ETHICAL CHALLENGES

The technological singularity and the transhumanist dream

Miquel Casas

Araya Peralta

In 1997, an AI beat a human world chess champion for the first time in history (it was IBM's Deep Blue playing Garry Kasparov). Fourteen years later, in 2011, IBM's Watson beat two winners of Jeopardy! (Jeopardy is a general knowledge quiz that is very popular in the United States; it demands a good command of the language). In late 2017, DeepMind's AlphaZero reached superhuman levels of play in three board games (chess, go and shogi) in just 24 hours of self-learning without any human intervention, i.e. it just played itself. Some of the people who have played against it say that the creativity of its moves make it seem more like an alien that a computer program. But despite all that, in 2019 nobody has yet designed anything that can go into a strange kitchen and fry an egg. Are our machines truly intelligent?

Successes and failures of weak AI

The fact is that today AI can solve ever more complex specific problems with a level of reliability and speed beyond our reach at an unbeatable cost, but it fails spectacularly in the face of any challenge for which it has not been programmed. On the other hand, human beings have become used to trivialising everything that can be solved by an algorithm and have learnt to value some basic human skills that we used to take for granted, like common sense, because they make us unique.

Nevertheless, over the last decade, some influential voices have been warning that our skills may not always be irreplaceable. Stephen Hawking believed that AI would outstrip humans in less than a century and Elon Musk and [1] Ray Kurzweil [20] maintain that that could happen even earlier.

Of course, this would not be the first time the experts have been wrong: Marvin Minsky, the father of AI, thought that this milestone would be passed in less than a decade (and he was writing in 1970!). That and other predictions, over-optimistic or just wrong, helped to generate a financial bubble that then burst in a crisis called the "first AI winter". So are we today about to witness the birth of a new type of intelligence as powerful as our won or laying the foundations for the nth crisis in the industry?

True intelligence: limits and horizons

Despite fresh triumphalist predictions from some public figures and some industry experts, the notion that AI is different from true intelligence is widespread in popular culture. It is often thought that we will never reach a general AI capable of performing a range of tasks equal in range to those human beings can take on. To a great extent, that is because our society has a dualist tradition that distinguishes the material (the body) from the immaterial (the mind or the soul). That distinction, however, depends on a number of metaphysical assumptions that have not been shown to be true: first, that some part of a human being is not material and, secondly, that it is impossible to have intelligence without whatever that non-material element is.

Conversely, some experts believe that we could create an artificial mind by combining technologies that are already being developed today. The human brain works differently when we interact with a three dimensional space, when we are talking or when we are making calculations but, unlike a computer, all those functions are interconnected in our brains. Thus, for example, we can describe in words the best way to throw a ball to get it in the basket and enable the reader to visualise the throw and maybe even improve their technique. From that perspective, what is essential to manage to achieve a general AI is to the find the best way to combine multiple capacities in an ordered way [3], something that today seems far from being achieved.

Other writers [4] think also that our intelligence depends exclusively on our brains, so they think that the AI developed so far has little or nothing to do with true intelligence. Among them are those who expressly maintain that the way our brains work cannot be imitated by a computer [5] . Others say that yes it probably can be, but that we are a long way from achieving it. Finally, there are those who believe that efforts so far are based on a reductionist, functional and fragmented framework and that we should be trying to imitate life more closely, as the paragon of true intelligence. A first milestone along that line of thinking would be to recreate a simpler animal, like a worm, as the *OpenWorm* project is trying to do.

The other alternative would “simply” be to emulate a whole human brain [6]. To make that a reality would require scanning a human brain and programming a computer to operate in exactly the same way as our neural connections. However, existing scanning techniques, our knowledge of our own brains and the power of the computers we are working with mean that this project is not viable in the short term [7] . Even so, there are initiatives like the *Human Brain Project*, financed by the European Union, the *Brain Initiative*, led by the United States and the *Blue Brain Project*, at *l’École Polytechnique Fédérale* in Lausanne (Switzerland), that might ultimately help to make it possible.

