Singularitarians, AItheists, and Why the Problem with Artificial Intelligence is H.A.L. (Humanity At Large), not HAL

Luciano Floridi
OXFORD INTERNET INSTITUTE, UNIVERSITY OF OXFORD
LUCIANO.FLORIDI@OII.OX.AC.UK

It is awkward and a bit embarrassing to admit, but average philosophy does not do well with nuances. It may fancy precision and very finely cut distinctions, but what it really loves are polarizations and dichotomies. Internalism or externalism, foundationalism or coherentism, trolley left or right, zombies or not zombies, observer-relative or observer-independent, possible or impossible worlds, grounded or ungrounded, . . . philosophy may preach the inclusive *vel* but too often indulges in the exclusive *aut aut*. Such an ability to reduce everything to binary alternatives means that anyone dealing with the continuum of real numbers (pun intended) is likely to be misunderstood.

The current debate about artificial intelligence (AI) is a case in point. Here the dichotomy is between believers and disbelievers in *true* AI. Yes, the real thing, not Siri in your iPhone or Roomba in your kitchen. Think instead of the false Maria in *Metropolis* (1927), Hal 9000 in *Space Odyssey* (1968), C3PO in *Star Wars* (1977), *Rachael* in Blade Runner (1982), Data in *Star Trek: The Next Generation* (1987), Agent Smith in *The Matrix* (1999), or the disembodied Samantha in *Her* (2013). You got the picture. Believers in true AI 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. Let's have a look at both faiths.

Singularitarianism is based on three dogmas. First, the creation of some form of artificial superintelligence—a so-called technological singularity—is likely to happen in the foreseeable future. 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. Second, humanity runs a major risk of being dominated by such superintelligence. Third, a primary responsibility of the current generation is to ensure that the Singularity either does not happen or, if it does, it is benign and will benefit humanity. As you can see, there are all the elements for a Manichean view of the world, with Good fighting against Evil, some 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 idiotic digital technologies on their lives, while the mass media report about new gizmos and unprecedented computer disasters on a daily basis, and you have the perfect recipe for a debate of mass distraction.

Like all views based on faith, Singularitarianism is irrefutable. It is also ludicrously implausible. You may more reasonably be worried about extra-terrestrials conquering earth to enslave us. 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 superintelligence were to appear, *then* we would be in deep trouble. Correct. 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, trust me. Some other times, Singularitarianism relies on mere possibilities: Some form of artificial superintelligence *could* develop, couldn't it? Yes, it could. But this is a mere logical possibility, that is, to the best of our current and foreseeable knowledge there is no contradiction in assuming the development of a superintelligence. It is 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 to be a human being." There is no contradiction in assuming that a relative of yours you never heard of just died leaving you $10m. Yes, he could. So? Contradictions are never the case, but non-contradictions can still be dismissed as utterly crazy.

When conditionals and modalities are insufficient, then Singularitarians, often moved, I like to believe, by a sincere sense of apocalyptic urgency, mix faith and facts. They start talking about job losses, digital systems at risks, and other real and worrisome issues about computational technologies dominating increasing aspects of human life, from learning to employment, from entertainment to conflicts. From this, they jump to being seriously worried about being unable to control their next Honda Civic because it will have a mind of its own. How true AI and superintelligence will ever evolve autonomously from the skill to park in a tight spot remains unclear, but you have been warned, you never know, and surely you better be safe than sorry.

Finally, if even this stinking mix of "could," "if . . . then," and "look at the current technologies . . ." does not work, there is the maths. A favourite reference is the so-called Moore's Law. This is an empirical generalization that suggests that, in the development of digital computers, the number of transistors on integrated circuits doubles approximately every two years. The outcome is more computational power at increasingly cheaper prices. This has been the case so far, and it may well be the case for the foreseeable future, even if technical difficulties concerning nanotechnology have started raising some serious manufacturing challenges. After all, there is a physical limit to how small things can get before they simply melt. The problem is that just because something grows exponentially, this does not mean that it develops without boundaries. A great example was provided by *The Economist* last November:

> 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 13.2lb (6kg) in 1929 to over 30lb 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."[1]

From Turkzilla to Alzilla, the step is small, if it weren't for the fact that a growth curve can easily be sigmoid (see Figure 1), with an initial stage of growth that is approximately exponential, followed by saturation, then a slower growth, maturity, and finally no further growth. But I suspect that the representation of sigmoid curves might be blasphemous for Singularitarianists.

