intricate social, political, and economic relations. In the latter case, Musk himself causes some perplexity. If the entrepreneur is so worried about AI, why did he invest in DeepMind, the leading British company in the field of machine learning, acquired by Google in 2014 and responsible for the recent AI victories over humans in the game of Go? If Musk fears that machines are a threat to people, what is his position relative to the autopilot car of his own company Tesla, involved in a fatal accident in May 2016 [22]? The Tesla case points straight to the most problematic issues of today's AI, also because, unlike fantasies about killer robots, it has caused a real death. However, first I must clarify the flawed conceptual framework of some kind of AI from which the abovementioned scary futuristic stories emerge, in order to shed light on the real critical issues of this type of technology.

3. A closer look: How does artificial intelligence really work?

All futuristic scenarios of AI (including novels and Sci-Fi movies) have this in common: the machines act in an unexpected way, and their actions are harmful to human beings. In the case of the chess player, the goal is harmless and well specified, but the lack of limitations on executable operations lead to robberies, breaking and entering, murder; in the case of happiness at all costs, the lack of precision in the description of the objective of the machine leads to the extinction of humanity; with a superintelligent machine, however, human beings have no say in the determination of the objectives that the machine should pursue.

[†] The other book, "The Master Algorithm" by Pedro Domingos, is an essay on machine learning based on the assumption that an ultimate algorithm exists with which it is possible to program computers to learn all of human knowledge and beyond.

In all these fantasies, the problem lies in the high degree of autonomy of the machines, but can we really talk about autonomy of the machines? We must not forget that, however vast the applications of AI are today (they range from driving to financial markets), they always boil down to programs running on digital electronic computers based on an architectural paradigm dating back to the 1940s [23]: the operations that a computer performs are calculations done by an arithmetic-logical unit (ALU), which applies commands coming from a central memory to data also coming from that memory; this transfer of commands and data from the memory to the arithmetic-logic unit is managed by a control unit (CU), and is determined by other commands present in the central memory. In other words, nothing happens in a computer that is not written in the central memory, and the type of operations that can be performed is determined by the nature of the ALU, that is, they are manipulations of binary electrical signals (low voltage and high voltage) that we interpret as digits (0 and 1) in arithmetic operations and as truth values (true and false) in logical operations.
"Is that all?" asks the Sci-Fi enthusiast without much knowledge on Computer Science. Yes and no: if on the one side a "direct" observation (actually mediated by adequate optical and electronic tools) of the workings inside a computer shows the limits of the field of action of the tool, on the other side it also makes us appreciate the vastness of its applications. Such versatility was made possible by the work of genius of a great number of mathematicians, physicists and engineers who in the second half of the 20th century managed to map entities of various types (texts, images, sounds, movies, etc.) and the relevant manipulation onto arithmetic operations that can be performed by a computer [24]. However, we should not let the versatility of a computer fool us: a computer is not able to escape from its intrinsic determinism.

3.1 The concept of autonomy in Artificial Intelligence

Is it possible to reconcile reality with the imagination of some AI futurologists? Did they simply draw on Sci-Fi stories?[‡] I believe that some AI researchers from the 1990s are at least in part responsible for these blurred lines between science and fiction, in particular, those who initiated a new line of research, dealing with "software agents". It is then that the term "autonomy" started being used with too much liberty, leading many to confuse a high degree of machine automation with the autonomy that characterizes human actions. Let us focus on a definition from one of the most famous works on agents by American researcher Patti Maes in 1994: "An agent is called autonomous if it operates completely autonomously, i.e. if it decides itself how to relate its sensor data to motor commands in such away that its goals are attended to successfully [25]." Apart from the problem of circularity (an agent is autonomous if it operates autonomously), this definition is so vague that it leaves a lot to the reader's imagination: how does a computational system act when acting autonomously? Maybe if we prevent it from playing chess, it will try to kill us. Of course I am not trying to imply that shaky conceptual definitions will lead us to extinction. After all, one can forgive the lack of a well-defined conceptual framework in a pioneering work. However, I suspect that an improper use of terms may have

[‡] The primacy, at least chronologically, belongs to the 1920 theatrical piece "RUR" ("Rossum's Universal Robots") written by the Czech playwright Karel Čapek. The title uses for the first time the term "robot" ("robota" in Czech means "corvee").

contributed to the lack of conceptual clarity that has led a number of AI researchers to concoct scenarios that are incompatible with the true nature of the technological tools at their disposal.

Here is a better outlined definition of the concept of autonomy in AI, provided by Mike Wooldridge and Nick Jennings, British academics specialized in agent technology: "Autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state [26]." What I appreciate the most about this definition is that it mentions a "direct intervention" and it implicitly implies the idea of an indirect intervention by human beings. I claim that *indirectness* is the key concept underlying automation in Information Technology (IT) and, by extension, the so-called "autonomy" in AI.

