Would Super-Human Machine Intelligence Really Be Super-Human?

Philip Larrey

Abstract Given recent advances in the field of artificial intelligence, the notion of creating a digital machine capable of not only logical operations, but also of complex inferences and other processes usually associated with human thought, must now be considered. This chapter attempts to provide a speculative basis for such a consideration. The chapter attempts to defend the position that although machine "intelligence" will always differ from human intelligence in nature, it will exceed human intelligence in significant ways that will require a serious and profound reflection on the meaning of thought itself.

1 Introduction

Recently, several renowned personalities have weighed in on the theme of "super-human artificial intelligence" in the popular press. The famous astrophysicist from Britain, Stephen Hawking, thinks that artificial intelligence could end mankind[1]; the founder of Tesla, CEO of SpaceX and Silicon Valley guru, Elon Musk, warned us at MIT that artificial intelligence is our biggest existential threat[2]; a meeting of the minds took place at the beginning of 2016 in Puerto Rico called: "The Future of AI: Opportunities and Challenges",[3] and James Barrat worries us with the very title of his recent work, *Our Final Invention.*[4]

[1]Cf. http://www.bbc.com/news/technology-30290540.
[2]Cf. http://webcast.amps.ms.mit.edu/fall2014/AeroAstro/index-Fri-PM.html. Centennial Symposium, at the 1:07:26 mark.
[3]Cf. their website: http://futureoflife.org/misc/ai_conference. Organized by the Future of Life Institute (Boston), many of the key people developing artificial intelligence attended.
[4]Cf. James Barrat, *Our Final Invention. Artificial Intelligence and the End of the Human Era*, St. Martin's Press, New York 2013.

P. Larrey (✉)
Pontifical Lateran University, Vatican City, Italy
e-mail: plarrey@uni.net

G. Dodig-Crnkovic and R. Giovagnoli (eds.), *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines*, Studies in Applied Philosophy, Epistemology and Rational Ethics 28, DOI 10.1007/978-3-319-43784-2_19

Aside from the apocalyptic scenarios which make for very good sci-fi films, the issue of super-human artificial intelligence is now on contemporary man's table. Specialists in the field have known for some time that in a very real way, we are sharing existence with other types of non-human intelligence. Aristotle understood that animals are intelligent (some exhibit more intelligence than others)[5] and that they also had 'souls', in the sense of a life principle. Thus, we have the term 'animate objects' connoting those objects which are alive, i.e. contain a soul (*anima* in Latin). The difference between animal intelligence and human intelligence is due to the degree of *being* (*esse* in Latin) of the vital principle in humans: the life principle ('form') in humans is capable of operations that exceed the potentiality of the body ('matter'), and therefore is capable of existing without the body, whereas the forms of animals cannot exist without the body. Such was Aristotle's philosophical argument in favor of the immortality of the human soul.

Aristotle also postulated the existence of purely spiritual beings, as forms which have no material substratum. These are the pure forms that inhabit the celestial domain and are responsible for the motion of the heavenly bodies. Thomas Aquinas will call these pure forms 'angels' and will conclude that they are not only intelligent, but also much more intelligent than human beings because they do not need to turn to the senses in order to possess knowledge: they 'know' by virtue of their essence.

With the birth of computers operating on binary logical systems, the term 'artificial intelligence' was coined, and was meant to connote logical operations achieved by software programs running on silicon chips. For a while, it was fashionable in philosophical circles (especially in the cognitive sciences) to conceive the relationship between the mind and the brain as similar to the relationship between software and hardware: the brain acts like a hard drive for the mind's software (read: program). Hilary Putnam called this *functionalism*: a view that he once held and later abandoned (as have almost all philosophers). It was the American philosopher from UC Berkeley, John Searle, who devised the very useful distinction between *strong AI* and *soft AI*, in order to draw clearer boundaries between what the human intellect does and what computers do.[6]

Perhaps it was unfortunate that computer engineers and philosophers called what a binary computer achieves (applying concepts created by the great British mathematician, Alan Turing) 'intelligence', albeit 'artificial'. Yet, the name has stuck. Such was the rationale behind the term "super-human machine intelligence" to connote the future evolution of AI which, it is assumed, will surpass or exceed human intelligence. How we get to this level of super-human intelligence is usually explained through a series of extrapolations, starting with what AI is capable of doing now, and assuming that as computers get faster, more powerful and cheaper

[5]The "smartest" animals in the animal kingdom may in fact not turn out to be the closest biologically to humans, but rather birds. Cf. Noah Strycker "In almost any realm of bird behavior—reproduction, populations, movements, daily rhythms, communication, navigation, intelligence, and so on—there are deep and meaningful parallels with our own", xii.

