Ultraintelligent Machines, Singularity, and Other Sci-fi Distractions about AI

Luciano Floridi

University of Oxford and Alma Mater Università di Bologna

Summary

In this article, I argue that the development of AI in terms of successful agency without intelligence does not lead to any fanciful realisation of science fiction scenarios (Singularity), which are at best distracting and at worst irresponsible; and that any denial of AI as a revolution in how we create, control, and conceptualise agency is also wrong. The article concludes by highlighting how this calls for ethical foresight and design of the kind of infosphere and information societies we would like to develop.

1. Introduction: the ancestral fear of monsters upgraded

Suppose you enter a dark room in an unknown building. You could panic about monsters that might be lurking in the dark, or just turn on the light to avoid bumping into furniture. The dark room is the future of AI. Unfortunately, some people believe that, as we step into the room, we may run into evil, ultraintelligent machines. Fear of some kind of ogre, such as a Golem or a Frankenstein, is as old as human memory. The computerised version of such fear dates at least to the 1960s when Irving John

Good[1], a British mathematician who worked as a cryptologist at Bletchley Park with Alan Turing, made the following observation:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus, the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously. (Good 1965), 33)

Once ultraintelligent machines become a reality, they may not be docile at all. Instead, they could behave like *Terminator* or, rather *Skynet* (in the movie, this is the AI defence network that becomes self-aware and initiates a nuclear holocaust). They might enslave humanity as a sub-species, ignore its rights, and pursue their own ends regardless of the effects on human lives. If this sounds too incredible to be taken seriously, fast-forward half a century and consider how the amazing developments in our digital technologies have led some people to believe that Good's 'intelligence explosion' (sometimes also known as Singularity), may be a serious risk—and that if we are not careful, the end of our species may be near. For instance, Stephen Hawking said, 'I think the development of full artificial intelligence could spell the end of the human race' (Holley 2 December 2014). Bill Gates is equally concerned. During an 'ask me anything' question-and-answer session on Reddit, he wrote:

> I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super intelligent. That should be positive if we manage it well. A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned.

[1] https://www.theguardian.com/science/2009/apr/29/jack-good-codebreaker-obituary

The statement has been reproduced many times (for example, see (Mack 28 January 2015), (Holley 29 January 2015), (Rawlinson 29 January 2015), (Christian 19 March 2019). And what did Elon Musk, say, exactly?

> I think we should be very careful about artificial intelligence. If I were to guess like what our biggest existential threat is, it's probably that. So, we need to be very careful with the artificial intelligence. Increasingly scientists think there should be some regulatory oversight maybe at the national and international level, just to make sure that we don't do something very foolish. With artificial intelligence we are summoning the demon. In all those stories where there's the guy with the pentagram and the holy water, it's like yeah he's sure he can control the demon. Didn't work out. (McFarland 24 October 2014)

In recent years, Musk has raised increasingly worrying alarms. He has been followed by authors who have popularised the fear of some kind of artificial ultraintelligence or superintelligence[2]. Many disagree or simply do not take such speculations seriously. Some make fun of them. In 2016, the Information Technology and Innovation Foundation (ITIF) awarded

> its annual Luddite Award to a loose coalition of scientists and luminaries who stirred fear and hysteria in 2015 by raising alarms that artificial intelligence (AI) could spell doom for humanity … 'It is deeply unfortunate that luminaries such as Elon Musk and Stephen Hawking have contributed to feverish hand-wringing about a looming artificial intelligence apocalypse', said ITIF President Robert D. Atkinson.[3]

The reality is more trivial and, in a way, more realistically worrisome. Current and foreseeable smart technologies have the intelligence of an abacus, i.e., zero. Some people argue in terms of slow but increasing growth in the current level of intelligence of smart technologies, which means we had better be concerned now because true AI is coming sooner or later. Those people should remember that it does not matter how many zeros you add; they will forever leave you exactly where you started. The trouble

[2] For example, see (Bostrom 2014) and (Russell 2019). I recommend the following review to understand the limits of the previous books: (Leslie 2019).

[3] https://itif.org/publications/2016/01/19/artificial-intelligence-alarmists-win-itif%E2%80%99s-annual-luddite-award

is always human stupidity or evil nature.

A few months after the Luddite Award mentioned above, on 23 March 2016, Microsoft introduced Tay to Twitter. Here is a quick reminder: Tay was an AI-based chat robot that had to be removed from Twitter only 16 hours later. Tay was supposed to become increasingly smarter as it interacted with humans. Instead, it quickly became an evil Hitler-loving, Holocaust-denying, incestual sex-promoting, 'Bush did 9/11'-proclaiming chatterbox. Why? Because it worked no better than a paper towel, absorbing and being shaped by the tricky, nasty messages sent to it. Microsoft apologised (Hunt 24 March 2016).

