big data and cognitive computing

MDPI

Article

From Big Data to Deep Learning: A Leap Towards Strong AI or 'Intelligentia Obscura'?

Stefan Strauß

Institute of Technology Assessment (ITA), Austrian Academy of Sciences, Vienna 1030, Austria; sstrauss@oeaw.ac.at

Received: 1 June 2018; Accepted: 16 July 2018; Published: 17 July 2018

check for updates

Abstract: Astonishing progress is being made in the field of artificial intelligence (AI) and particularly in machine learning (ML). Novel approaches of deep learning are promising to even boost the idea of AI equipped with capabilities of self-improvement. But what are the wider societal implications of this development and to what extent are classical AI concepts still relevant? This paper discusses these issues including an overview on basic concepts and notions of AI in relation to big data. Particular focus lies on the roles, societal consequences and risks of machine and deep learning. The paper argues that the growing relevance of AI in society bears serious risks of deep automation bias reinforced by insufficient machine learning quality, lacking algorithmic accountability and mutual risks of misinterpretation up to incrementally aggravating conflicts in decision-making between humans and machines. To reduce these risks and avoid the emergence of an intelligentia obscura requires overcoming ideological myths of AI and revitalising a culture of responsible, ethical technology development and usage. This includes the need for a broader discussion about the risks of increasing automation and useful governance approaches to stimulate AI development with respect to individual and societal well-being.

Keywords: artificial intelligence; deep learning; autonomy; automation; bias; algorithmic accountability; Turing test; ELIZA effect; technology assessment

1. Introduction

Creating intelligent machines has always been a vision of mankind. Rapid technological progress in the field of artificial intelligence (AI) being made over the last few years makes this vision more tangible. Today, the spectrum of applications directly or indirectly equipped with machine "intelligence" is broader than ever: sophisticated algorithms do not just defeat humans in complex games like chess, Jeopardy or Go, but support various kinds of human tasks ranging from Web search, text, image or voice recognition or even predicting trends. There is a general increase in automated systems and cognitive computing benefiting the cluster of "smart" technologies, robotics, "ambient intelligence", the "Internet of Things", remote piloted systems (drones), self-driving cars, assistive technologies etc. In each of these different developments, approaches of machine learning (ML) play an essential role which usually involves the processing of large amounts of raw data. The trend of big data and "datafication" with the basic aim to gather machine-readable data from contexts in everyday life [1–3], thus stimulates developments in the field of AI and vice versa. Recent progress being made in machine learning with so-called deep learning promise more flexibility and efficiency to structure and process raw data.

But how great is this potential factually, what are the prospects and limits, societal consequences and risks of deep learning and similar machine learning approaches? This paper critically examines the wider societal implications of AI along these questions and discusses ethical problems related to developments toward (self-)learning machines. The analysis combines two main perspectives:

the first perspective is on the emergence of AI, its main characteristics and, in particular, the basic functioning and limits of machine learning. The second perspective is on issues of human–machine interaction related to the Turing test which implies automated imitations of human behaviour. Both perspectives are considered when discussing societal consequences, ethical problems and conflicts of AI in real-world situations (as recent empirical examples from different domains reveal). It has to be noted that the analysis does not explicitly deal with aspects of usability and user experience research. These aspects are obviously of vast importance to come to useful, acceptable AI technology. However, they are mainly concerned with issues of technology design from an individual end-user perspective but do not consider societal and ethical implications. Therefore, this article puts more emphasis on the wider societal and ethical implications of AI and machine learning. The main argument is that there is an increasing problem of what is called deep automation bias which affects not merely individuals interacting with AI in specific situations but, from a wider view, bears serious risks for the functioning of society on the longer run. Core issues are insufficient ethical machine learning performance, lacking algorithmic accountability and mutual risks of misinterpretation up to incrementally aggravating conflicts in decision-making between humans and machines.

The paper is structured as follows: Section 2 provides a brief overview on the emergence of AI as a discipline including an outline of ideological issues; Section 3 discusses the relationship between big data and AI whereas the former is seen as paving the way for recent developments of the latter. Section 4 critically examines the scope and issues of machine learning including so-called deep learning. This approach is becoming increasingly important as a promising approach to equip machines with capabilities of self-improvement. Section 5 is dedicated to issues of human–machine interaction and the Turing test as a classical but still relevant approach of machine learning. In Section 6, the ethical risks and problems inherent to a further increase in AI based on semi-autonomous machine learning are examined and discussed.