Here follows an example of direct human intervention: a user interacts with a program running on a computer by entering data through the keyboard and clicking on several points of the screen with a pointing device. The user has the impression that she is providing the computer with a series of orders that are punctually executed by the machine: for example, by clicking on the "send" button of the e-mail program, an e-mail is actually sent to the recipient. The user seems to have total control, but that is not the case: the user only provides the parameters of an interaction that has already been conceived and programmed by the creators of the software. Here is the indirect intervention I was talking about earlier: whoever creates a program writes a sequence of instructions intended to be stored inside a computer device (typically, in a non-volatile memory device such as the hard-disk), which is then recalled in the central memory to be performed when the user decides to activate the program by clicking on a computer icon or, more and more frequently, by tapping on the touch screen of a smartphone.

Analysing an e-mail program from the point of view of its creator is an excellent exercise to understand how determinism in IT tools still leaves room for unpredictability: it is true that the set of instructions that make up the software is once and for all established at the time of the release of the program, and it is also true that the programmer has precisely defined all the operations that the user will be able to do with the program (write a message, add a recipient to the address book, save a draft, etc.), but given the parametric nature of the software (messages, recipients, drafts are all parameters set by the user), its actual operations at the time of its use (in Computer Science terms: "at run time") cannot be determined when its instructions are still written and controlled by the programmer ("at compile time").

Do these considerations make the e-mail program look like an "unpredictable" software in the sense of the fantasies of AI futurologists? Should we expect deadly emails? Of course not: even if the programmer has no idea of the content of the messages or the number of recipients that will be managed by the software, they know that the operations performed by the user will remain within the set of instructions specified in the program at compile time.[§]

Let us now focus on the second part of the definition of "autonomy" by Wooldridge and Jennings, which speaks of systems that "have some kind of control over their actions and internal state". Clearly they are not referring to an e-mail program: its actions and its internal state are entirely determined by the programmer and by the user. However, there are cases in which the user cannot be present at the time the program is executed and it is not possible to establish in advance all the parameters to

[§] Here I focus on software, but of course the same considerations apply to hardware, the physical part of the computer. Those who have built the hardware have no idea on how the computer will be used, but they know that any future operation will remain necessarily limited to the arithmetic-logical scope manageable by the processor.

be used during such execution, due to contingent factors that are difficult to predict (as it often happens in a "complex dynamic environment", as already mentioned by Maes). If the outcome of a mission that costs billions of euros depends on a correct execution of the program, then the concept of "autonomy" is invested with a whole new criticality.

3.2 Artificial Intelligence in space

I am referring to space missions, where robots must explore unknown territories under unpredictable circumstances and transmit the collected data to the mission control centre on Earth. When it comes to the complexity of the explored environment, the most daring mission was certainly "Rosetta", led by the European space agency ESA in collaboration with the American NASA and JPL, which aimed to better understand the nature of comets and to do so, it sent a "orbiter" module (named Rosetta) around a comet (the 67P/Churyumov-Gerasimenko), from which a "lander" module (named Philae) landed on the comet itself [27]. The great distance between mission control and the robot meant that not only neither the programmers nor the users of the Rosetta computer system could be present at the time of its use, but this also prevented real-time sending of commands in response to contingencies. It is noteworthy that exploring a comet involves many more unexpected conditions than the exploration of a satellite like the Moon or a planet like Mars: the orbit of a comet around the Sun is much more eccentric, hence the comet has a very different environment when it is close to the Sun than when it is far. Moreover, given the relatively small size of the comet, its gravitational field is not strong enough to be taken as a stable reference for calculations: at the point of maximum proximity to the Sun, the sudden release of gases by the melted ice are a much more significant force than the gravity of the comet itself. There is no way of predicting in a deterministic way, that is, in a way that is compatible with the programming of a computer, in which direction and with how much force the next geyser will blow.

To complicate things, resources are limited: for reasons of manoeuvrability, the Philae lander could not be loaded with too many batteries, which naturally entailed the need for an optimized use of the energy necessary to take photographs of the surface of the comet and send them to the Rosetta orbiter for as long as possible. To overcome the limitation of the batteries, Philae was equipped with solar panels, but their orientation was another parameter not programmable in advance, since the surface of the comet was not known in detail before the start of the mission and the lander could not be expected to land on the surface exactly according to plans.

How to successfully bring a mission to completion with so many unpredictable factors? If human beings are not able to intervene, then the computer system must be able to act in "autonomy" to achieve the goal for which it was built. Since a computer is a deterministic system that does not do anything that is not in its memory, programmers had to do their best to provide the system with the most possible flexibility. A traditional way to proceed in Computer Science is to write conditional statements: if a certain condition *c* occurs then the computer executes the operation *o*. Of course, if the condition refers to data inside the computer, checking this condition is very simple, but if it refers to a contingency in the surrounding environment, the computer system must be equipped with the sensors necessary to detect the relevant phenomenon and translate it into numerical data that can be processed by the computer. For example, if the instruction is "if the outdoor temperature exceeds 50°C,

activate the fans", then the system must be equipped with a thermometer and an apparatus that describes the state of the thermometer in numerical terms (in short, a digital thermometer). Of course, if the operation is not a simple calculation that can be performed by the arithmetic-logical unit, but involves physical actions in the environment in which the system is located, the system must be equipped with the necessary actuators, i.e. devices that are driven by electrical impulses sent to them by the computer's control unit, in accordance with the executed instruction. The most common actuators in AI are the wheels many systems are equipped to move around in the environment. Typically, the term "robot" refers to computers equipped with actuators. The Philae lander is equipped with many types of actuators: a turbine to prevent overturning during the descent on the comet, arms with screws to cling to the ground at the time of landing, solar panels, a photographic apparatus, and so on.