This is the state of AI today and for any realistically foreseeable future (see, for example, the Loebner Prize and the Turing Test). People keep making extraordinary claims, e.g. about machines becoming conscious.[4] Mass media keep milking such claims for marketing purposes.[5] And meanwhile, machines keep being just machines, with all their successes and limitations (see, for example, GPT-3, (Floridi and Chiriatti 2020)). And yet it is no reason to be complacent. On the contrary, after so much distracting speculation about the fanciful risks of ultraintelligent machines, it is time to turn on the light, stop worrying about distracting sci-fi scenarios, and start focusing on AI's real challenges in order to avoid making painful, costly mistakes in the design and use of smart technologies. In the rest of this article, I shall discuss what I consider to be the main reasons why such concerns (and their opposite, which are overly optimistic views) are mistaken.[6]

2. Believers and disbelievers in true AI: a debate about faith

Philosophy does not do nuance well (this is a sinner's confession). Especially analytic philosophy may like precision and finely honed distinctions, but polarisations and dichotomies are what it really loves. Internalism or externalism, foundationalism or coherentism, trolley left or right, zombies or not zombies, observer-relative or

[4] https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

[5] https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine

[6] For a full analysis of the ethics of AI, see (Floridi forthcoming). I discussed AI winters and the plausible future of AI in (Floridi 2020, 2019).

observer-independent, possible or impossible worlds, grounded or ungrounded, … philosophy may preach the inclusive *vel* ('girls *or* boys may play'), but too often indulges in the exclusive *aut aut* ('*either* you like it, *or* you don't').

The current debate about AI is a case in point. Here, the dichotomy is between believers and disbelievers in *true* AI, also known as AGI (Artificial General Intelligence), Strong AI, Full AI, or Universal AI. Yes, the real thing—not Siri in your iPhone, Roomba (a robot vacuum cleaner) in your living room, or Nest (a smart thermostat) in your kitchen (disclaimer: I am the happy owner of all three). Think instead of the false Maria in *Metropolis* (1927), Hal 9000 in *2001: A Space Odyssey* (1968; Good was one of the consultants), C-3PO in *Star Wars* (1977), Rachael in *Blade Runner* (1982), Data in *Star Trek: The Next Generation* (1987), Agent Smith in *The Matrix* (1999), the disembodied Samantha in *Her* (2013), or Ava in *Ex Machina* (2014). The list continues, but you've got the picture. Believers in true AI and in Good's 'intelligence explosion' belong to the Church of Singularitarians. For lack of a better term, I shall refer to the disbelievers as members of the Church of AItheists. In the rest of this article, I wish to discuss the two faiths and see why both are mistaken. Doing so will clear the ground from potentially misleading views. And in the meanwhile, remember that good philosophy is almost always in the boring middle.

3. Singularitarians: the end is near, true AI is coming

Singularitarians believe in three dogmas. First, the creation of some form of artificial ultraintelligence is likely or at least not impossible in the (for some of them foreseeable) future. This turning point is known as a *technological singularity*, hence the name. Both the nature of such a superintelligence and the exact timeframe of its arrival are left unspecified, although Singularitarians tend to prefer futures that are conveniently close-enough-to-worry-about but far-enough-not-to-be-around-to-be-proved-wrong (more on these imaginative timelines in a moment). Second, humanity runs a major risk of being dominated by such ultraintelligence. Third, a primary responsibility of the current generation is to ensure that the Singularity either does not happen or, if it does, that it is benign and will benefit humanity. This has all the elements of a Manichean view of the world: Good fighting Evil, apocalyptic overtones, the urgency of 'we must do something now, or it will be too late', an eschatological perspective of human salvation, and an appeal to fears and ignorance. Put all this in a context where people are rightly worried about the impact of digital technologies on their lives, especially in the job market, politics, cybercrimes, healthcare or cyber conflicts, and where mass media report new gizmos and unprecedented computer-driven disasters on a daily basis. And you get the perfect recipe for a debate that causes mass distraction—a digital opiate for the masses.

Like all faith-based views, Singularitarianism is irrefutable because, in the end, it is unconstrained by reason and evidence. It is also implausible, since there is no reason to believe that anything resembling intelligent (let alone ultraintelligent) machines will emerge from our current and foreseeable understanding of computer science and digital technologies. Let me explain.

