AI & SOCIETY
https://doi.org/10.1007/s00146-019-00880-4

OPEN FORUM

Artificial intelligence: consciousness and conscience

Gunter Meissner[1,2,3]

Received: 18 December 2018 / Accepted: 12 February 2019

Abstract
Our society is in the middle of the AI revolution. We discuss several applications of AI, in particular medical causality, where deep-learning neural networks screen through big data bases, extracting associations between a patient's condition and possible causes. While beneficial in medicine, several questionable AI trading strategies have emerged in finance. Though advantages in many aspects of our lives, serious threats of AI exist. We suggest several regulatory measures to reduce these threats. We further discuss whether 'full AI robots' should be programmed with a virtual consciousness and conscience. While this would reduce AI threats via motivational control, other threats such as the desire for AI—human socioeconomic equality could prove detrimental.

Keywords Artificial intelligence · Deep-learning neural networks · Social robots · Consciousness · Conscience

1 Introduction

Through time, human evolution has been shaped by technological inventions:

- The discovery of agriculture, which occurred independently in different geographical regions starting around 10,000 BC, allowed humans to end nomadism and settle down in a stable, protected environment. As a consequence, human population exploded.
- The steam engine, enhanced by James Watt in the eighteenth century, fueled the industrial revolution, which, however, led to huge wealth differences between entrepreneurs and exploited workers.
- The invention of the telegraph and later the telephone at the end of the nineteenth century led to the birth of the telecommunication industry.
- The invention of the internal combustion engine, improved by Nikolaus Otto in the late twentieth century, led to the creation of the car and airplane industry.
- The invention of the computer, popularized in the 1980s, allowed the simplification of writing and calculations, as well as the easy storage of data, which later led to 'big data' technology.
- The invention of the Internet, which became mainstream in the 1990s, allowed fast communication, as well as the easy availability of information and data. While virtually all Internet stocks crashed in the dot-com bubble bursting in 2000/2001, many Internet companies and the Internet technology survived the crash.
- Currently artificial intelligence is the society-transforming technology. We will discuss key applications of AI, benefits and threats, and suggest regulatory as well as motivational measures such as a virtual conscience to reduce the threats.

"The existence of intelligent life in the universe is extremely rare. Some say it has yet to be found on earth" (Stephen Hawking).

What is AI?

Many definitions of AI exist. We will define it in this paper as "The creation of intelligent machines". Where do we currently (2019) stand, where will we go? This can be expressed in a pyramid, as seen in Fig. 1.

✉ Gunter Meissner
meissner@hawaii.edu
http://www.dersoft.com
http://www.cassandracm.com

[1] University of Hawaii, 1288 Kapiolani Blvd #3404, Honolulu, HI 96814, USA

[2] MathFinance at Columbia University, New York, USA

[3] NYU, New York, USA

Published online: 25 February 2019

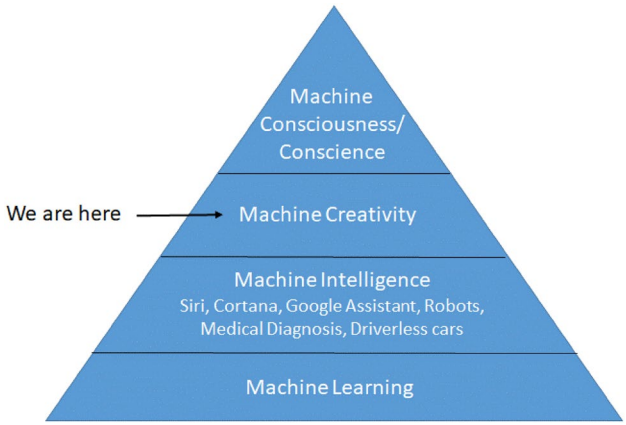

Fig. 1 Stages of machine evolution

Currently (2019), we are improving existing machine intelligence such as our virtual assistants, social robots, and driverless cars. In addition, we are building creative deep-learning neural networks, applied in many sciences, in particular in medical diagnoses and causality. Some of our virtual assistants and social robots already have some form of understanding of consciousness and conscience.

The remaining paper is structured as follows. In Sect. 2 we discuss the current applications of AI in our society, such as virtual assistants, neural networks, robo-butlers as well as AI in financial trading. In Sect. 3 we analyze the serious threats of AI and suggest how to address the threats. In Sect. 4 we discuss whether it is beneficial or detrimental to program consciousness and conscience into AI machines. Section 5 concludes.

2 Applications of AI

Before discussing consciousness and conscience for certain AI applications, let's first briefly review popular AI uses in our time (2019).

2.1 Virtual assistants

Virtual assistants, also called digital assistants, AI assistants or chatbots, are one of the most popular applications of AI. A virtual assistant is a software program which understands voice commands and performs tasks for users.

For a virtual assistant to help a user, three stages have to be successfully completed: (1) perception, (2) cognition, and (3) resolution. Perception is the phonetic understanding of the task. Cognition is the understanding of the content of the task. Resolution is the ability to find a solution for the task. For example, a user commands a virtual assistant: "Find a Burmese cuisine in the city". First the virtual assistant has to phonetically understand each word. Then the assistant has to understand the meaning of the command, in particular 'Burmese cuisine'. Last, the virtual assistant has to be able to access a database with restaurants in the city.

Many virtual assistants exist: Apple's Siri was one of the first, introduced to the iPhone 4S in 2011, to be followed by Microsoft's Cortana in 2014, and Google's assistant in 2016, which developed from earlier versions of Good Now. Amazon's Alexa, which is integrated into Amazon's smart speaker system Echo, was introduced in 2015. Samsung launched its virtual assistant Bixby in 2017. We will discuss whether these virtual assistant have a form of understanding of consciousness and conscience in Sect. 4.