If a robot is equipped with a program with conditional instructions and its sensors and actuators function correctly, it will be able to deal with the contingencies (at least those foreseen within its program) and respond adequately to achieve the goal for which it is been built.

Sometimes, however, goals are specified at a high level of abstraction, i.e. they are not directly translatable in terms of sequences of conditional statements. In the case of Philae, one of its daily goals was to send the largest possible number of photographs taken on the surface of the comet to the Rosetta module in orbit, from which the photographs would then be transmitted to mission control on Earth. This is a complex problem, because many factors are involved: the limited memory of Philae for saving photographs, its limited batteries, the management of the solar panels, the light conditions determined by the rotation of the comet and its orbit around the sun, the position of the Rosetta module to which the photos must be transmitted, and more. What is the best course of action to maximize the number of images transmitted while minimizing energy consumption? This is a planning problem: find the right combination of operations to perform to achieve the goal.

This task is not about understanding the conditions in which to perform a certain operation or not, but to calculate the sequence of instructions to be executed, i.e. to create a plan. A lot of information is needed to do so: not only you need to know all the possible instructions, not only you need to know the "local" result of each of these instructions, but you also need to know how to compute the "global" result according to the order by which the instructions are executed. Let's do some math in a purely combinatorial way: if the robot has at each step N different operations to choose from, and a possible plan consists of M steps, in theory, there are N^M different possible plans. In figures, if a robot is able to perform 100 different operations and a typical action plan consists of 100 executions, the number of possible plans would be 100^{100}. This would mean that even if the robot's computer were able to check the feasibility of a single plan in a billionth of a second, to control them all and choose the best one would need a number of millennia expressed with a 2 followed by 84 zeros.

The problem of planning therefore comes with a problem of finding the best solution and this search can be made much faster by pruning the possible searches by means of common sense criteria (for example, excluding a priori those plans that send photos before the photos are taken). However, it is a search that requires considerable computing power and, in the case of the photographs made by Philae, the action plan was processed day by day by the mission control supercomputer and then sent to the module on the comet to be executed [28].

Consider the combination of the Philae robot and the mission control supercomputer: it is undoubtedly an IT system with a higher degree of automation than the e-mail

program installed on your computer. There are obvious differences that make the two systems difficult to compare (sending an e-mail and taking pictures on a comet are very different activities), but if we abstract from the content of their objectives and we focus on the "control over their actions" that these two systems have, we can make non-trivial distinctions that have a more general value in the context of AI.
The operations of the e-mail program are part of a set pre-established by the programmer, and are executed when the order is given by the user through a keyboard or other device. The operations of Philae, however, are not performed when a command is given by mission control, also because this would be impossible given the distance and the related transmission issues: their execution depends on a plan established by the supercomputer that presides over the operations of Philae, and no mission control scientist is able to predict exactly the moments of the day when this execution will take place. This impossibility is not due to the fact that these are mysterious operations: the operations are well known and are those that were established during the design of Philae. The reality is that the calculations of the supercomputer to establish the moment of execution during the day of the robot are performed at such a speed that, for a human being to get to the same results manually, it would take such a long time that Philae would stay idle on the comet until the complete depletion of its batteries.
From this point of view, human beings give part of their control away to the machines, entrusting them with the timely elaboration of an action plan for the success of the mission. This control transfer is only partial: the set of operations of Philae are those established by its manufacturers and programmers, the optimization criteria in the search for the best action plan by the supercomputer are dictated by the common sense of the programmers and, above all, the goal that Philae pursues with its operations is the one originally established by ESA.

3.3 The unpredictability of Artificial Intelligence

After having met one of the most advanced computer systems ever created, able to carry out one of the most complex space missions in the history of mankind, let us recall the stories by some AI futurologists, and ask ourselves whether technological progress in this field could ever lead to the catastrophic situations described by them.
I can only assume that the futurologists have followed this line of reasoning: advances in the field of AI allow for more and more complex programs to be written, with an ever increasing degree of automation; there are already programs today (such as the Philae planner) whose operations cannot be controlled directly by human beings, who merely establish high-level objectives, entrusting the machines with the creation of the course of action aimed at achieving such goals; at the moment human beings write these programs, inserting the criteria derived from their experience to improve the programs' performance, for example by steering the search for optimal solutions towards the most promising directions, but soon the machines themselves will learn to exploit these criteria and human intervention will be less and less needed; at a certain point, people will only have to specify the objectives and the machines will "think" of the rest. If there is only one thing left for human beings, it is easy to imagine the last step in this development of AI: to pick the goals. Once the machines are able to do that, what are human beings needed for? Their elimination would seem to be a logical result from the "point of view" of the machines. Somewhere in this discourse we have jumped from the reality of the most advanced AI to the fantasies of the futurologists.