[6]John R. Searle, *Mind. A Brief Introduction*, Oxford University Press, 2004 65 ff.

(thanks to Moore's Law of accelerating returns), they will eventually achieve "super-human level intelligence".

Everyone knows that it is easy to extrapolate. The more difficult question is simply this: what is intelligence, and what is *human* intelligence? This more profound question has perplexed philosophers and non-philosophers for as long as we can remember.

A very opportune place to begin to address such a question in this context is with Nick Bostrom, from Oxford University.[7] In 2014, he published his very insightful work, *Superintelligence. Paths, Dangers, Strategies.*[8] In chapter three of that book, Bostrom identifies "Forms of superintelligence" and opens the chapter with this startling affirmation: "We also show that the potential for intelligence in a machine substrate is vastly greater than in a biological substrate. Machines have a number of fundamental advantages which will give them overwhelming superiority. Biological humans, even if enhanced, will be outclassed" (52). So, just what are the 'forms' that such a superintelligence could take?

1. Speed superintelligence. "The simplest example of speed superintelligence would be a whole brain emulation running on fast hardware. An emulation operating at a speed of ten thousand times that of a biological brain would be able to read a book in a few seconds and write a PhD thesis in an afternoon. With the speedup factor of a million, an emulation could accomplish an entire millennium of intellectual work in one working day" (53). This type of super-human capacity is easy to extrapolate by simply doing the math on speed. Yet it assumes that a super-human level of intelligence can be achieved by speeding everything up. If intelligence is measured by speed, then of course it is obvious that if an artificial intelligence can do it faster than biology-based intelligence, it is by definition 'super-human'. Yet there are underlying presuppositions which are controversial.

Simply emulating a human brain will not necessarily produce intelligence, and many AI experts agree because we do not understand how the brain produces intelligence (and much less how the brain would 'cause' consciousness—if, in fact, it does). Ben Goertzel, the founder and leading intellectual at the Open Cog Project (perhaps the best known group specifically dedicated to the development of artificial *general* intelligence), states the following: "My current feeling is that brain emulation won't be the fastest or best approach to creating human-level AGI. One 'minor problem' with this approach is that we don't really understand how the brain works yet, because our tools for measuring the brain are still pretty crude. Even our theoretical models of what we should be measuring in the first place are still hotly debated".[9] Although we have made much progress in the speed of computers (and more will come), it is clear that for general intelligence, speed is not a panacea.

[7]At Oxford, Nick Bostrom is Director of the Future of Humanity Institute and has recently received $10 million from Elon Musk, who expressed deep interest in such research.

[8]Nick Bostrom, *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, 2014.

[9]Ben Goertzel, *Ten Years To the Singularity If We Really, Really Try … and Other Essays on AGI and Its Implications*, CreateSpace Independent Publishing Platform, 2014, 112.

For very limited and narrow AI uses such as playing chess, data mining and hugely powerful search engines, fast computers are extremely important and they already are better at accomplishing their tasks than human beings. This comes as no surprise. But from these applications to extrapolate to a superintelligence is unwarranted.

The problem is really philosophical in nature.

Let us look at the two specific examples which Bostrom offers: reading a book and writing a PhD thesis. If 'reading' a book means digitizing the content and having it reside in some sort of memory (like RAM or on a hard drive), then computers can already do this and very quickly. But Bostrom knows that the real issue is deeper. The assumption is that a computer simulation of a human brain would in fact do everything the brain does, i.e. read and write a PhD thesis. Yet there is a decisive difference between *understanding* something and *simulating an understanding* of something. With the advent of the 'semantic web', sufficiently fast computers with proper software are going to achieve this simulation of *understanding meaning*, yet they will not really understand meaning.

John Searle's Chinese Room thought experiment is very illustrating in this sense.[10] The hypothetical scenario is a man in a room who does not speak Chinese. Chinese speakers outside the room slide slips of paper with Chinese characters on them, asking questions to the man inside. The man inside then consults a series of rule-books which indicate to him which Chinese characters he must write down on slips of paper to properly answer the question … in Chinese, which of course he does not understand. The Chinese speakers on the outside receive the slips of paper and are convinced that they were written by a Chinese speaker.