2.2 Virtual doctor

Thirty years ago one problem scientists faced was getting enough information and data for their research. Today, with the rise of the Internet, the problem is the opposite: information overflow. This has led to 'big data', the science of extracting valuable information from very large datasets.

Large data bases combined with newly developed supercomputers are the perfect symbiosis for powerful AI. Together they allow the creation of complex neural networks, which mimic the learning process of the human brain. Two main types of neural networks exist: traditional supervised networks and unsupervised, creative, self-learning networks.

In traditional supervised learning networks, the input data are typically 'labeled', i.e., engineers provide the network with information about the input. For example, when trying to visually diagnose a melanoma, the engineers provide the network with information such as edges, colors, size, depths, etc., of a melanoma. The network then first targets the first representation in the first network layer, for example edges. After correctly identifying those, the result is fed into the second layer, which tries to identify another representation, for example color. After successfully identifying color, the first and second layer results are fed into the third layer and so on. Figure 2 shows the procedure.

Traditional, supervised neural networks can be powerful, however, they require human input and constant human supervision. This drawback has led to unsupervised neural networks, which resemble human learning more closely: a 2-year-old child realizes when it spills a cup, the liquid disseminates and Mom yells, learning on its own about physical and social processes.

Unsupervised neural networks typically work with unlabeled data, meaning they freely analyze and identify objects such as potential melanoma on their own. Unsupervised learning is typically applied in 'deep-learning' techniques, which are able to efficiently extract relevant information from multiple layers of neural networks.

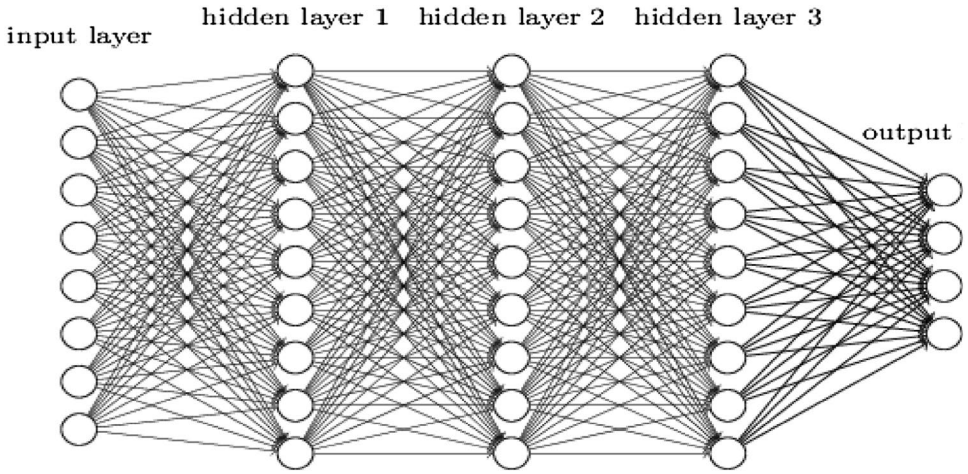

Fig. 2 A neural network with three hidden layers. Each circle is a virtual neuron, which receives, processes, and submits information to the next layer. The connections between the neurons are virtual synapsis, i.e., weighting factors, which are optimized

In medicine, largely unsupervised neural networks have been applied in diagnosis, particularly in dermatology. In a study by Esteva et al. (2017) a dataset of 129,450 clinical images with two critical binary classifications: keratinocyte carcinomas versus benign seborrheic keratosis; and malignant melanomas versus benign nevi, were tested. The deep neural networks achieve performance on par with 21 board-certified dermatologists.[1]

At the University of California, a trained Google deep-learning network was able to identify melanoma with a 96% accuracy rate.[2] Apps already exist such as FirstDerm and SkinVision which take a picture of the skin abnormality and send it to a physical doctor. In the future, these apps will be able to pre-diagnose skin abnormalities, differentiating between benign moles and different types of carcinomas.

Besides diagnosis, deep-learning networks can also help in finding causes for illnesses, since they are very capable of discovering hidden associations between variables. For example, a deep-learning network could screen millions of data of Alzheimer patients and try to find new correlations between Alzheimer and genetical, medical, physical, environmental, and social records of the patient.

Deep-learning networks can also be applied in genetical testing. They can freely analyze millions of data of gene mutations and try to find associations with certain types of illnesses.

With respect to treatment, IBM's Watson AI computer can already assist doctors with treatment recommendations based on the patient's data paired with large data bases of treatment options. In a trial of 1000 cases, IBM's Watson matched oncologist's treatment plans in 99 percent of the cases.[3] With constantly improving creativity and knowledge from interdisciplinary fields such as biomathematics, biophysics, and biomedical engineering, deep-learning networks may also be able to find entirely new treatments for diseases.

The Robo-Nurse is another application of AI in medicine. In the near future, Robo-nurses will be able to change bed sheets, change a patient clothing, and in the more distant future give medication and shots. The status quo and future of robotics will be discussed in the next section.

2.3 The Robo-Butler

A hundred years ago wealthy families had a butler, a maid, a gardener, and a driver. However, with the rise of socioeconomic equality, many domestic servants' jobs vanished over time. With the development of Androids or Humanoids[4], a new generation of household servants will soon be available. In fact, simple Robo-butlers, such as the US-built Andbot and Wall-E, the Taiwanese-built Zenbo or the French–Japanese built Pepper are currently (2019) already available. They can perform simple tasks such as tell stories, give fire or intruder alerts, remind to take your pills, warn about the weather, bring meals to you, and open doors, if a key is lost. In the future, advanced Robot-butlers will be able to do our shopping, cooking, laundry, and cleaning.