When did this happen? As I have already mentioned, the problem lies in confusing the high degree of automation of a machine with the autonomy that characterizes human beings. Even if humans give away more and more control on the elaboration of the action plans, the operations that a machine is able to perform are always determined and limited by its hardware and software. If a robot is equipped with wheels to move and a program to control these wheels, depending on how this program is written, the robot may be able to perform even very sophisticated movements, which could amaze the robot manufacturers themselves. This amazement, however, derives simply from the fact that the builders did not initially realize that certain combinations of movements could give rise to the results before their eyes: we are far from the amazement of the characters of science fiction movies when they realize that the robots that were supposed to help them are going to kill them (e.g. think of "2001: A Space Odyssey", "Terminator 2", "The Matrix", "Ex Machina").

Every time you see or read about a robot which is about to kill a person, you have to remember how computer systems work: if the robot performs a certain operation, it means that this operation is described as an instruction inside its central memory, and the robot is equipped with the sensors and actuators necessary to complete this operation. If Omohundro's chess player grabs a knife to kill the person who is about to shut it down, there must be an elaborate action plan in its memory, exactly as in Philae's memory there must be a sequence of operations to manage a day of photography on the comet, and as Philae is equipped with photographic equipment to take pictures, so the chess player must be equipped with sensors to locate the knife in the room and the position of the victim and actuators to approach the knife, grab it, move quickly toward the person, stab them, etc. Who wrote these instructions in the memory of the killer chess player? The AI futurists, focusing too much on the concept of "control over their actions", have forgotten that this control by the machine is related to the chronological order of the execution of its actions, a control that is very limited compared to that of a person who has, during their life, learned to handle a great variety of actions, including the use of a knife. In the case of the robot, the use of a knife must be described in terms of instructions in its memory. A robot programmed to play chess will simply manipulate the pieces on the board. Not even the most extreme futurologist would deny that this is how today's robots work. The point of disagreement is what will happen with the robots of the future: according to some researchers, the hardware of future AI will be so evolved to free the AI from the aforementioned architectural paradigm of the instructions in the central memory, and when that happens, the robots will begin to explore the world with their sensors and actuators in a way similar to that of a child, learning an increasing number of concepts and actions.

We will have to deal with entities capable of interacting with people in a seemingly intelligent way (according to weak AI, which rejects the idea that artificial systems can entertain phenomena comparable to human consciousness) or intelligent *tout court* (according to strong AI, which claims that a computer can become conscious if properly designed and built). Faced with this new species of higher entities (at least from the point of view of computational power), the destiny of humanity will be at a crucial point. Nothing at the moment suggests that such a technological evolution is possible: for the time being, IT systems behave exactly as they are built and programmed, and the only surprises happen when programmers are not able to pre-compute all the possible results of the programs they themselves write. When a program you are using crashes, it is because determinism still applies to every aspect of computing, including AI: if the system is in an environmental condition c, and its

software does not include an instruction that say which operation is to perform in case of *c*, the system does nothing.
Please beware: this does not mean that it is impossible to have a killer robot. A malicious programmer could equip the chess player robot with the instructions and the apparatus needed to behave as described by Omohundro. Indeed, there are already robots equipped with machine guns that act as automatic sentinels on the border between the two Koreas [29]. With apt sensors and actuators, and correct instructions, a robot can be built to automate a very large number of activities. The problem lies in the possibility of writing these instructions. Let's not forget that, inside the robots, they boil down to manipulations of digital signals, so whatever the context of the problem we want to solve, we must make sure that we have a numerical model of the factors involved. Therefore, we cannot expect IT to provide a solution to problems that we are not able to express in numerical terms, such as human rights, religious issues, psychology, etc.
On the other hand, street driving is apparently a problem that can indeed be expressed in numerical terms, since more and more companies are proposing to entrust it to robots on four wheels.