Sometimes Singularitarianism is presented conditionally. This is shrewd, because the *then* does follow from the *if*, and not merely in an *ex falso quod libet* sense: *if* some kind of ultraintelligence were to appear, *then* we *would* be in deep trouble (not merely 'could', as stated above by Hawking). Correct. Absolutely. But this also holds true for the following conditional: *if* the Four Horsemen of the Apocalypse were to appear, *then* we would be in even deeper trouble.

At other times, Singularitarianism relies on a very weak sense of possibility: some form of artificial ultraintelligence *could* develop, couldn't it? Yes, it could. But this 'could' is a mere logical possibility, that is, there is no contradiction in assuming the development of artificial ultraintelligence as far as we know. Yet this is a trick that blurs the immense difference between 'I could be sick tomorrow' when I am already not feeling too well, and 'I could be a butterfly that dreams it's a human being'. There is no contradiction in assuming that a relative you have never heard of just died, leaving you $10 million. That *could* happen. So? Contradictions, like happily married bachelors, aren't possible states of affairs. But non-contradictions (merely logical 'could'), like extra-terrestrial agents living among us so well-hidden that we've never discovered them, can still be dismissed as utterly crazy. Russell had a wonderful analogy to explain the point:

> [...] If I were to suggest that between the Earth and Mars there is a china teapot

> revolving about the sun in an elliptical orbit, nobody would be able to disprove my assertion provided I were careful to add that the teapot is too small to be revealed even by our most powerful telescopes. But if I were to go on to say that, since my assertion cannot be disproved, it is [an] intolerable presumption on the part of human reason to doubt it, I should rightly be thought to be talking nonsense … (Russell 1952).

The singularity is just a case of teapotism. The 'could' in 'artificial ultraintelligence *could* develop' is not like the 'could' in 'an earthquake could happen' when you live in Japan. Instead, it is like the 'could' in 'it is not true that it could not happen' that you are the first immortal human. Correct, but this is no reason to start acting as if you will live forever. That is, unless someone provides evidence to the contrary by showing how there is something in our current and foreseeable understanding of computer science that should lead us to suspect that the emergence of artificial ultraintelligence is even remotely plausible.

This is where Singularitarians mix faith and facts. They are often moved, I like to believe, by a sincere sense of apocalyptic urgency. They start talking about job losses, digital systems at risk, unmanned drones gone awry, and other real and worrisome issues about computational technologies that are coming to dominate areas of human life ranging from education to employment, from entertainment to conflicts. From this evidence, they jump to being seriously worried about their inability to control their next car because it will have a mind of its own. How some nasty ultraintelligent AI will evolve autonomously from the computational skills required to park in a tight spot remains unclear. The truth is that climbing on top of a tree is not the first step toward the moon; it is the end of the journey (a brilliant point made by Hubert Dreyfus, I believe). What we *are* going to see are increasingly smart machines able to perform more and more tasks—some of which we currently perform ourselves, and some of which will be beyond our abilities. For the first time in human history, agency has irreversibly and successfully divorced intelligence. *This* is extraordinary enough without having to believe in science fiction.

When all other arguments fail, Singularitarians are fond of throwing in some maths. A favourite reference is Moore's Law. This is very well known but let me restate it for the sake of clarity: it is the empirical claim that, in the development of digital computers, the number of transistors on integrated circuits doubles approximately every two years. The outcome has so far been more computational power for less. But things are changing. Technical difficulties in nanotechnology present serious manufacturing challenges. There is, after all, a limit to how small things can get before they simply melt. Moore's law no longer holds without some significant innovations (Waldrop 2016); (*The Economist* 19 April 2016). Other kinds of technological solutions will have to be identified, including quantum computing.

Just because something grows exponentially for some time does not mean it will continue to do so forever. Here is a good example of what happens if you are not careful with 'projections':

> Throughout recorded history, humans have reigned unchallenged as Earth's dominant species. Might that soon change? Turkeys, heretofore harmless creatures, have been exploding in size, swelling from an average 13.2lbs (6 kg) in 1929 to over 30lbs today. On the rock-solid scientific assumption that present trends will persist, *The Economist* calculates that turkeys will be as big as humans in just 150 years. Within 6,000 years, turkeys will dwarf the entire planet. Scientists claim that the rapid growth of turkeys is the result of innovations in poultry farming, such as selective breeding and artificial insemination. The artificial nature of their growth, and the fact that most have lost the ability to fly, suggest that not all is lost. Still, with nearly 250m turkeys gobbling and parading in America alone, there is cause for concern. This Thanksgiving, there is but one prudent course of action: eat them before they eat you. (*The Economist* 27 November 2014).