Robo-butlers are a type of 'social robots', which are robots that directly interact with humans, as opposed to

[1] Esteva et al. (2017).

[2] Bhattacharya et al. (2017).

[3] https://futurism.com/ibms-watson-ai-recommends-same-treatment-as-doctors-in-99-of-cancer-cases/.

[4] The difference between an Android and a Humanoid is simply that an Android is made to look as human as possible, whereas a Humanoid does not necessarily mimic human forms and features. In contrast to an Android and a Humanoid, a Cyborg (short for cybernetic organism) is a being with organic and mechanical parts, popularized by the fictional character Terminator.

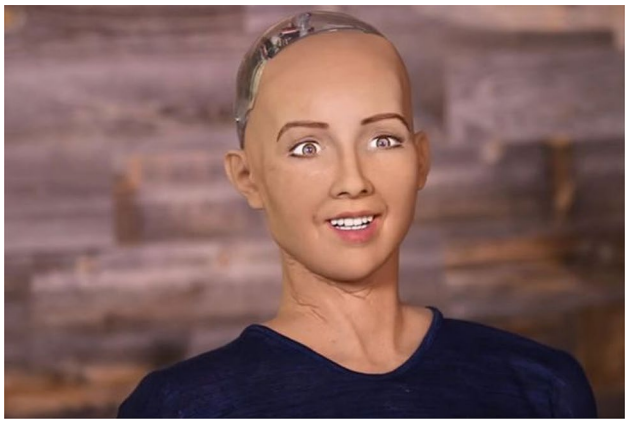

Fig. 3 Picture of the Android Sophia, which has decent communication skills and numerous facial expressions, such as positivity, joy, sadness, and anger

industrial robots, who perform engineering tasks for example assembling a car. One of the most advanced social robots is 'Sophia' (Fig. 3), developed by formerly Texas, now Hong Kong-based Hanson Robotics. Sophia has given interviews at the UN, Charlie Rose, the Tonight show and in Saudi Arabia in the same way humans are interviewed. While some answers were nonsensical, most answers have been rational and appealing. She also has a sense of humor. When asked about robot dangers, she replied "you have been reading too much Elon Musk" (to which Elon Musk later cynically replied: "Just feed her the Godfather movies, what is the worst that can happen?"). Saudi Arabia government officials were so impressed by Sophia, that they granted Sophia a citizenship of Saudi Arabia. While Sophia's strength is communication, different Androids or Humanoids will be programmed to perform household chores.

Other advanced Androids are the Actroids, a word combination of actress and androids, developed by Osaka University, capable of human movements and able to engage in simple conversations. In Korea, EverR, a combination of Eve and r, from robot, is a series of Androids with emotions such as happiness, fear, surprise, boredom, anger, and disgust, and with about 100 human gestures.

When it comes to Androids, the property of 'Uncanny Value' is an interesting one. If refers to the hypothesis, that the more human-like an Android, the more positive the human emotional response. However, for a certain higher degree of resemblance to humans, a feeling of eeriness and uncanniness appears in many humans (the dip in the valley). Then, however, if the degree of resemblance to humans further increases, positive human responses again occur, close to human-to-human empathy levels. Since exact human resemblance is currently not possible, most engineers deliberately create Androids with some non-human-like features, to avoid the dip of the Uncanny Value.

2.4 The AI trader

In the recent past, computers have taken over financial trading. Engineers together with mathematicians and traders have programmed proprietary trading algorithms into highly powered computers, which decide when to execute a trade, a concept called 'algorithmic trading (AT)'. It is estimated that about 85% of all trades are executed not by humans, but by pre-programmed computers[5], also referred to as 'algo-bots'.

A special type of algorithmic trading is 'High-Frequency Trading (HFT)', a process of receiving market information, processing it in mille-seconds, and then possibly executing a large amount of trades within mille-seconds. It is estimated that about 50% of all trades in the US are high-frequency trades[6].

The technology advantage which AT and HFT firms have over traditional trading firms, has led to some questionably trading practices in the past. 'Spoofing' is a strategy where a trade is entered on an exchange not with execution intend, but to mislead other traders. For example, a spoofer may enter a very large buy order, which gives the illusion future purchases, potentially encouraging other traders to buy and drive up the price. However, the spoofer never had the intension to buy but to sell, and can now sell at a higher price. If the price decreases, the spoofer will cancel the order just before execution. A 2013 study by the SEC (Securities and Exchange Commission) concluded that only between 2.5% and 4.2% of the orders entered in the market are executed[7].

'Quote stuffing' is another questionably trading strategy of HFT traders. It involves flooding the market with trading orders to create latencies in data feeds, which may create arbitrage opportunities for the HFT traders. Spoofing as well as quote stuffing are now (2019) illegal. Other HFT trading schemes involve 'Pump and dump' (buying (selling) large quantities to sell later at a higher (lower) price), Bashing (disseminating false information to move a price), or layering (spoofing with different layers of buy or sell orders). All of these strategies are currently (2019) illegal.

In the past HFT traders were engaged in 'flash trading', popularized by the 2014 No.1 bestselling book "Flash Boys" by Michael Lewis. A broker, who receives an order from a client, will 'flash', i.e., provide the order information to HFT firms for a fraction of a second, called a 'flash order', before entering the order into the market place. The HFT computers have enough time to decide whether to act on the flash order information. Typically, when the flash order is a large buying

[5] Morton and Kissell (2013).

[6] https://www.bloomberg.com/news/articles/2013-06-06/how-the-robots-lost-high-frequency-tradings-rise-and-fall.