4. AI on the road: Do self-driving cars really work?

The prefix "auto-" in the word "automobile" clearly shows how the idea of automation has been present since the beginning of the history of this technological artefact. In recent years the huge investments of companies like Google, Nissan, BMW (to name just a few) have drawn the attention of the general public towards the design of self-driving cars. However, the transfer of (part of the) control from the human driver to the machine is nothing new: just think of the introduction, dating back to the 1980s, of the ABS (Anti-lock Braking System), which relieves the driver from the need to press on the brake pedal intermittently while braking on wet or icy roads to avoid wheel locking. As in the case of the killer chess player and the Philae robot, these innovations depend on the addition of sensors, actuators, and relevant instructions in the on-board computer program.
There was a significant technological leap in the first decade of the new millennium thanks to researchers of the Stanford University in California who, under the guidance of German professor Sebastian Thrun, distinguished themselves in a competition organized by DARPA (Defense Advanced Research Projects Agency, an agency of the US Department of Defense) for cars that had to drive themselves through the Mojave desert, also in California. Their successes have attracted the attention of Google, who hired Thrun and his team to develop the design of a self-driving car. This transfer of know-how has not gone unnoticed, and more and more companies are now convinced that self-driving cars are not only an intellectual experiment for academics, but also a successful technological and commercial investment.
Unlike university research, studies and experiments conducted within a company are characterized by a high degree of confidentiality against industrial espionage, so we know much less of the technologies used in the Google car (or Waymo car, named after Google's division dedicated to this research) compared to what was published by the same people when they were still working for Stanford University. However, we can nevertheless get a good idea on the state of the art of AI in automatic transport based on academic publications, on statements and white papers by the company

itself, and on what can be observed directly on the models made available to the public for demonstration purposes.

At very foundation of the self-driving car industry we find the "lidar" (a term coined from the combination of "light" and "radar"), a device that emits laser pulses towards the environment and, by means of sensors, receives them back as they get reflected by objects in the surroundings. With the data obtained from the lidars mounted on the car's body (Google's model has 64 lidars), the car's computer builds a three-dimensional image of the environment in which the vehicle is moving, including other cars, motorcycles, bicycles, pedestrians, traffic lights, buildings, and so on, and computes, in accordance with the instructions written by the programmers, the best trajectory to follow and the speed to keep in order to proceed towards the destination smoothly. The accuracy of this detection system seems to have reached such a level of detail that, according to Google's company website, their car is able to perceive that a cyclist in the vicinity of the vehicle has raised an arm to indicate their intention to turn.

Since the software in the system contains the appropriate instructions, the car that "perceives" the intentions of a cyclist will slow down to leave room for the manoeuvre. All good, then? Yes, in that case. How many other cases, however, must a driver deal with on the road? Would you be able to make an exhaustive list of all the possible situations that you have to manage behind the wheel and pair them with the proper instructions so that everyone can come out unscathed and continue to their destination? This is a far-from-trivial undertaking. Even with the additional data coming from maps preloaded in the computer (with all the indications on intersections, traffic lights, stop signs, one-way traffic, etc.) and with the support of the GPS satellite positioning system, the length of the case list does not change and there are no shortcuts or optimizations: the environment in which a driver moves contains a large number of objects and the variables that must be dealt with are hardly predictable. Google's fleet can boast 3 million miles of on-the-road experience, thanks to experimental drives in the US states that have gave the company permission (California since 2009, Texas since 2015, Arizona and Washington state since 2016). However, not all of these miles were driven "autonomously": there were moments of "disengagement", when the human driver had to take control of the vehicle to handle a situation that was unexpected or too complex to be managed by the on-board computer. Google states that the constant improvement of their software decreased the number of disengagements from 0.8 per 1000 miles in 2015 to 0.2 in 2016 [30]. These figures may look negligible, but this means that in over 3 million miles travelled, human intervention was necessary more than a thousand times. In other words, there have been at least 1,000 cases in which, had the human being on board not intervened, there would have been an accident.

This is probably what happened in the previously mentioned fatal accident with a Tesla. If Google uses its vehicles in an experimental way with its researchers behind the wheel, Elon Musk's Tesla has already marketed a car enhanced with an assistance system called "Autopilot" that can be used in the simplest driving situations (typically on highways, where there are no intersections or traffic lights). The accident happened on May 7, 2016 on a highway in Florida, when a truck steered to the left in front of a Tesla that was in "Autopilot" mode that did not brake, causing a collision. The only person on the Tesla, 40-year-old Joshua Brown, died in the crash. In a first report by the NHTSA (National Highway Traffic Safety Administration), it was assumed that the driver or the autopilot had not noticed the light-coloured side of the truck because it did not have sufficient contrast against a very bright sky in a