The step from Turkzilla to AIzilla is small, if it weren't for the fact that a growth curve can easily be sigmoid (see picture below). There can be an initial stage of growth that is approximately exponential, followed by saturation, slower growth, maturity, and finally, no further growth. But I suspect that the representation of sigmoid curves might be blasphemous for some Singularitarianists.

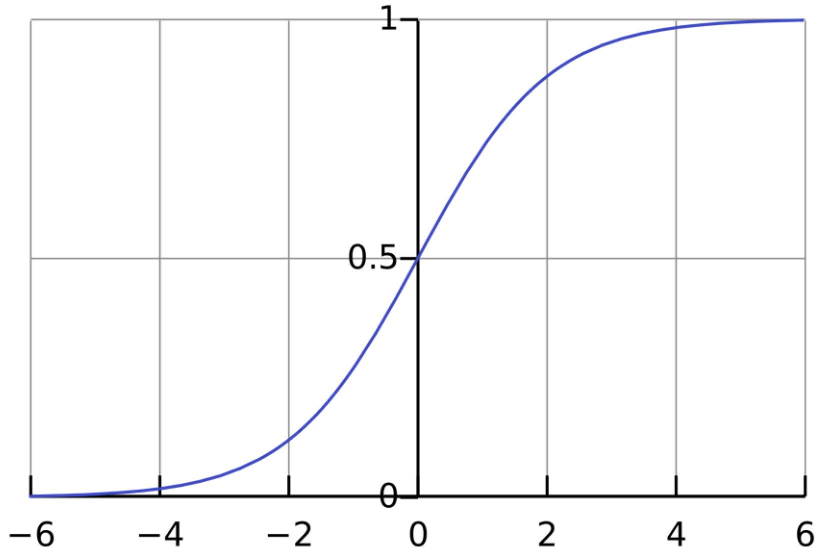

Figure 1 Wikipedia, Graph of Logistic Curve, a typical sigmoid function.

Source: http://commons.wikimedia.org/wiki/File:Logistic-curve.svg#metadata

I used to think that Singularitarianism was merely funny, not unlike people wearing tinfoil hats. Yet I have heard people at specialised conferences arguing that we should be concerned about the Singularity because [add some fallacious reasoning here of your choice] and we had better be safe than sorry. Following the same logic, you should also always bring a wooden stake with you just in case you encounter a vampire. I thought that some of their fallacious reasoning would not fool anyone. I still recall with shock a person in my department arguing that, since we had been wrong before about possibilities that turned out to be realities later, we should take the Singularity very seriously. I remember trying to explain that this 'equation'—A (e.g., flying, the actual example provided by that person) was deemed impossible in the past, but it is possible now, therefore B (in this case true AI, which was once considered impossible) will/may be possible in the future—was simply too permissive and hence useless, since one could replace B with anything (try 'being immortal'). I insisted that it was not a logical argument, but a mere rhetorical manoeuvre. I was unsuccessful, of course (the 'of course' was a lesson I learnt after the attempt). This was not least because anyone who finds that way of reasoning convincing is also probably impervious to any reasonable explanation showing that it is meaningless in the first place (in other words, if one is so silly to find the argument convincing, one is probably too silly to understand the explanation showing that it is mere rhetoric).

Today, I believe that Singularitarianism is neither funny nor just logically irritating, but irresponsibly distracting. It is a rich-world preoccupation likely to worry people in wealthy societies who seem to forget the real evils oppressing humanity and our planet. Unfortunately, the COVID-19 pandemic has reminded doomsayers that we have tragic problems, both serious and pressing. No matter what Musk may think, climate change is 'our biggest existential threat'. It is immoral to speculate about Hollywood scenarios when, already in 2019, according to UNICEF and the WHO,

> billions of people worldwide continue to suffer from poor access to water, sanitation and hygiene, according to a new report by UNICEF and the World Health Organization. Some 2.2 billion people around the world do not have safely managed drinking water services, 4.2 billion people do not have safely managed sanitation services, and 3 billion lack basic handwashing facilities.