The man in the room does not understand Chinese at all, he has no idea what the characters *mean*, but he uses the rule-books to answer the questions. This is exactly what a computer is achieving. Searle concludes: "[T]he implemented syntactical or formal program of a computer is not constitutive of nor otherwise sufficient to guarantee the presence of semantic content; and secondly, simulation is not duplication".[11] To drive home the point, he also recalls his famous example of digestion: a commercial computer can certainly simulate the digestive process that happens in the body, but it is not really digesting anything. There is a big difference.

Although the Chinese Room experiment is quite dated today, Searle believes it is still valid as describing the essential difference between artificial intelligence and human intelligence. "My reason for having so much confidence that the basic argument is sound is that in the past 21 years I have not seen anything to shake its fundamental thesis. The fundamental claim is that the purely formal or abstract or syntactical processes of the implemented computer program could not by themselves be sufficient to *guarantee* the presence of mental content or semantic content

[10]Cf. the insightful work edited by John Preston and Mark Bishop, *Views into the Chinese Room. New Essays on Searle and Artificial Intelligence*, Oxford 2002.

[11]John Searle, "Twenty-One Years in the Chinese Room", in *Views into the Chinese Room*, *cit*., 52.

of the sort that is essential to human cognition".[12] Could the brain emulation trick a human observer into believing that it is, in fact, *understanding* the meaning in the text? The short answer, I believe, is yes, at least for a sufficiently fast computer with the proper software. At this point, we would have a machine that successfully passes the Turing Test (which to date–no computer has yet achieved, even though there have been news-worthy attempts[13]).

However, as Searle concludes, "The 'system', whether me in the Chinese Room, the whole room, or a commercial computer, passes the Turing Test for understanding Chinese but it does not understand Chinese, because it has no way of attaching any meaning to the Chinese symbols. The appearance of understanding is an illusion".[14] Returning to Bostrom's examples, we would be assuming that "reading a book" or "completing a PhD dissertation" implies *understanding*, that very subtle, complex cognitive activity unique to human beings. On just about any comprehension of a theory of meaning, the human cognitive process captures meanings, evaluates them, compares and contrasts them and interprets. This is why a good text to read is not simply the product of rote memory or the repetition of things already stated. It goes much further: it implies a cognitive activity that *understands* and advances understanding in some significant way. The human intellect is capable of this kind of activity because the human being is *conscious*, it is aware and even further it is *self-aware*.

David Chalmers, when he was teaching in Arizona in 1994, called this the "hard problem of consciousness", and still today leaders in cognitive sciences do not seem to have progressed very much.[15] Perhaps one reason why progress does not seem to occur in this field is due to a philosophical option: that of reductionism. Reductionism, simply stated, proposes that consciousness is *reducible* to brain states (patterns of neurons firing and synapses exchanging information), and that the brain *causes* consciousness. If reductionism turns out to be the correct assumption, then we should be able to solve the 'hard problem' with more sophisticated technology and software.[16] As Searle states, "[t]he point, however, is that any such artificial machine would have to be able to duplicate, and not merely simulate, the causal powers of the original biological machine … An artificial brain would have to do something more than simulate consciousness, it would have to be able to *produce* consciousness. It would have to cause consciousness".[17] "In order to create

[12]*Ibid*., 51.

[13]Cf. for example, a computer program called Eugene Goostman: http://www.bbc.com/news/technology-27762088. Most specialists in the AI field contested the published results.

[14]John Searle, *cit*., 61.

[15]Cf. Oliver Burkeman's insightful article in the *Guardian* quite recently: "Why can't the world's greatest minds solve the mystery of consciousness?", http://www.theguardian.com/science/2015/jan/21/-sp-why-cant-worlds-greatest-minds-solve-mystery-consciousness.

[16]Such would seem to be the goal of Ray Kurzweil who now works for Google as head of engineering. Cf. his recent work, *How to Create a Mind. The Secrets of Human Thoughts Revealed*, Penguin Books, 2013.