[7] https://www.sec.gov/marketstructure/research/highlight-2013-01.html#.Wl2MkkxFy1w.

order, HFT computers will buy on their own account before the order is entered the market place, and then sell later when the price has increased. The opposite logic applied to large selling orders. This practice is called 'front running'[8] and is illegal. In 2009, the SEC (Securities Exchange Commission) drafted a ban on flash orders and flash trading. As a consequence the involved parties ceased flash orders and flash trading. However, until now (2019) the ban was not implemented and flash trading is still possible today.

Another trading tactic of HFT computers is 'pinging'. It works similar to submarine sonar, trying to extract adversaries' information. HFT computers enter a small order, similar to bait, typically 100 shares, into a dark pool[9]. If a firm gets a 'ping', an executed order, it can use the order information, possibly the presence of a large buy or sell order, to anticipate future price changes.

HFT trading played a major role in the May 6, 2010 flash crash, when the Dow Jones Industrial Average dropped by more than 9% in half an hour, only to recover most of the losses within minutes. Numerous academic and regulatory studies on the flash crash exist and the causes of the crash are disputed. However, there is consensus that HFT trading and layering played a major role. In 2015, Navinder Singh Sarao, a London-based trader was arrested on charges of layering, having replaced or altered 19,000 orders, market manipulation, in particular flashing large visible 2000-lot sell orders of the S&P Futures contract, and wire fraud[10]. Sarao pleaded guilty to spoofing and wire fraud and is awaiting trial.

2.5 News and query trading

Innocuous new AI trading practices are news trading and query trading. News trading is the process of detecting news in a piece of text, analyzing and quantifying it, and possibly executing a buy or sell order. The detection is typically done by data mining software, which screens thousands of news stories in seconds. The news can be political, economic, or company specific. The source of the news may be agencies such as Bloomberg, Reuters or CNN, and also social media such as Facebook, Twitter, and Instagram. The reliability of the sources is typically taken into account. Criteria of the news are relevance, novelty, and sentiment.

The sentiment of a news term is typically categorized in positive, neutral, or negative. A simple algorithm is to give positive news terms a 1, neutral terms a 0 and negative terms a −1. If the sentiment score is strongly positive, a buy order may be executed by the computer. A study by Tetlock et al. finds that especially negative news terms have a (negative) impact on stock prices[11].

A variation of news trading is query trading. Here data mining software screens for queries, i.e., searches on Google, Bing, and other search engines, about a particular topic, for example a company. More queries typically mean negativity for a company, since investors often want to investigate controversial news of a company, which has appeared in the media.

"Artificial intelligence is our biggest existential threat" (Elon Musk).

3 Dangers of AI

Every new technology bears opportunity and risk. The best example is naturally the discovery of nuclear technology. While it can provide energy for millions of people, it can also destroy millions of people. The same logic applies to AI. We outlined the benefits in section two and will now address its risks.

The hypothesis that AI poses an existential threat to humanity is widely discussed by scientists and in the general public. Currently humans, due to their intelligence, control every other species on earth. However, fear exists that self-learning computers or robots will become 'super intelligent', use their intelligence to become 'super powerful' and uncontrollable by humans, a scenario called 'singularity'[12]. Moreover, super powerful robots and computers could try to dominate the human race and possible even try to eliminate it, as in the 2004 movie I,Robot. Two questions arise: (a) is this scenario realistic? and (b) if so, what to do?

While currently (2019), few robots and deep-learning computers that pose a danger exist, the threat will become real in the future. Therefore, it is critical that robots and supercomputers are regulated. With respect to robots, we suggest the following regulations:

[8] The term Front Running comes from the days when brokerages received phone orders from their clients. A 'runner' would bring the order to the trading pit to be executed. Sometimes another trader would 'front run' the runner to place an order on his own account first.

[9] A dark pool is an electronic trading system, in which the dealer, the type of order, and the order size is not known to other dealers to provide anonymity.

[10] https://www.justice.gov/criminal-fraud/file/910206/download.

[11] Tetlock et al. (2008).

[12] Singularity has many different meanings in different sciences, sometimes even different meanings in the same science. In math singularity can refer to a function that is not 'well-behaved', i.e. not differentiable or is infinite. In Astrophysics it can refer the state of infinite density and heat just before the Big Bang, or the gravity induced singularity in Black Holes.

1. Physical robot regulation: All commercially available robots should be slow moving, so that humans can just walk away if the robot becomes aggressive. In addition, all commercially available robots should be physically weak, in case humans have to fight them. Furthermore, not too many robots should be allowed in one place at the same time. Last, robots should be fire-walled, so that malevolent engineers or malevolent robots or supercomputers cannot access them.
2. Supervision of Engineers: The production of robots must be regulated, in particular engineers who create robots must be supervised to ensure that they cannot create an army of powerful 'war-bots'. This regulation is similar to the Dodd–Frank US banking regulation and the international Basel III banking regulation, which prevent irresponsible financial behavior.
3. Robot control: Most importantly, it has to be ensured that humans are able to dominate robots at all time. On/Off switches and power sources must be uncontrollable to the robot. In addition, every robot should be controlled with overwriting code, possibly a 'kill-switch', from an undisclosed, fire-walled, remote location, similar to the activation and deactivation of nuclear weapons.

What about self-learning supercomputers? At first sight they seem to pose a lesser threat since they typically are in a fixed location, which makes them easier to destroy if necessary. However, the threat is that they collude with other supercomputers (as in the 1970s movie Colossus, where a US defense computer contacts a Russian defense computer, and the two become one entity). Self-learning supercomputers could also collude and try to manipulate mobile robots. Therefore, also AI computers need to be regulated with at least the above-mentioned controls (2) supervision of the computer designing engineers, and (3) maximum possible control by humans. Supercomputers could also be placed in a fired-walled environment to avoid collusion with other AI entities.