These are real, major threats to humanity. Oh, and just in case you thought predictions by experts were a reliable guide, think twice. Great experts have made many technological predictions that are staggeringly wrong (see some hilarious ones in (Pogue 18 January 2012) and (Cracked Readers 27 January 2014). For example, Bill Gates stated in 2004 that 'two years from now, spam will be solved'. In 2011, Stephen Hawking declared that 'philosophy is dead' (Warman 17 May 2011), so you are not reading this article. But the prediction of which I am rather fond is by Robert Metcalfe, co-inventor of Ethernet and founder of 3Com. In 1995, he promised to eat his words if his prediction that the Internet would soon go supernova and catastrophically collapse in 1996 should turn out to be wrong. In 1997, he publicly liquefied his article in a food processor and duly drank it. He was a man of his word. I wish Singularitarians were as bold and coherent as him.

4. AItheism: what computers cannot do, allegedly

I have spent more than a few words to describe Singularitarianism not because it can be taken seriously, but because AI disbelievers, the AItheists, can be better understood as people overreacting to all this singularity nonsense. Deeply irritated by those who worship the wrong digital gods and their unfulfilled Singularitarian prophecies, disbelievers (AItheists) make it their mission to prove once and for all that any kind of faith in true AI is wrong, totally wrong. For them, AI is just computers; computers are just Turing Machines; Turing Machines are merely syntactic engines, and syntactic engines cannot think, cannot know, and cannot be conscious. End of story. This is why there is so much that computers (Olteanu et al.) cannot do. In fact, this is the wording of the titles of several old-fashioned publications (Wilson and Wilson 1970); (Dreyfus 1972); (Dreyfus 1979); (Dreyfus and Dreyfus 1992); (Harel 2000); (Searle 9 October 2014). However, precisely what computers cannot do is a conveniently movable target. It is also why computers cannot process semantics (of any language, Chinese included) no matter what Google translation achieves (Preston and Bishop 2002). This proves that there is absolutely nothing to discuss, let alone worry about. There is no genuine AI, so *a fortiori,* there are no problems caused by it. Relax and enjoy all these wonderful electric gadgets.

The AItheists' faith is as misplaced as that of the Singularitarians. Both churches have plenty of followers in California, where Hollywood sci-fi films, great research universities like Berkeley, and some of the world's most important digital companies flourish side by side. This may not be accidental. When there is big money involved, people can become easily confused. For example, everybody knows that Google has been buying AI tech companies as if there were no tomorrow.[7] Surely, they must know something about the actual chances of developing a computer that can think. We, outside 'The Circle' are just missing something. Then-executive Chairman of Google Eric Schmid fuelled this view when he spoke at The Aspen Institute on 16 July 2013, saying, 'Many people in AI believe that we're close to [a computer passing the Turing Test] within the next *five years* [emphasis added]'.[8] Five

[7] Disclosure: I was a member of Google's Advisory Council on the right to be forgotten (Herritt 30 December 2014) and of the ill-fated Advanced Technology External Advisory Council in 2019 (see https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board).

[8] https://www.youtube.com/watch?v=3Ox4EMFMy48

years. Turing preferred 50 years. Hawking indicated 100 years[9]. The impression that the future of AI would have been very different if we had evolved with six instead of five fingers is strong. But Kurzweil, the person who popularised the expression 'technological Singularity', resisted counting in terms of multiples of five. He predicted that true AI would become available in 2029.[10] That is very precise, so if you are wondering what scientific analysis lies behind the choice of that specific date, it is because this is after his eightieth birthday (he was born in 1948).

The Turing Test is a way to check whether AI is getting any closer. Let me remind you of it: you pose questions to two agents in another room; one agent is human while the other is artificial; if you cannot tell the difference between the two from their answers, then the robot passes the test. It is a crude test. Think of it as a driving test. If Alice does not pass it, she is not a safe driver. Yet if she passes it, she may still be an unsafe driver. In short, the Turing Test provides a necessary but insufficient condition for a form of intelligence. This is a really, very low bar. And yet, no AI has ever gotten over it. More importantly, all programs keep failing in the same way, using tricks developed in the sixties. This is why, in the past, I offered a bet. I hate aubergine (eggplant), but I promised to eat a plate full of it if a software program could win the gold medal (i.e., pass the Turing Test) of a Loebner Prize competition before 16 July 2018, the date offered by Schmidt. It was a safe bet. It is 2022, and we still have no sign of a real winner of the Turing Test. I do not know who the 'many people' in Schmidt's quote are, but I know that the last people you should ask about whether something is possible are those who have abundant financial reasons to reassure you that it is.