[17]John Searle, *cit*., 56.

consciousness you have to create mechanisms which can duplicate and not merely simulate the capacity of the brain to create consciousness".[18] Here, of course, Searle is assuming that the brain *causes* consciousness, which may or may not be true. The philosophical jury is still out on the issue. According to one long-standing tradition in philosophy, the brain 'houses' consciousness, but consciousness itself would be caused by the soul, or by the principle of being which gives existence to the subject.[19] As stated above, such a principle resides in all animate objects (which are composed of form and matter), and therefore on such an assumption, we can attribute consciousness also to non-human life forms (such as birds, dogs and cats). Ask any owner of a dog if their pet is conscious, and the answer will be obviously yes. The level of consciousness would be less than for humans, yet it would be present nonetheless.

One of the most impressive demonstrations of computer generated 'intelligence' from an historical view point was IBM's *Deep Blue*, which beat the world's number one Grand Master of chess, Gary Kasparov on May 11, 1997: a feat that many at the time considered impossible. Although the event was historical in many senses, *Deep Blue* still had not given evidence of "super-human intelligence": it was simply better at playing chess. "In the case of Deep Blue, the machine did not know that it was playing chess, evaluating possible moves, or even winning and losing. It did not know any of these things, because it does not know anything".[20] Ordinary computer chess programs on your laptop now reach *Deep Blue* levels at playing chess. Yet we would generally not say that the program "knows how to play chess".

Another IBM experiment recently challenged our conception of artificial intelligence, namely the super computer called *Watson*, which in 2011 defeated the two most successful players of *Jeopardy!* and was awarded a million dollars. The intriguing element here is that in the quiz show, one must come up with the questions to the answers which are given. It would seem that being able to achieve this would require the machine to *understand* human language. When faced with

[18]John Searle, *cit*., 68.

[19]Cf. Thomas Aquinas, *Summa Theologica*, Q. 75, art. 2: "Therefore, the intellectual principle, which we call the mind or the intellect, has an operation in which the body does not share. Now only that which subsists in itself can have an operation in itself … We must conclude, therefore, that the human soul, which is called intellect or mind, is something incorporeal and subsistent." Also, "Now it is clear that the first thing by which the body lives is the soul. And as life appears through various operations in different degrees of living things, that whereby we primarily perform each of all these vital actions is the soul. For the soul is the primary principle of our nourishment, sensation, and local movement; and likewise of our understanding. Therefore this principle by which we primarily understand, whether it be called the intellect or the intellectual soul, is the form of the body. This is the demonstration used by Aristotle (*De Anima* ii, 2). But if anyone says that the intellectual soul is not the form of the body he must first explain how it is that this action of understanding is the action of this particular man; *for each one is conscious that it is himself who understands.* But one cannot sense without a body: therefore the body must be some part of man. It follows therefore that the intellect by which Socrates understands is a part of Socrates, so that in some way it is united to the body of Socrates." Id., *ST*, Q. 76, art. 1. My emphasis.

[20]John Searle, *cit*., 65.

this clue given during competition: "A long tiresome speech delivered by a frothy pie topping", *Watson* came up with the correct question: "What is meringue harangue?" This was quite impressive. The builders of *Watson* even admitted that they are not sure *how* the machine arrived at the proper answers, given the various subroutines operating within the software program. Yet again, using Searle's distinction, one can still suggest that *Watson* is simulating having understood, yet it really does not understand anything. In no way does this diminish the amazing capability of the machine, which is now being used at the Memorial Sloan Kettering Cancer Center in Manhattan to help diagnose cancer in patients. It has been reported that 90% of the nurses concur with its analysis. It simply tells us that there is a difference between what *Watson* does and what the human intellect does.

It also serves as a warning, already mentioned by Bostrom: machines are better than human beings in many cognitive functions, and they are getting even better, even without solving the "hard problem of consciousness". He concludes: "Although these systems [such as *Watson*] do not understand what they read in the same sense or to the same extent as a human does, they can nevertheless extract significant amounts of information from natural language and use that information to make simple inferences and answer questions. They can also learn from experience".[21]

The second type of possible superintelligence analyzed by Bostrom is called 'Collective superintelligence', and consists of "[a] system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system".[22] From a theoretical point of view, this second type of superintelligence does not differ radically from the first type: instead of only one, there are many, and they collaborate with each other in order to solve problems. Within the field of AI research, this is a plausible outcome of the many actors who are developing general super-human artificial intelligence.[23] However, because of the different platforms currently being developed, it is not clear how 'separate' intelligent machines would communicate harmoniously with each other. As an example, much development is being carried out in the field of quantum computing, using qubits instead of binary bits, capable of housing information in superimposed states (and not simply as ones and zeros). A working prototype is already being used at the NASA Ames Research center, sponsored in part by Google, and is called *D-Wave Two*. The Chief Scientist there is Eric Ladizinsky, who is very articulate, and states that this machine is a thousand times more capable of computing than a traditional machine. It utilizes

[21]Bostrom, *cit*., 71.