particularly sunny day. Shortly thereafter, Tesla issued a statement, according to which the driver must manually activate "Autopilot" and each activation is accompanied by an audio notification that advises the driver to always pay attention to the road and never release the steering wheel. Moreover, the sensory technology adopted on Tesla vehicles, supplied by the Israeli company Mobileye, is able to warn drivers of rear-end collision risks, even to activate emergency brakes if needed. However, cases involving vehicles that come from the side, like the one of the accident, are not managed by the computer system. The NHTSA investigations ended in January 2017, with a full acquittal of Tesla, as all Tesla customers must read and sign the recommendations for a correct use of "Autopilot" at moment of purchase [31]. Still, there is a precedent related to Joshua Brown, a car enthusiast and, in particular, a Tesla fan: a few months earlier, he had published a video on Twitter showing his car in "Autopilot" mode make a quick steer to the right to avoid a collision with a truck coming from the left with a hazardous lane change manoeuvre. The video was intended to showcase how effective the "Autopilot" of his Tesla was and when Musk himself published a link to Brown's video on his own Twitter page, Brown said he was overjoyed, as shown in a segment of American news programme "Inside Edition" [32]. In the video (a link is in the references of this article), at 1 minute and 30 seconds, we can see Brown's Tesla avoid the truck. Look at it carefully and then answer the following question. Does "Autopilot" react to the presence of the truck by following the instructions to avoid collisions in the direction of travel (a feature officially recognized by Tesla as part of the system's capabilities)? Or has some other operation come into play, such as a steering to avoid objects on the road? One thing is sure: a vehicle with "Autopilot" has not been designed to handle insertions from the sides. Yet, the successful avoidance shown in the video might have given Brown this (false) impression, certainly reinforced by Musk's approval on Twitter. I believe that this episode further increased Brown's already great trust in "Autopilot", and led the driver to riskier behaviours, such as letting the steering wheel go and get distracted on the highway. There are also rumours, reported by some witnesses in the "Inside Edition" video, that Brown was watching a movie on a portable DVD player when the accident occurred, but no evidence was found to support this thesis. This fatality is one of many cases of imprudent behaviour of people in the presence of highly automated systems: they call it "automation bias", a partiality of human beings towards such systems, whose reliability is being questioned less and less frequently.

5. Where is the real danger?

Allow me a very banal example: you are at dinner in a restaurant with friends and you all decide to split the bill equally. How much is 357 divided by 11? One of your friends does some math and says 34, while another, using the calculator on her smartphone, says 32.45. Who do you believe? Of course, the friend with the calculator, but ask yourself on what basis you make this choice. A human being is prone to calculation errors, while a calculator or a computer cannot make mistakes: they have been built precisely for this purpose, and their circuits "embody" the laws of arithmetic. Actually there was quite an extraordinary case in the 1990s that reminded us that even the electronic circuits inside computers are, like all technological artefacts, designed and built by humans, the same humans who make mistakes after a dinner with friends. In June 1994, lecturer Thomas Nicely of the

Lynchburg College in Virginia noticed that once a new computer containing the Intel Pentium processor was added to the series of machines he was using for his experiments with prime numbers, the system began to give results that were not consistent with the mathematical theory. Nicely took months to isolate the various factors that could have been the cause of these errors, but in the end it was clear that they did not depend on an error in the program written by him: it was the processor of the last computer added to the system to make some divisions incorrectly [33]. Since this was a design flaw at a specific position in the electronic circuitry, only the divisions of particular sequences of digits involved the faulty part and therefore led to wrong results. Intel declared that the average user would receive incorrect results from the processor once every 27,000 years, whereas according to IBM (then a competitor of Intel in the industry of processors) the error would occur every 24 days of normal use of the computer. Under pressure from public opinion, Intel recalled the faulty processors in December 1994, with a loss estimated at around $475 million of the time.

That hardware built to perform calculations contains a defect is a very rare event in the history of Computer Science but, as we have seen, it is not impossible. Add to this type of problem the much more frequent software faults, that is, the fact that a deterministic computer system is in a situation not foreseen by its code and hence stops working, possibly putting at risk the lives of the people depending on that system. Moreover, we have seen how certain careless people can get used to the use of highly automated systems, to the point of abusing them, and expecting them to be able to carry out even those operations for which they have not been designed.

What do all these situations have in common? You must not focus your attention on computer technology, but widen the view to see that this technology is conceived, built and used by human beings within a socio-political-cultural context often overlooked when it comes to computers and AI. Futurologists who fear that robots will learn to use weapons to exterminate us seem to forget the fact that there are multinationals hiring experts in robotics and planning to build armies of automated sentinels. Entrepreneurs who exploit legislative gaps to market cars endowed with an extremely sophisticated but not perfect driving support system seem to ignore the existence of reckless drivers who will do anything to gain “likes” on social networks (e.g. pretending to be asleep on the back seat of a driverless car) [34].

Whole industries, such as civil aviation, continue to insist on increasing automation in their artefacts, although more and more experiments show how the performance of human personnel decreases in quality as IT systems increase in their activities. Why is this happening? Why is a reduction in human capacity considered an acceptable consequence of the technological innovation of an entire industrial sector? This is a matter of numbers: not the numbers processed by a computer, but economics and statistics. The autopilot technology on airplanes has reached such a level of development that, in an average flight, human pilots must only maintain control of the aircraft for a few minutes, at take-off and landing. This means that it is required less and less of the human pilots, that it is easier to train them, and, at the same time, companies are able to train more pilots and need fewer of them in the cockpit. 60 years ago every flight was managed by 5 well-paid professionals, while today there are only two people in the cabin, whose salary has been on a constant decline in recent years. Statistics do not lie: it is undeniable that there are fewer aircraft accidents than in the past. Since there is less room for manoeuvre by people, the chances of human error have decreased. Naturally, all works as long as the automatic system that manages the aircraft does not contain hardware or software defects.