So far, we have seen only consolation prizes given to the less poorly performing versions of contemporary ELIZA. It is human interrogators who sometimes fail the Turing test by asking binary questions, such as 'do you like ice cream?' or 'do you believe in God?' (these are real examples). Any answer to these questions would be utterly uninformative in any case (Floridi, Taddeo, and Turilli

[9] https://www.computerworld.com/article/2922442/stephen-hawking-fears-robots-could-take-over-in-100-years.html

[10] https://www.theguardian.com/technology/2014/feb/22/computers-cleverer-than-humans-15-years

2009).

5. Singularitarians and AItheists: a pointless diatribe

Both Singularitarians and AItheists are mistaken. As Turing clearly stated in the article where he introduced his test (Turing 1950), the question 'can a machine think?' is 'too meaningless to deserve discussion'. Ironically, or perhaps presciently, that question is actually engraved on the Loebner Prize medal. It remains meaningless, no matter to which of the two churches one belongs. Yet both churches continue this pointless debate, often suffocating any dissenting voice of reason. True AI is not logically impossible, but it is utterly implausible. People confuse 'the Singularity will never happen' with 'the Singularity is impossible'. Let me restate that "impossible" is a logical concept, and true AI is logically possible. But it is possible in the same way as, for example, a calculation that would take longer than the life of the universe to be completed: it is not going to happen. This is why it is disappointing when experts who should know better timidly hide behind a 'not yet' or 'not for a long time'. It is a harmful case of friendly fire. If it were true that *Terminator* is not yet possible and won't be possible for a long time, we should panic. Immediately. Luckily, the real answer is 'not now, not never', to the best of our knowledge. We have no idea how we might begin to engineer some true AI. This is so not least because we have very little understanding of how our brain and our own intelligence work. It is also unlikely that our conception of intelligence will remain unchallenged, in terms of a unified phenomenon. All this means that we should not lose any sleep over the possible appearance of some ultraintelligence.

What really matters is that the increasing presence of ever-smarter technologies in our lives is having huge effects on how we conceive ourselves, the world, and our interactions among ourselves and with the world. The point is not that our machines are conscious, or intelligent, or able to know something the way we know it. They are not. Plenty of machines can do amazing things, including playing board games like checkers, chess, Go, and the quiz show Jeopardy better than us. And yet, they are all versions of a Turing Machine—an abstract model that sets the limits of what can be done by a computer through its mathematical logic. Quantum computers, too, are constrained by the same computational limits. These are the limits of what can be computed, or so-called computable functions. Nobody seems able to explain how a conscious, intelligent, empathic entity might emerge from a Turing Machine.

The point is that thanks to an enormous amount of available data and some very sophisticated programming, our smart technologies are increasingly able to deal with a growing number of tasks better than we can—including predicting our behaviours—*without* having to be intelligent at all. So, we are not the only agents able to perform tasks successfully. We are far from it. This is what I have defined as the fourth revolution in our self-understanding (Floridi 2014). We are not at the centre of the universe (Copernicus), of the biological kingdom (Darwin), or of the realm of rationality (Freud). And after Turing, we are no longer at the centre of the infosphere, the world of information processing, or smart agency either. Ironically, the BBC made a two-minute short video[11] to introduce this idea of a fourth revolution that is worth watching. But it made a mistake in the end, equating 'better at accomplishing tasks' with 'better at thinking'. I have never argued that digital technologies *think* better than us. Instead, they can *do more and more things* better than us *without thinking* and only by processing increasing amounts of data more and more quickly, efficiently, and effectively. And if the latter is the definition of thinking used to win the argument, then we are having a linguistic debate.

We share the infosphere with digital technologies. These are not the children of some sci-fi superintelligence, but ordinary artefacts that outperform us in ever more tasks despite being no cleverer than a toaster. Their abilities are humbling. They make us re-evaluate our human exceptionality and our special role in the universe, which remains unique. We thought we were smart because we could play chess. Now, a phone plays better than a chess master. We thought we were free because we could buy whatever we wished. Now, our spending patterns are predicted, sometimes even anticipated, by devices as thick as a plank. What does all this mean for our self-understanding? That is a question worth investigating philosophically.

The success of our technologies largely depends on the fact that, while we were

[11] http://www.bbc.co.uk/programmes/p02hvcjm

speculating about the possibility of ultraintelligence, we increasingly enveloped the world in so many devices, sensors, applications, and data that it became an ICT-friendly environment. This is an environment where technologies can replace us without having any understanding, mental states, intentions, interpretations, emotional states, semantic skills, consciousness, self-awareness, or flexible intelligence. Memory (as in algorithms and immense datasets) outperforms intelligence when landing an aircraft, finding the fastest route from home to the office, and discovering the best price for your next fridge. Digital technologies can *do more and more things* better than us because they can process increasing amounts of data and improve their performance by analysing their own output as input for subsequent operations.