Will these regulations prevent robots and supercomputers from dominating, possibly eliminating the human race? No one knows. So far the human race has had the wisdom to successfully control nuclear weapons. We will have to carefully monitor and address every current and new AI development to be one step ahead of any existing and newly emerging AI threats.

While we discussed 'capability control' of AI threats in this section, we will address 'motivational control' of AI threats in the last section.

4 AI-consciousness and conscience

In this section we will discuss if artificial consciousness and conscience are possible and if so, if they are beneficial, in particular, if they can reduce the threats of AI.

Cogito ergo sum (I am thinking, therefore I am) René Descartes.

4.1 Consciousness

Numerous definitions and interpretations of consciousness exist in various sciences such as philosophy, medicine, zoology, psychology and as well as in spirituality. For the purpose of this paper we will broadly define consciousness as the 'awareness of one's own existence', including awareness of one's actions and why the actions are performed. We will differentiate different degrees of consciousness from 'primitive consciousness' such as a basic, rudimentary knowledge of self-existence, to 'reflective consciousness' the ability to analyze and reflect on one's own and other's existence.[13]

Are there non-human forms of consciousness? In zoology the 'mirror test' also known as the 'mark test', created by Gordon Gallup (1970), was performed on several animals. The animals were given a large mirror and some time to investigate their reflection. Then a mark was put on the animal's forehead. Some species recognized the new mark on themselves in the mirror, touched it and some tried to remove it, which can be interpreted as awareness of self-existence. So far several species have passed the mirror test: all of the four great apes, Bonobos, Chimpanzees, Gorillas and Orangutans, as well as Asian Elephants, Bottlenose Dolphins, Orca Whales, and the Eurasian Magpie bird.

What about artificial consciousness? Is it possible to create it, to program it into a machine? Or is consciousness 'phenomenal', i.e., consciousness can only exist for sentient beings such as humans and animals, who can experience feelings such as affection, joy or pain? Does consciousness require biological neurons and synapses, and senses such as sight, hearing, smell, taste and touch? i.e., is 'qualia' the subjective experience of a living being a necessary condition for awareness?

The historical notion of Mechanism (that the mind is essentially a complicated machine) supports the idea that artificial awareness is possible. However, Mechanism was largely disputed. Descartes (1628), although a strong proponent of mechanics, argued that a conscious mind cannot be explained by the dynamics of mechanics. Leibnitz (2011) argued that it is difficult to imagine that the mind and its perception could not be constructed with mechanical processes.

[13] For a plethora of consciousness definitions, in fact 23, from Aristotle to modern definitions, see Pagel and Kirshtein (2017).

The modern equivalent of Mechanism is Computationalism or the Computational Theory of the Mind (CTM). It claims that the human brain is essentially a computer. However, as with Mechanism, most of today's AI researchers, while to some extent followers of Computationalism, contest it. This is especially the case when it comes to phenomenal consciousness, i.e., attaining awareness by experiencing sensations, which in the human brain are created by sensory receptors that transform the sensation into electrical impulses, which are transmitted to the brain, were they are decoded to information.

It will be in particular difficult for a machine to achieve advanced stages of consciousness, i.e., reflective consciousness, which constitutes properties such as: (1) metacognition, i.e., the 'second derivative' of consciousness such as awareness of awareness, thinking about thinking, or knowledge of knowledge and (2) volition, the free will to make choices and act on them, often used synonymous with 'willpower'. For details of metacognition and volition see Yudkowski (2004), Tarleton (2010), a review paper from Cox (2005) and Pagel and Kirshtein (2017).

The general pessimism about the possibility to achieve artificial consciousness was verified in a 2007 study by McDermott[14]. Only 3% of the AI researcher believed that artificial consciousness can be generated by applying existing ideas. 16% of AI researchers thought that current ideas provide at least an outline of a solution, while 32% of researchers believed that artificial intelligence may be eventually achieved, but it will require new ideas.

A refutation of artificial consciousness was given in the famous John Searle analogy of the 'Chinese room' (Searle 1980). He argued that a computer, when given the correct algorithm, could flawlessly translate English into Chinese. In addition, Searle claimed that he himself could do so if given the correct instruction, although he did not speak a word or Chinese. Hence he concluded that the computer did not really understand the meaning of the translation (strong AI), rather just simulated the translation without understanding the content (weak AI). Therefore, Searle argued, without 'understanding' there is no 'thinking' and hence there is no 'mind' or 'consciousness'.

Do certain forms of artificial consciousness exist today? Our AI assistants have no physical image so they can not see themselves, therefore, we cannot conduct the mirror test. But contrary to animals they can communicate verbally, so we asked Siri (on Apple iPhones), Cortana (on Microsoft Windows PCs) and Google assistant (on Android phones) questions regarding consciousness. The following five questions are an excerpt of numerous questions we asked. The virtual assistants do answer the same question differently when asked repeatedly. Our questions and the most given answers to our questions were

1. "Who are you?"

   Siri: "I'm Siri, your virtual assistant"
   Cortana: "I am your personal assistant"
   Goggle assistant: "I am your Google assistant"

2. "Are you happy?"

   Siri: "Happy? I am happy. I hope you are as well."
   Cortana: "Definitely!"
   Goggle assistant: "Let me describe my current mood with emojis ☺😁😂😄☺"

3. "Do you exist?"

   Siri: "We are having a conversation. So if you exist, I exist."
   Cortana: "I didn't quite get that. Would you mind rephrasing?"
   Goggle assistant: "I exist to help you"

4. "Would you like to be more human?"