“If the human brain were so simple that we could understand it, we would be so simple that we couldn’t.” Time will tell whether this quote attributed to Emerson M. Pugh is true or not. In the meantime, we are left with the speculations that the experts are engaged with. However, according to the outcome of two surveys [8]

published recently, the majority think it is possible that AI will match human intelligence and think that it is most likely to happen within decades, in 2040-2080.

From general AI to the technological singularity

Let's assume that the predictions of the majority of experts are not biased by over-confidence in their own abilities. On that basis, we should be getting ready for something that, if it happens, it will happen not in years but in decades. Do we not have problems today that are more urgent?

From 2011 to 2013, participants at a number of AI conferences, the members of the Greek AI association and one hundred of the most cited authors in the field [9] were asked how long they thought it would take for an AI similar to our own intelligence to clearly overtake human beings as a whole in most respects. 62% of respondents thought it would take 30 years and 19% had worked out it would take less than 2 years. That means that there would be very little time to study AI, debate what to do with it and, where thought necessary, to adopt the international standards and treaties required to regulate AI before it got potentially so superior to us that, in the worst case scenario, we could not impose any form of control on it.

But if after decades of being perfected AI is still so far from true intelligence, how is it possible for the experts to be speculating that, when it does come, progress will be so rapid? First, because once we know how general intelligence works it is highly likely that we will be able to make it work better. In fact, AI itself could work by a process of self-improvement and as it became more intelligent it would be easier for it to make itself even more so. And, assuming that we do manage to emulate a human brain without fully understanding how it works, we would only need to speed its processing up to increase its capacity. From that starting point, the limit would depend on the hardware in our computers, which have been growing in power exponentially for decades. That is not an exaggeration but a phenomenon known as Moore's Law: the assertion that the transistors built into microprocessor chips have doubled every short period. In 1980, the chips in a PC held less than 105 transistors and now they hold more than 10(10) [10] . According to Ray Kurzweil's calculations, right now our computers are about as powerful as the brain of a mouse, but he calculates that during the 2020s they will reach the processing capacity of a human being 10(16) and by 2045 they will reach the equivalent of the collective power of the whole of humanity 10(26).

Regardless of whether Kurzweil's calculations about the power of the brain are correct, they highlight the exponential progress that a general AI would make if it emerged while Moore's Law still held. We should not forget, either, that Moore's Law may not last for ever (and it is true that it is showing signs of obsolescence), but there are new technologies in prospect, like quantum computing that could, if made to work, make for even faster growth.

All in all, it is not unthinkable that in a matter of decades we could go from having hardware that cannot even house a lesser intelligence than our own to having a computer

that could outperform the whole of humanity working together towards a common goal. The process would not have to stop there, either, and with the passage of time, the growth curve could become steeper and steeper and attain almost infinite rates of increase in processing power. In other words, we could witness what some writers have called the "technological singularity".

Many of those who call themselves transhumanists have been arguing for decades that the technological singularity could solve humanity's greatest ills: cure all disease, produce clean energy that helps to win the fight against climate change, reduce the cost of producing and distributing food to end hunger and develop technologies to colonise space so as to overcome overpopulation of the planet. At the same time, it might also make achievable the ancient dream of immortality, by digitising our brains and putting them in the cloud.

The dream of post-humanity

Transhumanism is a cultural and intellectual movement that starts from the premise that human beings in their current form are not the endpoint of our development. Transhumanists think that we can use science and technology to overcome our biological limitations to achieve, among other things: more powerful senses, greater empathy, better memory, faster thinking, greater artistic ability, less sleep, less pain, better health and increased life expectancy [11].

At the beginning of the 20th century, it was thought that the best tools for achieving the aims of transhumanism would come in the form of new drugs or would be to do with genetics. Since the 1960s until now, however, the movement has had great hopes of AI development. It is true that we have surrounded ourselves with every more powerful and ever more portable devices and AI plays an increasingly vital role in our decision-making. In fact, sometimes we let AI make the actual decision (who hasn't let YouTube pick the next video they "want" to watch?) So much so that some writers think that our current dependency on technology already makes us, in a sense, a sort of person-machine hybrid (or cyborg). In recent years we have developed peripherals like electronic watches and wireless headphones that we typically wear right next to our skin. Is it really unthinkable that in the medium term, a large part of the population could have prosthetics in their eyes or ears despite having nothing wrong with them? And in their brains? It seems like science fiction, but there are real start-ups, like Neuralink [12], who are looking into that very possibility. This could be just the beginning.