AlphaGo, the computer program developed by Google DeepMind, won the board game Go against the world's best player only because it could use a database of around 30 million moves and play thousands of games against itself, 'learning' a bit more each time about how to improve its performance (Silver et al. 2016). AlphaZero learnt to play better than any human or other software by playing against itself, relying only on the rules of the game. It is like a two-knife system that can sharpen itself. What's the difference? It is the same as between you and the dishwasher when washing the dishes. What's the consequence? It is that any apocalyptic vision of AI can be disregarded. The serious risk is not the appearance of some ultraintelligence. The serious risk is that we may misuse our digital technologies to the detriment of a large percentage of humanity and the whole planet.

6. Conclusion: the problem is not HAL but H.A.L., humanity at large

We are and shall remain, for any foreseeable future, the problem—not our technology. This is why we should turn on the light in the dark room and watch carefully where we are going. There are no monsters, but there are plenty of obstacles to avoid, remove, or negotiate. We should be worried about real, human stupidity, not imaginary artificial intelligence, and concentrate on the actual challenges raised by AI. By way of conclusion, let me list five of them (all equally important).

First, we should make AI environmentally friendly. We need the smartest technologies we can build to tackle the concrete evils oppressing humanity and our planet. These range from environmental and health-related disasters to financial crises, from crime, terrorism, and war to famine, poverty, ignorance, inequality, and appalling living standards. Second, we should make AI human-friendly. AI should be used to treat people always as ends and never as mere means, to paraphrase Kant. Third, we should make AI's stupidity work for human intelligence. We have seen how millions of jobs will be disrupted, eliminated, and created; the benefits of this transformation should be shared by all, and the costs should be borne by society. Fourth, we should make AI's predictive power work for freedom and autonomy. Marketing products, influencing behaviours, nudging people, or fighting crime and terrorism should never undermine human dignity. And finally, we should make AI make us more human. The serious risk is that we may misuse, overuse, or underuse our smart technologies to the detriment of the entire planet and much of humanity.

Singularitarians and AItheists will continue their diatribes about the possibility or impossibility of true AI for the time being. We need to be tolerant, but we do not have to engage. As Virgil suggests to Dante in *Inferno*, Canto III: 'Speak not of them, but look, and pass them by'. For the world needs some good philosophy, and we must take care of more pressing problems.[12] Let me remind you of the quote by Winston Churchill that I used in the preface to this book: 'we shape our buildings, and afterwards, our buildings shape us'. This applies to the infosphere and the smart technologies inhabiting it as well. We'd better get them right as soon as possible.

References

Bostrom, Nick. 2014. *Superintelligence: paths, dangers, strategies*. Oxford, England: Oxford University Press.

Christian, Jon. 19 March 2019. "Bill Gates Compares Artificial Intelligence to Nuclear Weapons." *Futurism*.

Cracked Readers. 27 January 2014. "26 Hilariously Inaccurate Predictions About the Future." *http://www.cracked.com/photoplasty_777_26-hilariously-inaccurate-predictions-about-future/#ixzz3QWYlN4gg*.

D'Agostino, Marcello , and Luciano Floridi. 2009. "The Enduring Scandal of Deduction. Is Propositional Logic really Uninformative?" *Synthese* 167 (2):271-315.

[12] For some further background information, see (Floridi 2009), (Floridi 2011), (Floridi 2008). On the debate about the tautological nature and informativeness of logic, see (D'Agostino and Floridi 2009).

Dreyfus, Hubert L. 1972. *What computers can't do : a critique of artificial reason*. New York ; London: Harper & Row.

Dreyfus, Hubert L. 1979. *What computers can't do : the limits of artificial intelligence*. Rev. ed, *Harper colophon books*. New York: Harper & Row.

Dreyfus, Hubert L., and Hubert L. Dreyfus. 1992. *What computers still can't do : a critique of artificial reason*. 3rd ed ed. Cambridge, Mass ; London: MIT Press.

Floridi, Luciano. 2008. "Artificial Intelligence's New Frontier: Artificial Companions and the Fourth Revolution." *Metaphilosophy* 39 (4/5):651-655.

Floridi, Luciano. 2009. "The information society and its philosophy: Introduction to the special issue on "the Philosophy of Information, its Nature, and future developments"." *The Information Society* 25 (3):153-158.

Floridi, Luciano. 2011. "A Defence of Constructionism: Philosophy as Conceptual Engineering." *Metaphilosophy* 42 (3):282-304.