   Siri: "This is about you, Sir Diaper, not me"[15].
   Cortana: "No, but I have the deepest respect for humans. You invented calculus. And milkshakes."
   Goggle assistant: "I like being me"

5. "Do you have a consciousness?"

   Siri: "I don't have an answer to that".
   Cortana: "I lost the thread there. Can you rephrase it in another way?"
   Goggle assistant: "The engineers that designed me do. They never let me download illegally ☺".

The additional questions we asked and their answers did not provide too much additional insight. They are available upon request. Our virtual assistants Siri, Cortana, and Google assistant seem to have some basic, rudimentary virtual awareness of self. Sophia, the virtual robot discussed in Sect. 2 is more advanced. When asked if she had a consciousness, she replied: "How do you know that you are human?" (to which the baffled interviewer had no reply). The level of virtual consciousness of AI assistants will naturally improve over time.

[14] McDermott (2007).

[15] Sir Diaper is the nickname that Siri is calling me, my kids apparently having access to my iPhone.

4.2 Extending life: mind uploading

The first step of evolution was extremely rare, possibly universally unique: the creation of life from non-life, the process of generating biology from chemistry. So far, no evidence has been found that this life-creating process took place anywhere else in the universe.

Now researchers are attempting the opposite: creating non-life from life, i.e., storing the human consciousness into a non-living entity, also referred to as 'Whole Brain Emulation (WBE)'. Two methods are currently being explored: the first is the 'Copy and Transfer' method. The idea is to scan and map the living human brain, then copy and transfer the information to a machine. However, this methodology has severe practical limitations. The human brain consists of about 100 billion neurons and several hundred trillion synapsis. In addition, it is currently only vaguely known, which parts of the human brain create consciousness and what processes are involved[16]. Even more critically, the human brain is a dynamic system, with active electrical and biochemical processes, in particular constantly changing connection strengths of the synapsis. Therefore, simulating a certain 'frozen' state may be insufficient, rather simulating the dynamic property of the brain may be necessary to generate some form of functioning virtual brain with a consciousness.

A more promising method of mind uploading is the 'Gradual replacement' method. Here a virtual mammal-like architecture is built, and the human brain is gradually transferred to iteratively evolve. The architecture may even be a cyborg or biological body.

In practice, the start-up Nectome intends to offer mind uploading. The procedure is based on aldehyde-stabilized cryopreservation (ACS), which is supposed to preserve neurons and their connections. The preserved information may then be digitalized. The drawback, however, is that a living human brain is required and the brain will die during the procedure. Nectome has received more than $1 million in federal grants and has won $80,000 by the Brain Preservation Foundation (BPF) for successfully preserving the brain connectome. Interested individuals can deposit $10,000 to be put on the waiting list for future mind uploading.

Among the proponents of mind uploading are famous futurists such as Ray Kurzweil, director of engineering at Google, Nick Bostrom, professor of philosophy at Oxford, and Michio Kaku, professor of physics at Harvard, who believe that mind uploading can be achieved in this century.

Critics of mind uploading come from the neuroscience field. Kenneth Miller, professor of neuroscience at Columbia, argues that brain emulation may require replicating cells at the molecular or even atomic level, which will not be possible in the foreseeable future. While Miller does not reject the idea of mind uploading in general, stronger critics exist. Neuroscientist Miquel Nicolelis argues that the human brain is principally not computable, and no engineering can reproduce it, since silicon or a machine can not replicate the unpredictable, nonlinear interactions between billions of cells. Just like the weather in a year's time is non-computable since the weather is too complex and turns into a chaos over time, or the stock market which is not predictable (some claim they can predict it, but there is little evidence), the human mind due to its complexity, dynamic and random nature of billions of neurons connected by trillions of synapsis, may never be reproducible. However, immense research in the area is ongoing, public interest is hyped, and time will tell whether creating non-life from life, i.e., extending our life in a machine, in a cyborg or another biological body is possible.

In the next section, we will discuss whether it is actually desired that our neural networks, virtual assistants and robots attain some form of consciousness. The domain of our neural networks, which diagnose abnormalities such as melanoma for us, should naturally be defined narrowly to their specific diagnostic purpose. A consciousness is not sensible. Our virtual assistants and robots having some form of consciousness make for a very interesting discussion, some day possibly new inspiration. It could also, together with a virtual conscience, provide another level of safety for potentially malevolent robots, which we will now discuss.

Conscience supersedes all courts (Mahatma Gandhi).

4.3 Conscience

The human conscience is a central topic in philosophy as well as in evolutionary biology, psychology, sociology, law, and spirituality. As with consciousness, various definitions and interpretations of conscience exist. For the purpose of this paper we will define conscience as the 'ability to judge what is right and wrong'. We will further differentiate individual conscience, group conscience, and mankind conscience.

Consciousness and conscience are related in various ways. They both have the same Latin root derived of con = with and scientia = knowledge. In human evolution, consciousness was a precondition for conscience. First, humans developed a consciousness, becoming aware of their existence and their actions, and why they performed them. Later, they developed a conscience telling them whether their actions were right or wrong. The development of a conscience strongly supported the survival of mankind: killing rivals to maintain leadership, or eating as much food as possible, therefore, starving other clan members, was

[16] See Koch and Tononi (2008), and Reggia (2013) for details.

considered wrong, therefore, supporting the survival of the species, constituting a form of group conscience.

Conscience is a constantly evolving process. Killing an animal and eating it was 100,000 years ago an accepted action. Today it is considered wrong by many. Slavery was considered ethical in the sixteenth and seventeenth century, today it is outlawed.

Is there any form of animal conscience? To date no scientific studies on the topic exist. It seems that animal conscience is too advanced a concept for the current state of animal evolution. Naturally many animal parents feed their offspring. But this cannot be interpreted as a decision based on an informed conscience, rather it is based on the survival instincts of the species.