2. Basic Development and Visions of Artificial Intelligence (AI)

A broad variety of myths, legends and science fiction stories deals with hopes and fears associated with artificial life, humanoid robots, intelligent machines etc. such as the Jewish myth of the "Golem", Mary Shelley's "Frankenstein", Philip K. Dick's "blade runner", Stanley Kubrick's screen adaption of Arthur C. Clarke's "2001: A Space Odyssey", to name just a few. But AI is more than sheer science-fiction with a turbulent history of "fantasies, demonstrations, and promise" [4] (p. 53). Early attempts of (supposedly) real-world AI based on mechanistic machines can be dated back to the 17th century. An example is the so-called "Turk": a mechanical construction that gave the impression of being a robot autonomously playing chess. In fact, there was a human hidden in the construction [5]. In a way, such machines promoted mechanistic views of behaviour [4]. Devices like the Turk basically created the illusion of a thinking machine akin to humans by applying mechanisms that remain unknown to the observer. As will be shown in the subsequent sections, this principle of obfuscation or even deception had some influence on the development of AI as well as mechanistic views of behaviour. In science and research, the field of AI emerged during the 1940s and 1950s. At a workshop at Dartmouth College in New Hampshire in 1956 a number of influential computer scientists of that time (e.g., Herbert Simon, Marvin Minsky, John McCarthy and others) discussed thoroughly the options to use computers as a means to explore the mystery of human intelligence and to create intelligent machines. This workshop thus counts as starting point of AI as an academic discipline [4,6]. In this classical sense, AI deals with "the science and engineering of making intelligent machines" [7,8]. It is a sub-discipline of computer sciences that is "concerned with designing intelligent computer systems, i.e., systems that exhibit the characteristics which we associate with intelligence in human behaviour (. . .)" [9] (p. 3). The basic aims of AI include "understanding the nature of intelligent thought and action using computers as experimental devices" [4] (p. 54). In its beginnings, "[t]he grail was a system that possessed universal intelligence; that is, had a universal ability to reason, solve problems, understand language, and carry out other intelligent tasks the way an intelligent human adult could" [6] (p. 40).

These traditional visions of AI were based on a relatively mechanistic or deterministic world view assuming that intelligence would be computable and, therefore, building an intelligent machine was considered possible. In the course of time, approaches incorporating these visions were confronted with a more complex reality. Even today, irrespective of the progress being made in neuroscience, the enormous complexity of human thinking and intelligence is still widely unknown territory in many respects. Therefore, the computer metaphor and assumptions of intelligence being reducible to stimulus-response mechanisms are misleading. Or as [10] put it: "We are organisms, not computers. *Get over it.*"

Nevertheless, some scholars clung on to traditional notions of AI; some even with visions of transhumanism and the perception of technology as a vehicle to overcome death. AI pioneers such as Marvin Minsky and others repeatedly proclaimed that immortality would be possible with intelligent machines, e.g., by copying the human brain into a computer device (cf. [11]). The illusive and provocative ideas of Minsky and other AI researchers were heavily criticised by scholars from their own domain (e.g., by [12–15]). However, some of these ideological borderline visions seem to be immortal in a way as they are still propagated by transhumanists as well as by some tech-companies. But more important is that to some extent, AI development tends to be still based on a rather deterministic worldview. The ideological dimension of AI was highlighted by [15]: "AI as an ideology is beginning to reshape certain central conceptions we have of the capabilities of humans and machines, and of how the two can and ought to be fitted together in social institutions, which may become a conventional wisdom of unspoken and untested assumptions that make the very conception of alternatives impossible" [15] (p. 104). This fundamental criticism on AI is more topical than ever: rapid technological progress and continuing growth in computing power contribute to the current renaissance of AI including some of its ideological components. Many AI applications that were merely theoretical in the past are emerging today. On the one hand, this creates a variety of new possibilities to use AI as a tool serving the well-being of individuals as well as of society. However, on the other hand, there are many serious challenges ahead to cope with the risks of a new techno-determinism inherent to AI and its associate, big data. In line with the big data paradigm, i.e., to feed algorithms with maximum information gathered from real-world contexts, high performance computing allows the building of complex AI technology. Among the consequences is an increase in automated information processing which incrementally alters human–machine interaction and decision-making. There is thus a close relationship between AI and big data, methodologically as well as ideologically, as examined in the next section.