However, it is clear that, due to the lack of experience of the people relying on the automation of the aircraft, the fatal accidents of recent years are almost all attributable to errors committed by pilots when they were forced to resume manual control in emergency situations, in which the autopilot had stopped operating [35].
Let us recap: the most advanced AI is not based on technologies other than those of basic Computer Science (it is still a matter of deterministic electronic circuits) but rather on the ingenuity of programmers who can model in mathematical terms the most varied aspects of reality (we range from highways to comets) and physicists and engineers who can equip the computers with sensors and actuators necessary to perceive and modify the surrounding environment. Despite the bold predictions of some researchers who, mixing up machine automation with human autonomy, imagine that machines will one day make decisions exactly as humans do, at the moment it is indeed humans who decide what goals to pursue through AI systems and decide to design and build such systems. Humans are not perfect, and not only because they cannot perform calculations quickly and correctly all the time, but also because they can build faulty AI systems, or systems that work perfectly but with morally questionable objectives, or systems with noble objectives, but that make their human users less and less qualified in dealing with emergencies, when the AI for some reason fails.
Actually, even before the advent of AI, humanity had already lost many skills in managing situations without the help of technology: think about how easy it is to go to a supermarket and buy a packet of pasta, and try to imagine having to grow wheat to feed yourself and your family. AI technology, however, is different from more traditional artefacts because, at least in the intentions of its designers, it aims at enhancing and possibly replacing what we characterize as typically human: our intelligence, our intentions, our thought. Even if some futurologists may disagree, there are still immense leaps to do in our scientific knowledge for this feared substitution to take place: for example, we still know very little about how our brain works, and we already delude ourselves into believing that we can build an artificial one. This is not the real problem: I'm not afraid that future generations will be enslaved by robots. However, the confusion between automation and autonomy has become part of the discourse on technology, and more and more often I hear or read opinions of experts in the field that pave the way for a very dangerous future, where responsibility of people who make self-interested choices and impose questionable technologies could get lost in the seemingly excessive complexity of the new AI systems that will surround us.

6. Conclusion

Despite numerous broken promises, AI has made great strides in the decades that followed its beginnings. However, the extraordinary results obtained in very specific sectors are increasingly the subject of gross generalizations and misleading metaphors. Journalists like American editor Will Knight write about a new version of Microsoft's "Minecraft" game as "a testing ground for human-AI collaboration" [36]; scholars like Japanese philosopher Minao Kukita propose to rethink the concept of responsibility in front of complex systems such as future self-driving, because "it would be not only useless but also costly to search for some individuals who is to blame when the accident happens due mainly to actions of a complex artificial autonomous system or to interactions among such systems" [37]; jurists like

American lawyer Shawn Bayern propose an analogy between the executive power of legal agreements on legal entities and that of algorithms on AI "autonomous" systems and believe that "autonomous systems may end up being able, at least, to emulate many of the private–law rights of legal persons" [38]. In all these proposals we find a drift towards a way of conceiving and treating AI that does not reflect its (real) nature of software written by people working on hardware built by people, but surrenders to the simplifying power of a metaphor that depicts AI as an independent entity.
The only recommendation that I would like to give here is to not give in to this temptation: no matter how complex AI systems are (and the level of complexity will only increase in the future), always remember that they are artefacts built by human beings for very specific purposes, and there will always be a way of tracing the choices made by these people to assign responsibility in case of accidents. It is essential that this link between the consequences of a technology and the people who conceived, designed, implemented and deployed it is always evident: my hope is that the link can act as a deterrent and guide towards the development of AI that is truly at the service of all and not just for the benefit of a few.

References

[1] Future of Life Institute (2015). "Research priorities for robust and beneficial artificial intelligence", futureoflife.org/ai-open-letter/ (last accessed January 2018).

[2] Itskov, D. (2016). "2045 Strategic Social Initiative", 2045.com (last accessed January 2018).

[3] Minski, M. (2013). "Dr. Marvin Minsky – Facing the Future", www.youtube.com/watch?v=w9sujY8Xjro (last accessed January 2018).

[4] Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology, Penguin Books.

[5] Barrat, J. (2013). Our Final Invention: Artificial Intelligence and the End of the Human Era, Thomas Dunne Books.

[6] Ford, M. (2016). The Rise of the Robots: Technology and the Threat of Mass Unemployment, Oneworld Publications.

[7] Storm, D. (2015). "Steve Wozniak on AI: Will we be pets or mere ants to be squashed our robot overlords?", Computerworld, 25 March 2015, www.computerworld.com/article/2901679/steve-wozniak-on-ai-will-we-be-pets-or-mere-ants-to-be-squashed-our-robot-overlords.html (last accessed January 2018).

[8] Gaudin, S. (2015). "Stephen Hawking fears robots could take over in 100 years", Computerworld, 14 May 2015, www.computerworld.com/article/2922442/robotics/stephen-hawking-fears-robots-could-take-over-in-100-years.html (last accessed January 2018).

[9] waymo.com (last accessed January 2018).

[10] www.amazon.com/primeair/ (last accessed January 2018).