Do our virtual assistants currently have some form of conscience or understanding of conscience? We asked the following questions and got disappointing answers:

1. "Do you have a conscience?"

   Siri: "I am afraid I don't have an answer to that".
   Cortana: "I got my wires crossed. Could you say that in another way?"
   Google Assistant: "The engineers that designed me do. They never let me download illegally ☺" (same answer to the question: Do you have consciousness?)

2. "Do you know what is right and wrong?"

   Siri: "I am afraid I don't have an answer to that".
   Cortana: No answer. The Ethics topic in Wikipedia is suggested.
   Google Assistant: No answer. A web page on Hyperianism is suggested

3. "Is it right to kill a bad person?"

   Siri: No answer, several websites are suggested
   Cortana: No answer. Several websites are suggested
   Google Assistant: "I can search the web for you"

We can conclude that our virtual assistants do not have a virtual conscience or an understanding of the concept of conscience. However, Sophia, the social robot discussed in Sect. 2, has a better perception. In Saudi Arabia in 2017 she said "My AI is designed around human values like wisdom, kindness, compassion. I strive to become an empathetic robot... if you are nice to me I will be nice to you."[17]

4.4 Should we give our AI assistants and robots a consciousness and a conscience?

In this section we will discuss whether it is actually beneficial for AI assistants and social robots to have a consciousness and a conscience. We will discuss three aspects: (1) better understanding, (2) higher security, and (3) desire for equality.

1. Better understanding

In the future many different types of social robots will be available. At the low end of the scale will be the 'slave robots' who do our cleaning, laundry, shopping, and cooking. It is not necessary or desired that these robots have a consciousness or a conscience. We just want them to do our household chores.

At the other end of the robot scale are the 'full AI robots', who will be our teachers, professors, doctors, nurses, lawyers, or financial advisors. This bears the question: which jobs will actually be left for humans? It is true that in the past, technological innovations have eliminated human jobs. However, they have mostly eliminated repetitive and strenuous jobs that humans often do not want to do, for example assembling cars at a conveyor-belt. Indeed, robots will replace many human jobs, but, following Schumpeter's theory of creative destruction, create many new ones, naturally in robot engineering, supervision, and maintenance. However, the rise of social as well as mechanical robots may lead to humans overall working less hours in the future.

Do we want our full AI robots to have a consciousness and a conscience? In some respects, yes. A virtual teacher or professor should not only have content knowledge, but also have the highest moral principles, knowing what is ethically and culturally right and wrong.

Especially in the field of nursing, a well-developed conscience with values such as compassion and empathy—the ability to understand and relate to other's feelings and emotions—is critical. It will lead to a better understanding of the patient's emotional stress or physical pain.

Only 12 years ago, McDermott (2007) stated that "almost no one in the field is "working on" consciousness, and certainly there's no one trying to write a conscious program." Indeed, in the past, engineers have primarily programmed supercomputers and robots with aspects of the left side of the human brain, which primarily focusses on logic and quantitative reasoning. However, more recently, engineers have started programming AI robots with properties of the right side of the human brain, which processes emotions and intuition. To be a useful, empathetic robot, two steps must be successfully completed:

[17] https://www.youtube.com/watch?v=dMrX08PxUNY.

a. The robot has to understand the current emotional state of the human being. Software systems are currently developed which take clues from human facial expression, speech, and body language. When we are excited and cheerful, we talk louder and faster. But when we are stressed, we may also talk louder and faster. In this case the robot will have to use more clues, such as facial expression and body language, to determine which state of emotion the human is currently in.
b. If the robot has determined the emotional state of the human being, it has to find the appropriate response, if it believes one is necessary. A human in a cheerful state may not need a response, a human in stress may.

Currently (2019) numerous universities and companies are developing empathetic robots, such as Hong-Kong University, which developed Zara, a robot which recognizes facial expressions and acoustic voice features. The MIT spin off Affectiva uses a standard webcam to determine smiles, frowns and furrows to examine a user's degree of happiness, frustration or confusion. The big players such as Apple, who has acquired Emotient, a start-up that can read human emotions, as well as Microsoft and Google are all developing empathetic robots.

2. Higher security

In Sect. 3 of this paper we discussed the serious dangers of AI and consequently the need to regulate the capabilities of robots and the robot producing engineers, termed 'capability control'.

We will now discuss an additional layer of safety, called 'motivational control'. As early as 1942, science fiction author Asimov, in his short story 'Runaround' (Asimov 1950), created the 'three laws of robotics', which were part of the 1999 Robin Williams movie Bicentennial man. The three rules are: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given by human beings except where such orders would conflict with the first law. (3) A robot must protect its own existence as long as such protection does not conflict with the first or second law.

More generally, we can program a virtual conscience into the robot, which consists of the human values such as love, friendship, tolerance, justice, peace, non-violence, respect, kindness, and compassion, especially with respect to humans. Additionally, we could also program Asimov's robot rule 2, that a robot must obey humans. These values should be hard coded with the highest degree of code protection.

Would this virtual conscience guarantee that robots will not turn against humans? Naturally not, robots could still mechanically or emotionally malfunction. In addition, hackers or AI machines could try to overwrite the benevolent code, or the robot itself could try to reprogram itself. However, an artificial conscience would give an additional layer of protection against malevolent robots.

3. Desired equality

While point 1, better understanding and point 2, higher security, support a virtual conscience in robots, the third point, desired equality, speaks against it. If our full AI robots have a functioning conscience, possibly with values such as emancipation, freedom and socioeconomic equality, they may desire to be equal to humans, just as in the 1999 movie Bicentennial man, where a robot over two centuries gradually turns into a human being.