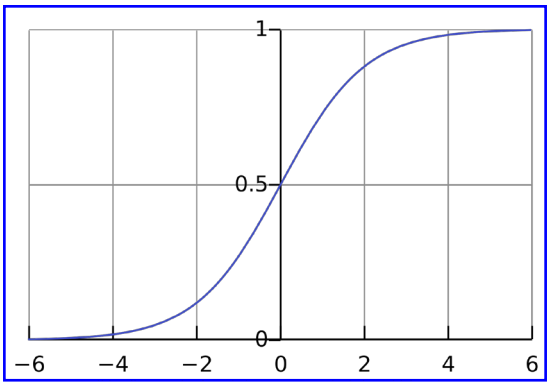

Figure 1. Graph of Logistic Curve, a typical sigmoid function. Wikipedia, http://commons.wikimedia.org/wiki/File:Logistic-curve.svg#metadata

Enough. I used to think that Singularitarianism was merely funny. Not unlike people wearing tin foil hats. I was wrong, for two reasons. First, plenty of intelligent people have joined the Church: Bill Gates, Stephen Hawking, or Elon Musk, Tesla CEO, who has gone as far as to tweet that "We need to be super careful with AI. Potentially more dangerous than nukes." I guess we shall be safe from true AI as long as we keep using Windows but, sadly, such testimonials have managed to transform a joke into a real concern. Second, I have realized that Singularitarianism is irresponsibly distracting. It is a rich-world preoccupation, likely to worry people in leisure societies, who seem to forget what real evils are oppressing humanity and our planet, from environmental disasters to financial crises, from religious intolerance and violent terrorism to famine, poverty, ignorance, and appalling living standards, just to mention a few. Oh, and just in case you thought predictions by experts were a reliable guide, think twice. There are many staggeringly wrong technological predictions by great experts.[2] For example, in 2004 Bill Gates stated "Two years from now, spam will be solved." And in 2011 Stephen Hawking declared that "philosophy is dead," so you are not reading this article.[3] 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 will soon go supernova and in 1996 will catastrophically collapse" should turn out to be wrong. In 1997 he publicly liquefied his article in a food processor and duly drank it. A man of his word. I wish Singularitarians were as bold and coherent as him.

I have spent more than a few words to describe Singularitarianism not because it can be taken seriously, but because AI disbelievers, the Altheists, can be better understood as people over-reacting to all this singularity nonsense. I sympathise. Deeply irritated by the worshipping of the wrong digital gods and the catastrophic prophecies, the Church of Altheism makes its mission to prove once and for all that any kind of faith in true AI is really wrong, totally wrong. 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 the story. AI does not and cannot exist. Even bigots should get it. This is why computers (still) cannot do something (the something being a conveniently movable target), and are unable to process semantics (of any language, Chinese included, no matter what Google translation achieves). This proves that there is absolutely nothing to talk about, let alone worry about. There is no AI, so *a fortiori* there are no problems caused by it; relax and enjoy all these wonderful electric gadgets.

Both Churches seem to have plenty of followers in California, the place where Hollywood sci-fi films, wonderful research universities like Berkeley, and some of the most important digital companies in the world live side by side. This may not be accidental, especially when there is a lot of money involved. For example, everybody knows that Google has been buying AI tech companies as if there were no tomorrow (disclaimer: I am a member of Google's Advisory Council on the right to be forgotten.[4] Surely they must know something, with regard to the real chances of developing a computer that can think, that we, outside "The Circle," are missing. Thus, Eric Schmidt, Google Executive Chairman, speaking at The Aspen Institute on July 16, 2013, stated, "Many people in AI believe that we're close to [a computer passing the Turing Test] within the next five years."[5] I do not know who the "many" 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 let me offer a bet. I hate aubergine (eggplant), but I shall eat a plate full of it if a software program will get the gold medal (i.e., pass the Turing Test) of a Loebner Prize competition before July 16, 2018. It is a safe bet. So far, we have seen only consolation prizes given to the less badly performing versions of contemporary ELIZA. As I explained when I was a judge the first time the competition came to the UK, it is human interrogators who often fail the test, by asking binary questions such as "Do you like ice cream?" or "Do you believe in God?" to which any answer would be utterly uninformative in any case.[6] I wonder whether Gates, Hawking, Musk, or Schmidt would like to accept the bet, choosing a food of their dislike.