[22]*Ibid*., 54.

[23]The current leader in the field of general (as opposed to specific) AI is probably *DeepMind*, owned by Google. There may be covert programs in different parts of the world, about which we know little or nothing. However, with the enormous resources at their disposal, Google is positioned as the likely leader in the race to produce a general super-human level of artificial intelligence. In his book presentation at UC Berkeley several months ago, Bostrom concurred that Google would likely be the first to create true superintelligence.

quantum states on a macroscopic level. This development will probably ensure that Moore's Law will continue to function into the foreseeable future.

As an interesting note, Bostrom admits: "nothing in our definition of collective superintelligence implies that a society with greater collective intelligence is necessarily better off. The definition does not even imply that the more collectively intelligent society is *wiser*".[24] This is a non-obvious truth that must be highlighted. Human history is replete with episodes of very intelligent people making very poor decisions and creating very harsh conditions for millions of people. Yet, the point refers to intelligence in general, and specifically to the ability to solve cognitive problems: not to construct a better society for human beings, laudable as that goal is.

The third and final form of superintelligence is called 'Quality superintelligence' and consists in "[a] system that is at least as fast as a human mind and vastly qualitatively smarter".[25] Here the difficulty is to define the term 'qualitatively' when referring to cognitive activity. Bostrom offers some examples to make the point. The first consists of non-human animal intelligence, which we know exists and which is 'qualitatively' inferior to human intelligence (and which has already been mentioned above, for example the case of bird intelligence). Interestingly, "[i]n terms of raw computational power, human brains are probably inferior to those of some large animals, including elephants and whales. And although humanity's complex technological civilization would be impossible without our massive advantage in collective intelligence, not all distinctly human cognitive capabilities depend on collective intelligence".[26] Thus, human intelligence is qualitatively superior to animal intelligence, and not because of pure computational power. In fact, a very important factor is often neglected when referring to computational power and that is computational architecture. The human brain has a more complex computational architecture than other animal brains, that might be actually much larger. In this sense, a superintelligence would need to exhibit intelligence qualitatively superior to humans.

Another example given deals with human intelligence's capability for complex linguistic representations, which gives humans an enormous evolutionary advantage over nonhumans. Most probably, linguistic capabilities were developed for communicative purposes. Thus, linguistic skills are part of human collective intelligence. Humans were constituted with the cognitive modules that enable linguistic representations and therefore became superior to the brutes. Furthermore, "were we to *gain* some new set of modules giving an advantage comparable to that of being able to form complex linguistic representations, we would become superintelligent".[27] The implication here is that were a machine to gain a similar set of modules, it would be considered 'super-human'.

[24] *Ibid*., 55.
[25] *Ibid*., 56.
[26] *Ibid*., 57.
[27] *Ibid*., 57.

It is not clear in what a similar set of modules would consist, in terms of cognitive activity. In many respects, certain machines are already 'super-human' in terms of brute strength (we can think of the large industrial machines or those used in agriculture, which have replaced millions of people in recent decades), speed, precision, sensing ability (like military-grade satellites which can 'see' hundreds of miles away or in the dark), etc. It would seem logical to eventually add 'cognition' to the list of things which machines do better than we do. And, as already mentioned, in some respects they already out perform us in many, isolated tasks, from the perspective of cognitive activity. Yet, to achieve *general* super-human intelligence, something more will be needed. And a sufficiently complex and powerful computer just may be able to come up with those 'extra modules' necessary to give the machine a cognitive advantage over humans. Ben Goertzel states: "As every software engineer knows, the design and implementation of complex software is a process that constantly pushes against the limitations of the human brain—such as our limited short term memory capacity, which doesn't allow us to simultaneously manage the states of more than a few dozen variables or software objects. There seems little doubt that a human-level AGI, once trained in computer science, would be able to analyze and refine its own underlying algorithms with a far greater effectiveness than any human being".[28]