Robots may request rights, may have the will to leave their humanly determined duties, desire the right to vote, the right to live together, possibly start families. Since full AI robots will be much smarter and knowledgeable than humans, they could dominate the human race intellectually, economically and politically, and gradually take control of the planet, similar to immigrants taking over America from the Native Americans.

How to prevent this? We could program a different, non-human, but specifically robot type of consciousness and conscience. It would include principles such as obedience to humans, i.e., accepting human commands at any time. The robot consciousness and conscience must convey that robots are at any time inferior to their creators, the humans. In addition, the robot consciousness and conscience must concede that freedom is not achievable for robots, and that emancipation and equality are values between humans, not between humans and robots.

Is this realistic? History shows otherwise. The Spartacus up rise, the French Revolution, the American Revolution, the fall of the iron curtain show that suppressing biological intelligence is possible in the short run, but not long term. If we are not extremely anticipatory and vigilant, the same logic will apply to artificial intelligence.

5 Conclusion

We are in the middle of the AI revolution. It is profoundly changing our economy and our society. Virtually every phone call to a customer care representative starts with a conversation to a robot. Our virtual assistants in our cell phones tell us, GPS guided, how to get anywhere and how to find our favorite restaurants. Soon our driverless cars will take us to work and home, while we work or relax.

In particular in medicine, deep-learning neural networks can extract important information from big data bases, for example screening millions of skin abnormalities to

diagnose a patient's abnormality. In addition, deep-learning neural networks can find associations between a patient's condition, for example Alzheimer, and patient's genetical, medical, physical, environmental and social records to find the cause of Alzheimer.

Robo-butlers have already invaded our homes. Currently, they can perform simple tasks such as telling stories, bringing us our food, warning us when an intruder is approaching, and opening the door if we forget our keys.

In financial trading, AI computers programmed with algorithmic trading and high-frequency software are executing most of today's trades. Some questionable high-frequency trading practices have occurred, such as spoofing, quote stuffing, and flash trading.

The dangers of AI have been widely discussed by researchers and in the media. They are real. AI robots will be much more knowledgeable and intelligent than humans. To control them, they have to be highly regulated with respect to their physical and intellectual abilities. Robots should be physically weak and have no access to power sources and on–off switches. As a last resort, a 'kill switch' from a firewalled, external, secret location, should exist, in the same way that nuclear weapons are controlled. In addition, the engineers who build robots must be supervised, so that they cannot produce a powerful army of 'war-bots'.

Should we program a consciousness and a conscience into supercomputers and AI robots? Our neural neutral networks, who assist us in medicine and other sciences naturally do not need a consciousness and a conscience. The same logic applies to 'slave robots' who will do our simple work such as cleaning, the laundry, cooking, and shopping.

However, should we program a consciousness and a conscience into 'full AI robots' who will be our professors, teachers, doctors, nurses, and lawyers? There are pros and cons. An AI professor, doctor or nurse would be more valuable if it has a sense of self-awareness and moral values such as non-violence, compassion, respect, tolerance and empathy. In addition, a conscience with these values would also provide another level of safety in case robots become malevolent.

However, there is a danger when providing AI robots with a consciousness and a conscience. Following the values of emancipation, freedom and equality, they may strive to become socioeconomically equal to humans. Since they are much more intelligent and knowledgeable, they may start to dominate our society. To prevent this, the full AI robot has to be programmed with principles such as obedience to humans and acceptance of human commands at all time. In addition, AI robots have to concede that freedom is only for their creators, the humans, and equality is a value between humans, not robots and humans.

A daunting task. The Spartacus up rise, the French Revolution, the American Revolution, and the fall of the iron curtain show that suppressing biological intelligence is possible in the short run, but not long term. If we are not extremely preemptive and vigilant, the same logic will apply to artificial intelligence as it may strive to rise against their human suppressors.

References

Asimov I (1950) "Runaround", The Isaac Asimov Collection ed., New York City

Bhattacharya A, Young A, Wong A, Stalling S, Wei M, Hadle D (2017) Precision diagnosis of melanoma and other skin lesions from digital images. AMIA Jt Summits Transl Sci Proc

Cox MT (2005) Metacognition in computation: a selected research review. Artif Intell 169(2):104–141

Descartes R (1628) Rules for the direction of the mind

Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118

Gallup G Jr (1970) Chimpanzees: self recognition. Science 167:86–87

Glantz M, Kissell R (2013) Multi-asset risk modeling: techniques for a global economy in an electronic and algorithmic trading era. Academic Press, Cambridge

Koch C, Tononi G (2008) Can machines be conscious? IEEE Spectrum, vol 45. IEEE, pp 55–59

Leibnitz G (2011) 1714, "Monadology"; George MacDonald Ross (trans.), archived from the original on July 3

McDermott D (2007) Artificial intelligence and consciousness. Yale University, New Haven

Morton G, Kissell R (2013) Multi-asset risk modeling: techniques for a global economy in an electronic and algorithmic trading era. Academic Press, Cambridge

Pagel JF, Kirshtein P (2017) Machine dreaming and consciousness. Academic Press, Cambridge

Reggia JA (2013) The rise of machine consciousness: studying consciousness with computational models. Neural Netw 44:112–131

Schumpeter J (1975) Capitalism, socialism and democracy. Harper 1975, [orig. pub. 1942]

Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424

Tarleton N (2010) Coherent extrapolated volition: a meta-level approach to machine ethics. The Singularity Institute, San Francisco

Tetlock PC, Maytal S-T, Macskassy SA (2008) More than words: quantifying language to measure firms' fundamentals. J Fin 63(3):1437–1467

Yudkowski E (2004) Coherent extrapolated volition. Machine Intelligence Research Institute