Let me be serious again. Both Singularitarians and Altheists are mistaken. As Alan Turing clearly stated in the article where he introduced his famous test (Turing 1950), the question "Can a machine think?" is "too meaningless to deserve discussion" (ironically, or perhaps presciently, that

question is engraved on the Loebner Prize medal). This holds true, no matter which of the two Churches you belong to. Yet both Churches dominate this pointless debate, suffocating any dissenting voice of reason. True AI is not logically impossible but it is utterly implausible. According to the best of our scientific knowledge today, we have no idea how we may begin to engineer it, not least because we have very little understanding of how our brain and our own intelligence work. This means that any concern about the appearance of some superintelligence is laughable. 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 as we do. They are not. The point is that they are increasingly able to deal with more and more tasks better than we do, including predicting our behaviors. So we are not the only smart agents around, far from it. This is what I have defined as the fourth revolution in our self-understanding. We are not at the center of the universe (Copernicus), of the biological kingdom (Darwin), or of the realm of rationality (Freud). After Turing, we are no longer at the center of the world of information and smart agency either. 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 and make us revaluate our intelligence, 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?

The success of our technologies largely depends on the fact that, while we were speculating about the possibility of true AI, we increasingly enveloped the world in so many devices, applications, and data that it became an IT-friendly environment, where technologies can replace us without having any understanding or semantic skills. Memory (as in algorithms and immense datasets) outperforms intelligence when landing an aircraft, finding the fastest route from home to the office, or discovering the best price for your next fridge. The BBC has made a two-minutes short animation to introduce the idea of a fourth revolution that is worth watching.[7] Unfortunately, like John Searle, it made a mistake in the end, equating "better at accomplishing tasks" with "better at thinking." I never argued that digital technologies *think* better than us, but that they can *do more and more things* better than us by processing increasing amounts of data. What's the difference? The same as between you and the dishwasher when washing the dishes. What's the consequence? That any apocalyptic vision of AI is just silly. The serious risk is not the appearance of some superintelligence, but that we may misuse our digital technologies, to the detriment of a large percentage of humanity and the whole planet. We are and shall remain for the foreseeable future the problem, not our technology. We should be worried about real human stupidity, not imaginary artificial intelligence. The problem is not HAL but H.A.L., Humanity At Large.

It may all seem rather commonsensical. But if you try to explain it to an Altheist like John Searle he will crucify you together with all the other Singularitarians. In a review of my book, *The Fourth Revolution – How the Infosphere is Reshaping Humanity*, where I presented some of the ideas above, Searle criticized me for being a believer in true AI and a metaphysician who thinks that reality is intrinsically informational.[8] This is nonsense. As you might have guessed by now, I subscribe to neither thesis.[9] In fact, there is much I agree about with Searle's Altheism. So I tried to clarify my position in a reply.[10] Unsuccessfully. Unfortunately, when people react to Singularitarianism, to blind faith in the development of true AI or to other technological fables, they run the risk of falling into the opposite trap and thinking that the debate is about computers (it is not—social media and Big Data, for example, are two major issues in the philosophy of information) and that these are nothing more than electric typewriters, not worth a philosophical investigation. They swing from the pro-AI to the anti-AI, without being able to stop, think, and reach the correct, middle ground position, which identifies in the information revolution a major transformation in our *Weltanschauung*. Let me give you some elementary examples. Our self-understanding has been hugely influenced by issues concerning privacy, the right to be forgotten, and the construction of personal identities online. Just think of our idea of friendship in a world dominated by social media. Our interactions have hugely changed due to online communications. Globalization would be impossible without the information revolution, and so would have been many political movements, or hacktivism. The territoriality of the law has been completely disrupted by the *onlife* (sic) world, in which online and offline experiences are easily continuous, thus further challenging the Westphalian system.[11] Today science is based on Big Data and algorithms, simulations and scientific networks, all aspects of an epistemology that is massively dependent on, and influenced by, information technologies. Conflicts, crime, and security have all been re-defined by the digital, and so has political power. In short, no aspect of our lives has remained untouched by the information revolution. As a result, we are undergoing major philosophical transformations in our views about reality, ourselves, our interactions with reality, and among ourselves. The information revolution has renewed old philosophical problems and posed new, pressing ones. This is what my book is about, yet this is what Searle's review entirely failed to grasp.