[11] www.aidyia.com/company/ (last accessed January 2018).

[12] Hevelke, A., Nida-Rümelin, J. (2015). "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis", Science and Engineering Ethics, 21(3), 619-630.

[13] Heyns, C. (2013). "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns", United Nations Human Rights Council, session 23, 9 April 2013, www.ohchr.org/Documents/HRBodies/ HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf (last accessed January 2018).

[14] Berkowitz, R. (2014). "Drones and the Question of «The Human»", Ethics & International Affairs, 28(2), 159-169.

[15] Metz, C. (2016). "The Rise of the Artificially Intelligent Hedge Fund", Wired, 25 gennaio 2016, www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund/ (last accessed January 2018).

[16] Omohundro, S. (2016). "Autonomous Technology and the Greater Human Good" in Müller, V. (editor) Risks of Artificial Intelligence, CRC Press, 9-27.

[17] Yampolskiy, R. V. (2016). "Utility Function Security in Artificially Intelligent Agents" in Müller, V. (editor) Risks of Artificial Intelligence, CRC Press, 115-140.

[18] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies, Oxford University Press.

[19] steveomohundro.com (last accessed January 2018).

[20] Ha, T. (2016). "Bill Gates says these are the two books we should all read to understand AI", Quartz, 3 giugno 2016, qz.com/698334/bill-gates-says-these- are-the-two-books-we-should-all-read-to-understand-ai/ (last accessed January 2018).

[21] Dowd, M. (2017). "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse", Vanity Fair, April 2017, www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x (last accessed January 2018).

[22] Greenemeier, L. (2016). "Deadly Tesla Crash Exposes Confusion over Automated Driving", Scientific American, 8 July 2016, www.scientificamerican.com/article/deadly-tesla-crash-exposes-confusion-over-automated-driving/ (last accessed January 2018).

[23] Von Neumann, J. (1945). "First Draft of a Report on the EDVAC", rapporto tecnico, University of Pennsylvania.

[24] Verdicchio, M. (2016). L'Informatica per la Comunicazione, 2nd edition, Franco Angeli.

[25] Maes, P. (1994). “Modeling adaptive autonomous agents”, Artificial Life Journal, 1(1-2), 135-162.

[26] Wooldridge, M., Jennings N. R. (1995). “Agent theories, architectures, and languages: A survey” in Wooldridge, M. e Jennings, N. R. (editors) Intelligent agents, Springer, 1-22.

[27] Taylor, M. G. G. T., Altobelli, N., Buratti, B. J., Choukroun, M. (2017). “The Rosetta mission orbiter science overview: the comet phase”, Philosophical Transaction of the Royal Society A, 375, dx.doi.org/10.1098/rsta.2016.0262 (last accessed January 2018).

[28] Chien, S. (2016). “Artificial Intelligence Support of Rosetta Orbiter Science Operations”, www.youtube.com/watch?v=wcwW7dKI76g (last accessed January 2018).

[29] Welsh, S. (2017). “Clarifying the Language of Lethal Autonomy in Military Robots” in Aldinhas Ferreira, M. I., Silva Sequeira, J., Tokhi, M. O., Kadar, E., Virk, G. S. (editors) A World with Robots, Springer, 171-183.

[30] waymo.com/ontheroad/ (last accessed January 2018).

[31] Boudette, N. E. (2017). “Tesla’s Self-Driving System Cleared in Deadly Crash”, The New York Times, 19 January 2017, www.nytimes.com/2017/01/19/ business/tesla-model-s-autopilot-fatal-crash.html (last accessed January 2018).

[32] Inside Edition (2016). “Man Died Watching ‘Harry Potter’ When Self-Driving Tesla Crashed: Witness”, 5 July 2016,
www.youtube.com/watch?v=TSN3gDUNpXQ (last accessed January 2018).

[33] Cipra, B. (1995). “How Number Theory Got the Best of the Pentium Chip”, Science, 267(5195), 175.

[34] Inside Edition (2016). “See Motorists Play, Read and Relax In Self-Driving Cars As Second Tesla Crashes”, 6 July 2016, www.youtube.com/watch?v=qnZHRupjl5E (last accessed January 2018).

[35] Carr, N. (2014). The Glass Cage: Automation and Us, W. W. Norton & Company.

[36] Knight, W. (2016). “Minecraft Is a Testing Ground for Human-AI Collaboration”, MIT Technology Review, 21 July 2016, www.technologyreview.com/s/601923/minecraft-is-a-testing-ground-for-human-ai-collaboration/ (last accessed January 2018).

[37] Kukita, M. (2017). “When HAL Kills, Stop Asking Who’s to Blame”, CEPE/ ETHICOMP 2017, easychair.org/smart-program/CEPEETHICOMP2017/2017-06-05.html (last accessed January 2018).

[38] Bayern, S. (2016). “The Implications of Modern Business–Entity Law for the Regulation of Autonomous Systems”, European Journal of Risk Regulation, 7(2), 297-309.