2 Common Sense

One area in which artificial intelligence systems have yet to make much progress is something so natural and so spontaneous that all humans exercise effortlessly, i.e., common sense. Although at times it is said that 'common sense' is not very common (because of the insane behavior of which many human beings are capable every day, which Erasmus writes of in his *The Praise of Folly*), this characteristic of human cognition is extremely difficult to replicate in a digital computer. With his groundbreaking paper, "Programs with Common Sense",[29] John McCarthy ushered in the era of attempting to formalize common sense knowledge so as to be used by a digital computer. "We shall therefore say that *a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows*".[30] In that paper, McCarthy described the theoretical basis for a program to 'learn' from experience and from

[28]Ben Goertzel, *Ten Years To the Singularity If We Really, Really Try … and other Essays on AGI and its Implications*, CreateSpace Independent Publishing Platform, 2014, 19. Cf. the recent article in the *New Scientist* which analyzes exactly this type of technology: http://www.newscientist.com/article/mg22429932.200-computer-with-humanlike-learning-will-program-itself.html#.VMpnW9KG9vA.

[29]Cf. McCarthy, J., "Programs with Common Sense". *Proc. of Conference on the Mechanization of Thought Processes*, 1959, 75–91.

[30]*Ibid*., 78. Italics in original.

the world as such. And he further stipulated that the mechanism through which the program 'learns' would also have to be *improvable*, i.e., as more and more information becomes available and assimilated, the way in which the program utilizes it has to become more complex and applicable. This was a profound insight, and has led to what is commonly known as 'machine learning' which is found in many software applications in use by Amazon, Netflix, Google and Facebook. Of course, the program is not 'learning' anything, for it is simply calculating (and recalculating) relationships based on acquired patterns in order to produce suggestions about what books or films you may like, or what kind of advertising would be best suited for you. Such software has been described as 'creepy', precisely because it gives you the impression that it is 'learning' on the basis of information given by you.

Producing authentic common sense in a digital format has proven to be quite elusive. There have been efforts in this direction, yet perhaps the most serious difficulty arises from the fact that common sense reasoning is not reducible to sheer logic (for programs are much better than humans from a strictly logical point of view). That is the whole point about the uniqueness of common sense knowledge: it is not based simply on laws or logical inferences. Another difficulty arises from the ambiguity of just what is common sense knowledge. What do we mean by common sense knowledge?

In his work, Antonio Livi has attempted to provide some clarification.[31] However, his primary interest lies with the relationship between common sense knowledge and belief in a divine being, which of course does not seem too applicable to a software program (although one could make the case for its importance). Thomas Reid was the 'grandfather' of the Scottish school of common sense, attempting to oppose David Hume's skepticism, founded on assumptions carried over from John Locke's notion of 'ideal theory'. Between the two world wars, Cambridge saw a flourishing of the philosophy of common sense,[32] and, more recently, Noah Lemos has returned to the question in his *Common Sense. A Contemporary Defense.*[33]

Seeing how important common sense knowledge is for human beings, a crucial task for AI engineers working on superintelligence would necessarily imply giving such an ability to machines. Can machines *simulate* common sense knowledge? To answer this question, a group of intellectuals have come up with the Winograd Schema Challenge, named after Terry Winograd, which would replace the Turing Test in order to assess the ability of a digital computer to pass as a human.[34] Every 2 years, a major convention is held to allow participants to try and win the

[31]Cf. Antonio Livi, *A Philosophy of Common Sense. The Modern Discovery of the Epistemic Foundations of Science and Belief*, trans. Peter Waymel, Davies Group Publishers, Aurora, 2013.

[32]Cf. John Coates, *The Claims of Common Sense. Moore, Wittgenstein, Keynes and the Social Sciences*, Cambridge University Press, 1996/2001.

[33]Noah Lemos, *Common Sense. A Contemporary Defense*, Cambridge University Press, 2004/2010.