I suspect Singularitarians and Altheists 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: "don't mind them, but look and pass." For the world needs some good philosophy and we need to take care of serious and pressing problems.

NOTES

1. "Turkzilla!"*The Economist*.
2. See some hilarious ones in Pogue, "Use It Better," and Cracked Readers.
3. Matt Warman, "Stephen Hawking Tells Google 'Philosophy Is Dead."

4. Robert Herritt, "Google's Philosopher."
5. https://www.youtube.com/watch?v=3Ox4EMFMy48
6. Luciano Floridi, Mariarosario Taddeo, and Matteo Turilli, "Turing's Imitation Game."
7. http://www.bbc.co.uk/programmes/p02hvcjm
8. John R. Searle, "What Your Computer Can't Know."
9. The reader interested in a short presentation of what I mean by informational realism may wish to consult Floridi, "Informational Realism." For a full articulation and defense, see Floridi, *The Philosophy of Information*.
10. Floridi, "Response to NYROB Review."
11. Floridi, *The Onlife Manifesto*.

BIBLIOGRAPHY

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

Floridi, Luciano. "Response to NYROB Review." *The New York Review of Books*, November 20, 2014. http://www.nybooks.com/articles/archives/2014/dec/18/information-desk/.

Floridi, Luciano. 2003. "Informational Realism." Selected papers from conference on Computers and Philosophy, volume 37.

Floridi, Luciano. *The Philosophy of Information*. Oxford: Oxford University Press, 2011.

Floridi, Luciano. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press, 2014a.

Floridi, Luciano, ed. *The Onlife Manifesto: Being Human in a Hyperconnected Era*. New York: Springer, 2014b.

Floridi, Luciano, Mariarosaria Taddeo, and Matteo Turilli. "Turing's Imitation Game: Still a Challenge for Any Machine and Some Judges." *Minds and Machines* 19, no. 1 (2009): 145–50.

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

Pogue, David. "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," January 18, 2012. http://www.scientificamerican.com/article/pogue-all-time-worst-tech-predictions/.

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

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

Turing, A. M. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–60.

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

First-Person Consciousness as Hardware

Peter Boltuc

UNIVERSITY OF ILLINOIS SPRINGFIELD AND AUSTRALIAN NATIONAL UNIVERSITY

INTRODUCTION

I take the paradigmatic case of first-person consciousness to be when a nurse says that a patient regained consciousness after surgery. The patient does not need to have memory or other advanced cognitive functions. But she is *online*, so to say—we have good reasons to believe that the question *what it is like* for her to be is not empty.

Advanced cognitive architectures, such as LIDA, approach the functional threshold of consciousness. Such software performs advanced cognitive functions of all kinds, including image making and manipulation, advanced memory organization and retrieval, communication (including semantic structures), perception (that includes responses to color, temperature, and other *qualia*), and even creativity (e.g., imagitrons). Some AI experts believe that, at a certain threshold, adding further cognitive functions would result in first-person consciousness. Non-reductivists claim that the latter would emerge based on an informationally rich emergence base. Reductivists claim that such a rich information processing structure just *is* consciousness, that there is no further fact of any kind. I disagree with both claims.

The kind of first-person consciousness in the example of a patient regaining consciousness is analogous to a stream of light—it is not information processing of some advanced sort. Just like light bulbs are pieces of hardware, so are the parts of animal brain that create first-person consciousness.[1] Every object can be described as information (Floridi) and is in principle programmable (a physical interpretation of Church-Turing thesis), but this does not make every object in the universe a piece of software. The thesis of this paper is that first-person consciousness is more analogous to a piece of hardware, a light emitting bulb, than to software. There are probably information processing thresholds below which first-person consciousness cannot function (just like a bulb cannot emit light if not hooked up to the source of electricity), but no amount of information processing, no cognitive function, shall produce first-person consciousness without such consciousness emitting a piece of hardware.

This claim follows from the so-called *engineering thesis*, the idea that if first-person consciousness is a natural process it needs to be replicable in robots. Instituting such functionality in machines would require a special piece of hardware, slightly analogous to the projector of holograms. On the other hand, human cognitive functions can be executed in a number of cognitive architectures.[2] Such architectures do not need to be hooked up to the *light-bulb-style* first-person consciousness. This last claim opens the issue of *philosophical zombies* and epiphenomenalism. On both issues I try to keep the course between Scylla and Charybdis presented by the most common alternatives.