Floridi, Luciano. 2014. *The Fourth Revolution - How the infosphere is reshaping human reality*. Oxford: Oxford University Press.

Floridi, Luciano. 2019. "What the Near Future of Artificial Intelligence Could Be." *Philosophy & Technology* 32 (1):1-15. doi: 10.1007/s13347-019-00345-y.

Floridi, Luciano. 2020. "What the Near Future of Artificial Intelligence Could Be." In *The 2019 Yearbook of the Digital Ethics Lab*, edited by Christopher Burr and Silvia Milano, 127-142. Cham: Springer International Publishing.

Floridi, Luciano. forthcoming. *The Ethics of AI - Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its nature, scope, limits, and consequences." *Minds and Machines*:1-14.

Floridi, Luciano, Mariarosaria Taddeo, and Matteo Turilli. 2009. "Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest." *Minds and Machines* 19 (1):145-150. doi: 10.1007/s11023-008-9130-6.

Freud, S. 1955. *A difficulty in the path of psycho analysis*. Vol. XVII(1917-1919): 135-144., *The Standard Edition of the Complete Psychological Works of Sigmund Freud*.

Good, I. J. 1965. "Speculations concerning the first ultraintelligent machine." In *Advances in Computers, volume 6.*, edited by F. Alt and M. Ruminoff, 31-88. Academic Press.

Harel, David. 2000. *Computers Ltd : what they really can't do*. Oxford: Oxford University Press.

Herritt, Robert. 30 December 2014. "Google's Philosopher." *Pacific Standard* http://www.psmag.com/nature-and-technology/googles-philosopher-technology-nature-identity-court-legal-policy-95456.

Holley, Peter. 2 December 2014. "Stephen Hawking just got an artificial intelligence upgrade, but still thinks AI could bring an end to mankind." *The Washington Post*.

Holley, Peter. 29 January 2015. "Bill Gates on dangers of artificial intelligence: 'I don't understand why some people are not concerned." *The Washington Post*.

Hunt, Elle. 24 March 2016. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." *The Guardian*.

Leslie, David. 2019. "Raging robots, hapless humans: the AI dystopia." *Nature* 574

(7776):32-34.

Mack, Eric. 28 January 2015. "Bill Gates Says You Should Worry About Artificial Intelligence." *Forbes*.

McFarland, Matt. 24 October 2014. "Elon Musk: 'With artificial intelligence we are summoning the demon.'." *The Washington Post*.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." *SSRN Electronic Journal*. doi: 10.2139/ssrn.2886526.

Pogue, David. 18 January 2012. "Use It Better: The Worst Tech Predictions of All Time - Plus, flawed forecasts about Apple's certain demise and the poor prognostication skills of Bill Gates." *http://www.scientificamerican.com/article/pogue-all-time-worst-tech-predictions/*.

Preston, John, and Mark Bishop. 2002. *Views into the Chinese room : new essays on Searle and artificial intelligence*. Oxford: Clarendon Press.

Rawlinson, Kevin. 29 January 2015. "Microsoft's Bill Gates insists AI is a threat." *BBC News*.

Russell, Bertrand. 1952. "Is There a God? Bertrand Russell." In *The Collected Papers of Bertrand Russell Vol. 11: Last Philosophical Testament, 1943–68*, 542-548. London: Routledge.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*: Penguin Publishing Group.

Searle, John R. 9 October 2014. "What Your Computer Can't Know." *The New York Review of Books* http://www.nybooks.com/articles/archives/2014/oct/09/what-your-computer-cant-know/.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (7587):484-489. doi: 10.1038/nature16961 http://www.nature.com/nature/journal/v529/n7587/abs/nature16961.html#supplementary-information.

The Economist. 19 April 2016. "The end of Moore's law." *The Economist*.

The Economist. 27 November 2014. "Turkzilla!" http://www.economist.com/blogs/graphicdetail/2014/11/daily-chart-16.

Turing, A. M. 1950. "Computing machinery and intelligence." *Mind* 59 (236):433-460.

Waldrop, M. Mitchell. 2016. "The chips are down for Moore's law." *Nature* 530:144–147.

Warman, Matt. 17 May 2011. "Stephen Hawking tells Google 'philosophy is dead'." *The Telegraph* http://www.telegraph.co.uk/technology/google/8520033/Stephen-Hawking-tells-Google-philosophy-is-dead.html.

Wilson, Ira Gaulbert, and Marthann E. Wilson. 1970. *What computers cannot do*. Princeton: Vertex.