[34]Cf. http://commonsensereasoning.org/winograd.html.

challenge. So far, no one has been able to. Let us look at one of the examples that has been set forth in order to test common sense knowledge in a computer. The machine would be asked the following question and would have to answer either 0 or 1, with the option of using either 'big' or 'small': "The trophy would not fit in the brown suitcase because it was too big (*small*). What was too big (*small*)? Answer 0: the trophy. Answer 1: the suitcase". For a human intellect, it is clear that if 'big' is used as the main adjective, then it refers to the trophy (answer 0); whereas if the preferred adjective is 'small', then it refers to the suitcase (answer 1). According to the promoters of the challenge, this is an example of common sense knowledge that humans routinely and effortlessly achieve, and AI has yet to be able to tackle. A human who answers these questions correctly typically uses his abilities in spatial reasoning, his knowledge about the typical sizes of objects, as well as other types of common sense reasoning, to determine the correct answer.[35] Such abilities are common to all conscious humans who do not suffer from some sort of mental disability. Computers (at present) lack such abilities. Common sense knowledge is vital for humans to be able to interact in the world, and without it, people would surely die (and quickly). "Standing in front of a speeding train" goes against common sense (unless of course one wanted to commit suicide), and one does not need to experience the effects of doing so before concluding that it would be unwise. This shows one of the special characteristics of common sense: it is not the result of experience (or trial and error), but rather it is connatural with human intelligence.

Some software engineers are attempting to 'teach' a computer common sense knowledge by entering lists of common sense statements. Perhaps the most successful thus far has been the MIT Media Lab's project called Open Mind Common Sense (OMCS) which uses their ConceptNet as an engine to make connections among the millions of common sense phrases that have been introduced by more than 15,000 participants. The project is directed by Catherine Havasi, one of the original founders who worked with Marvin Minsky.

The efforts of the people at Media Lab may prove to be successful at having an AI *simulate* common sense knowledge, yet the AI will still not possess common sense. One might respond by saying that common sense is not necessary for AI to achieve excellent results in interacting in the real world, and this would be an important distinction to bear in mind. But it also shows the limitations of AI in the real world which must utilize formal models of the world in order to interact with our reality. Up until now, human programmers have furnished these formalized models to the AI and have been able to achieve remarkable results.[36] It does seem clear that the human intellect does not use formalized models in the same way as AI does. The human intellect has direct contact with reality and objects of medium

[35] http://commonsensereasoning.org/winograd.html.

[36] It will be interesting to see if a sufficiently advanced AI will become capable of coming up with its own formalized models of reality and use those instead of the ones provided by human programmers. This is perhaps one of the goals of the DeepMind project, directed by Demis Hassabis and located in London.

range (i.e. those not too large—like planets—and those not too small—like atoms, the knowledge of which requires special tools and analysis) and through a process often known as 'abstraction', it is able to *understand the nature of things*. From a philosophical perspective, our knowledge of reality begins with common sense, for it is the inescapable initial relationship between our minds and reality, and it is one that is not learned through experience but rather is 'hardwired' in us.

The renowned historian of science, Steven Shapin, goes even further and claims that even the best scientific knowledge begins with common sense. He states: "In the 1850s, T.H. Huxley wrote that 'Science is, I believe, nothing but trained and organized common sense'. The whole of science, according to Albert Einstein, 'is nothing more than a refinement of every day thinking'. Max Planck agreed: 'Scientific reasoning does not differ from ordinary reasoning in kind, but merely in degree of refinement and accuracy'. And so did J. Robert Oppenheimer: 'Science is based on common sense; it cannot contradict it'."[37] Therefore, the basis of our most sophisticated scientific reasoning is actually something ordinary and practiced by all human beings, common sense. Proverbs such as "A stitch in time saves nine", or "Great oaks from little acorns grow", or "Stolen apples are sweetest" are phrases which contain great knowledge, expressed in simple form and applicable to the world at large. Such propositions would be difficult for an AI to harness, in part because they are not *always* true (albeit almost always) and in part because they are phrases which are applicable to reality through the intermediacy of the human thinker who captures the proverbial meaning of the sentence and applies it to a concrete situation (i.e. when debating whether to sew a sock or wait a couple of more days to do so).