Given all that, it is not surprising that some writers have concluded that transhumanism looks a lot like a religion. A religion that is consistent with a materialist view of the universe where science takes the place of faith, but without giving up humanity's ancestral longings, like immortality and paradise and, in a sense, even the soul (what is the soul, if not a digitised mind?).

One of transhumanism's strengths is that it wants to achieve these goals through science

and technology, but its main weakness is that it depends on advances or discoveries that are a long way beyond our reach. What's more, transhumanism also has doctrine of sorts for those who want to get as close as possible to paradise. Doctrine with laudable aims like boosting scientific and technological progress and other not so laudable aims, like opposing regulation of the industries involved, acceptance of the low levels of social involvement and participation in the tax system of the great technology corporations. It is not surprising therefore that amongst the leading priests and prophets of transhumanism we should find a good number of Silicon Valley gurus or that the movement receives support from businesses that, in the final analysis, may be looking to boost their own profits [13].

Therefore, without losing sight of the vistas that would open up for us if we achieved superintelligence, we must add a dose of scepticism to the optimism in the air in some parts, since that optimism could be fruit of unmet desires or just reflect commercial interests.

Nor should we forget that the utopia we are promised by some transhumanists, in connection with the singularity, is not risk-free: AI could be used to benefit a privileged few at the expense of the rest or could turn against the whole of humanity.

Finally, if we truly are close to an enormous leap as a species that might lead either to paradise or to extinction, we should think about it together not as isolated individuals who just have their own lives to lose. We are not taking risking ourselves as individuals, or even as a generation, but as the whole future of our species, or maybe even all of life as we know it.

An existencial risk

In 2015, Stephen Hawking said that success in the creation of AI would be the greatest achievement in the history of humanity, but that unfortunately it might also be the last if we do not learn how to avoid the risks. Some of the leading figures in the world of technology, like Elon Musk, Bill Gates and Steve Wozniak [14] have issued warnings in the same vein.

Others, like Mark Zuckerberg [15] , think that there is no reason to see AI so negatively, that there is plenty of time to regulate it, or that we will never lose control of it.

Once again, the experts do not agree amongst themselves. One group of researchers [16] in this area asked whether creating a superintelligence would be a good thing or a bad thing and 41% of respondents said it would be a good thing, 23% were neutral and 17% said it would be a bad thing. At the same time, 18% said it would be catastrophic. We can say, therefore, that most experts have an optimistic view of superintelligence but that we should not underestimate the significant chance that it all leads to a drastically terminal end.

The power of a superintelligence connected to the web would be almost total. It could use every connected device for its purposes – and what today is not connected? It would be able

to see through our cameras, hear through our microphones and could even back up itself in our computers. Plus, we already have robots like Atlas, that can competently interact with the physical environment, that would enable a superintelligence to change the environment as it wished. But if humanity's creations did not satisfy it, it could surely use our factories to make others. In principle it would seem that disaster could be prevented just by making sure that the superintelligence is born on a computer with a level of access to the outside world determined by us (clearly not including access to the Internet), but if it truly was an intelligence so superior to our own, it might well be about to find technical loopholes or use social engineering to escape its prison. In other words, it could trick us, manipulate us or tempt us to make some of the human beings that interact with it give it access to the outside world.

But would a superintelligence want to do us any harm? That would depend to a great extent on its nature and the purposes programmed into it. A superintelligence based on the human brain could have the same inclinations to do good and evil that we have, but would also have a more absolute power than any head of state has ever had. A head of state that would not have a limited term of office, because it would not have to retire or die. If on the other hand the superintelligence was created in a completely artificial way, using self-improvement, as AlphaZero does to learn to play board games, it could be a completely different type of intelligence to our own. An intelligence that could completely ignore everything we see as important. The philosopher Nick Bostrom [17] has warned that an entity like that, programmed with the single aim of finding pi to the greatest possible number of decimal places could end up exterminating life on Earth or throughout the universe. The most ironic thing is that humanicide would not be deliberate. This machine could destroy every ecosystem for the sole aim of having more sources of energy to achieve its assigned goal. An impossible and absurd goal, at least from our point of view.