Strangely enough, perhaps it will be precisely that type of knowledge which we as human beings take so much for granted that will ensure our uniqueness and importance in cognitive activity when seriously challenged by AI. "Many people suppose that computing machines are replacements for intelligence and have cut down the need for original thought", Norbert Wiener once wrote. "This is not the case".[38] "The more powerful the computer, the greater the premium that will be placed on connecting it with imaginative, creative, high-level human thinking".[39] Instead of the popular scenario of 'us versus them', the more probable outcome of advanced AI systems will be one of collaboration, where each type of 'intelligence' is able to maximize its own qualities in order to achieve its goals. This also makes sense from a business point of view. Stephen F. DeAngeli, President and CEO of the cognitive computing firm *Enterra Solutions* writes: "Although concerns remain that intelligent computers will continue to put workers out on the street, we believe

[37]Steven Shapin, *Never Pure. Historical Studies of Science as If It Was Produced by People with Bodies, Situated in Time, Space, Culture, and Society, and Struggling for Credibility and Authority*, The Johns Hopkins University Press, Baltimore 2014, 349–350.

[38]Norbert Wiener, "A Scientist's Dilemma in a Materialistic World" (1957), in *Collected Works*, vol. 4 (MIT Press, 1984), 709.

[39]Walter Isaacson, *The Innovators. How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution*, Simon & Schuster, New York 2014, 222.

that computers working with (not in place of) humans creates the most effective, efficient, and profitable working environment".[40]In his book on the protagonists of the digital era, Walter Isaacson shows complete agreement: "These ideas formed the basis for one of the most influential papers in the history of postwar technology, titled 'Man–Computer Symbiosis', which Licklider published in 1960. 'The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly', he wrote, 'and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today'. This sentence bears rereading, because it became one of the seminal concepts of the digital age".[41]

If at such a point we call this type of intelligence 'super-human', then the answer to the question that this paper asked at the beginning, "Would Super-human-machine intelligence really be super-human?" would be yes. However, if the intelligence doing the thinking is human, the answer would be no, simply because it is still human. Perhaps we will choose to refer to such intelligence as 'augmented', and yet perhaps not. For centuries we have used telescopes and microscopes to augment our knowledge, yet we usually do not refer to such knowledge as 'super-human'. In any event, as AI continues to develop, it will be fascinating to see what happens. Ours is truly an 'unknown future'.

References

1. Aquinas, T: Summa Theologica, pp. 75–76
2. Barrat, J.: Our Final Invention: Artificial Intelligence and the End of the Human Era. St. Martin's Press, New York (2013)
3. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
4. Burkeman, O.: Why can't the world's greatest minds solve the mystery of consciousness?. http://www.theguardian.com/science/2015/jan/21/-sp-why-cant-worlds-greatest-minds-solve-mystery-consciousness
5. Coates, J.: The Claims of Common Sense: Moore, Wittgenstein, Keynes and the Social Sciences. Cambridge University Press (1996/2001)
6. Goertzel, B.: Ten Years To the Singularity If We Really, Really Try … and Other Essays on AGI and Its Implications. CreateSpace Independent Publishing Platform (2014)
7. Isaacson, W.: The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution. Simon & Schuster, New York (2014)
8. Kurzweil, R.: How to Create a Mind: The Secrets of Human Thoughts Revealed. Penguin Books (2013)
9. Lemos, N.: Common Sense: A Contemporary Defense. Cambridge University Press (2004/2010)

[40]http://innovationinsights.wired.com/insights/2014/08/ai-systems-will-prove-useful-long-become-self-aware/.

[41]Walter Isaacson, *cit.*, 226.

10. Livi, A.: A Philosophy of Common Sense: The Modern Discovery of the Epistemic Foundations of Science and Belief, translated by Peter Waymel. Davies Group Publishers, Aurora (2013)
11. McCarthy, J.: Programs with Common Sense. In: Proceedings of Conference on the Mechanization of Thought Processes, pp. 75–91 (1959)
12. Preston, J., Bishop, M.: Views into the Chinese Room: New Essays on Searle and Artificial Intelligence, Oxford (2002)
13. Searle, J.R.: Mind: A Brief Introduction. Oxford University Press (2004)
14. Shapin, S.: Never Pure: Historical Studies of Science as if It Was Produced by People with Bodies, Situated in Time, Space, Culture, and Society, and Struggling for Credibility and Authority. The Johns Hopkins University Press, Baltimore (2010)
15. Strycker, N.: The Thing with Feathers: The Surprising Lives of Birds and What They Reveal About Being Human. Riverhead Books (2014)
16. Wiener, N.: A Scientist's Dilemma in a Materialistic World (1957). in Collected Works, vol. 4. MIT Press (1984)