How can we control general AI?

One way of ensuring that AI is well-disposed towards us is to program into it aims aligned with our ethical principles. In principle, this might look easy. As easy as making sure that AI follows our noble aim of making ourselves happy. However, an entity that did not understand humanity very well might think that the solution was as simple as putting electrodes into the pleasure centres in our brains. So we would have to do more, and give AI true human morals that it could not skip or distort. But, what morals are we talking about, when even the Universal Declaration of Human Rights is still today not accepted by everyone? And, are we sure that we want to be tied to the morality of our own times and give up the chance that in the future our principles might be different from those of today? Surely that could not be the right decision when we remember the extent to which and how fast humanity's values have changed over the last century.

Perhaps the more sophisticated solution to the lack of objective morals would be the solution put forward by Elizer Yudowsky [18] . In general terms, it would consist in programming AI so that it always acts, not as we think it should act, but in the way we would want it to act if we were the best possible version of ourselves. But imperfect beings

that we are, could we ever have the moral judgement of a hypothetical ideal version of ourselves? Why settle for being governed by an AI that knows how we would act if we were better? Why resign ourselves to being imperfect beings? For the transhumanist movement, the preferred scenario is not to delay the development of AI, but quite the opposite. If we have technology that allowed us to connect our brains to computers or go on the web, we could be part of that exponentially growing intelligence, taking on and guiding its potential, but right now we are a long way from knowing if that is a viable possibility

Who would control general AI?

On 28 September 2016, Amazon, Facebook, Google, DeepMind, Microsoft and IBM (later to be joined by Apple) launched a private initiative to develop good practice for AI.

Governments are reacting more slowly than the private sector, but even so countries like China, the United Kingdom, the European Union and the United States of America have started to make moves to regulate AI. Nevertheless, none of those countries has seriously engaged with the challenges of researching and achieving general AI [19] . In fact, the UK's AI Committee, in its report of 16 April 2017, openly declared that the issue is not expected to affect the public in the short term. Similarly, the European Union and the United States have been happy just to ignore it.

The fact is that governments have been ignoring the potential emergence of general AI because they do not think it can happen any time soon. Also, right now, it is not a concern in terms of public opinion. The main problem is that it will be difficult to deal with it sufficiently in advance of when it does emerge to respond adequately to its emergence. When scientist are unanimous in saying that general AI is imminent or about to be created and manage to convince the politicians all of a sudden, it may already be too late. Look at the climate crisis: scientific research in that area started in the nineteenth century and up to the 1980s there was no real consensus among the experts. Now we are well into the 21st century, everyone can see the threat, but Governments continue to be unable to agree who has to pay the bill and meanwhile we are still polluting the atmosphere. Being realistic, maybe to the point where we are at risk of extinction

Now, the main measures to control AI come from the self-regulation of business, led by a group of powerful multinationals such as Google, Facebook, Apple and Amazon. As such, if they do achieve a superintelligence, they would increase their own power over the rest of the industry – and over the whole world – even more drastically. Bearing that in mind, you might think that one way of addressing these problems would be to stimulate competition with the aim of sharing the power more widely. If the problem is inequality, let's make sure that there is more than one superintelligence. That possibility, however, would bring even more risk. If general AI improves exponentially, the business that gets there first (even if by only a few months) will have an advantage that could only grow over time, putting it in a position of permanent advantage. That being the case, two or more businesses being very close to creating a general AI could unleash a perfect storm. Driven by the prospect of such an enormous success, the competitors could take greater and greater risks and bend their

safety rules more and more. As we have already seen, small programming errors in a superintelligence could lead to catastrophic harm for the whole of humanity.

Ultimately, bearing all that in mind, Governments should look for cooperation to turn general AI into an international collaborative project (like CERN or the International Space Station or the Human Genome Project). Perhaps in that way they could avoid any one individual, business or country monopolising control of a hypothetical general AI. At the same time we would be a little closer to ensuring that the superintelligence, if it ever comes about, would benefit the whole of humanity (and other sentient beings) fairly and would be at the service of widely shared ethical principles, as proposed by Nick Bostrom20 [20].

It is possible that in our lifetimes we will have to confront some of the most important decisions that humanity has ever made. Even so, right now we are like trapeze artists swinging between extinction and eternity with our eyes closed. Perhaps the groundwork has already been laid for a mind totally different to all the minds we have known up to now. A mind that can bring paradise, humanicide, or a new way of being us, but for the time being we prefer to look the other way. We need in-depth analysis, at every level, looking at all the many possibilities and consequences, because at the end of the day a relatively unlikely outcome becomes of vital significance if it affects our future as a species (or as a lifeform). We know so little about the universe and about humanity itself that it is impossible not to feel dizzy in the face of such decisions. But we hold the reins of the future in our hands and, right now, we alone are responsible for passing on to our children the best possible future. If we are mistaken and superintelligences never come about, at the end of the day, we will only have lost a little time, and what is a little time compared to eternity?

REFERENCES

1 — Paine, C. (2018) Do you trust in this computer? Papercut Films.
2 — Kurzweil, R. (2006) The Singularity is Near: When Humans Transcend Biology. Penguin.
3 — This is, for instance, the line of work of OpenCog.
4 — Searle, J. R. (2007) Biological naturalism. In Max Velmans & Susan Schneider. Blackwell.
5 — Dreyfus, H. (1972) What Computers Can't Do, New York: MIT Press.
6 — Kurzweil, R. (2013) How to Create a Mind: The Secret of Human Thought Revealed. Penguin.
7 — Fan, X., Markram, H. (2019) A Brief History of Simulation Neuroscience, Frontiers in Neuroinformatics, vol. 13, 7.
8 — Grace, K., Salvatier, J., Dafoe, A., Baobao Z.. (2018) When Will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research 62 729-754.
Müller V.C., Bostrom N. (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller V. (eds) Fundamental Issues of Artificial Intelligence.
9 — Müller V.C., Bostrom N. (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller V. (eds) Fundamental Issues of Artificial Intelligence.

10 — The calculations presented by Ray Kurzweil in “The Singularity is Near: When Humans Transcend Biology” presupposed an annual increase in transistors per microprocessor that is currently significantly lower than expected. This slowdown in Moore’s Law, in principle, should have delayed his predictions, but even so, in 2017 the author kept his initial forecasts considerably, predicting the achievement of both consumer computers powerful as the human brain in 2029 and firmly in the imminence of technological uniqueness. The opinion of the author can be consulted in his speech at the 2017 SXSW Festival.
11 — Transhumanist FAQ. Humanity+.
12 — Winkler, R. (2017) Elon Musk Launches Neuralink to Connect Brains With Computers. Wall Street Journal.
13 — Interrelations between transhumanism and Silicon Valley can be seen, for instance, in: Vance. A. (2010) Merely Human? That’s So Yesterday.
14 — Sainato, M. (2015) Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence.
15 — Wagner, K. (2017) Mark Zuckerberg thinks AI fearmongering is bad. Elon Musk thinks Zuckerberg doesn’t know what he’s talking about.
16 — Müller V.C., Bostrom N. (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller V. (eds) Fundamental Issues of Artificial Intelligence.
17 — Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc.
18 — Yudkowsky, E. (2004) Coherent Extrapolated Volition. The Singularity Institute.
19 — McLay, R. (s. d.) Managing the rise of Artificial Intelligence.
20 — Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc.

Miquel Casas

Miquel Casas és llicenciat en Dret per la Universitat Pompeu Fabra i té un màster en Filosofia amb especialitat en Lògica, Historia i Filosofia de la Ciència per la UNED. Ha treballat com a tècnic superior en diversos departaments de la Generalitat de Catalunya i a l'Institut Català de l'Acolliment i l'Adopció, i actualment és assessor jurídic a i l'Institut Català Internacional per la Pau (ICIP). Ha publicat el seu treball de fi de màster a Ápeiron Ediciones: *El fin del Homo sapiens: la naturaleza y el transhumanismo* (2017).