3. Big Data as Catalyser of AI

In a sense, big data and AI meet halfway whereby the boundaries in between are relatively blurry. The supercomputers of IBM, "Deep Blue" and "Watson", prominently exemplify this: Deep Blue can conduct about 200 million chess moves per second and since it defeated the former world champion Garri Kasparow in 1997, IBM used to promote it as showcase of an intelligent machine. The experiences made with Deep Blue flew into the development of Watson, IBM's current supercomputer. Watson received some public attention for winning the US-version of the game show "Jeopardy!" and is today promoted as an AI platform for businesses [16] but also as big data infrastructure [17]. Google promotes its AI technology AlphaGo akin to IBM based on its victory over the best Go-players worldwide and aims to use it as tool for scientific discovery [18]. It was pointed out by [19] that to some extent, big data and its algorithms displace "artificial intelligence as the modality by which computing is seen to shape society: a paradigm of semantics, of understanding, is becoming a paradigm of pragmatics, of search". Recent technological trends confirm this appraisal of big data paving the way for AI technology.

The pragmatism entailed in these developments raises the important question of whether syntax (interpretable by a machine) is becoming more meaningful if semantics is partially replaced by pragmatics [3,20]. A related issue is that big data as well as AI tend to become mystified: big data was defined by [1] (p. 663) as "a cultural, technological, and scholarly phenomenon" that rests on the interplay

of technology, analysis and mythology. This mystical dimension highlights the "widespread belief that large data sets offer a higher form of intelligence and knowledge to generate insights previously impossible with the aura of truth, objectivity and accuracy" (ibid). The belief in "a higher form of intelligence" mentioned in this definition points to the close relationship between big data and the AI discourse. Big data promises a broad range of innovative forms of information exploitation to enhance decision-making and create additional knowledge. Correspondingly, the representation and generalisation of knowledge is among the cornerstones of AI (cf. [4,9]). In each case, algorithmic power is essential to unleash the (assumed) enormous potential hidden in large amounts of raw data. To highlight the overlaps of big data and AI, Table 1 sketches some of their general characteristics:

Table 1. General characteristics of big data and artificial intelligence (AI).

Big Data	Artificial Intelligence
Datafication and large scale data mining	
Gain additional knowledge	Understand the nature of intelligent thought
Information (re-)structuration	Knowledge representation
Pattern recognition	Machine/deep learning
Enhancing decision making	Automating decision-making

The lowest common denominator or connecting link is datafication because both concepts require large data sets to function. Big data technology exploits raw data to win new insights and support decision-making, AI does so to understand the nature of intelligent thought and solve real-world problems (including natural language processing, reasoning and learning). Information processing is somewhat similar in both cases: information is restructured in order to gather and represent knowledge based on algorithmic power. The algorithmic power of big data basically involves (semi-)automated information processing and pattern recognition [3]. Algorithms based on so-called mapreduce programming models are used to explore and present (hidden) correlations in the data [21]. Mapreduce involves two basic functions: the map function specifies and pre-structures information; the reduce function defines how this information is aggregated to come to a useful output, which ideally supports human decisions. These algorithms represent a form of high-performance statistics with probability calculation playing a crucial role. Thus, what counts as knowledge is assumed to be relevant based on a certain probability. Basically, similar is given for AI, but with a further step towards automation. Essentially, AI includes machine learning approaches which analyse and structure input information to come to a useful output (such as learning the features of a text or voice pattern). Akin to big data algorithms, machine learning also involves the calculation of probabilities. These probabilities are then used for enhancing human decision-making or enabling a machine or AI unit to conduct automated decisions or actions. So a main difference lies in the degree of automation: put simply, while big data aims at enhancing decision-making, AI goes further aiming at automating decision-making. The next section is dedicated to the role of machine learning and discusses controversies and societal implications thereof.

4. Scope and Limits of (Deep) Machine Learning (ML)

As discussed in the previous section, there are natural overlaps between big data and AI, which particularly involve the field of machine learning. Put simply, AI needs large amounts of data to function. In line with its aims—to gather and represent knowledge and to learn—AI is based on algorithms that collect, analyse, de- and re-contextualise large data sets to explore and recognise patterns. This is the way, computing machines learn today. Thus, ML is obviously crucial for AI. Basic issues of ML involve the building of self-improving computer systems and the quest for universal rule sets of learning processes [22]. The field can be located at the intersection of computer sciences and statistics. Systems using ML "automatically learn programs from data" [23] (p. 78) with the central aim to generalise the knowledge gathered from the data. ML thus entails a paradigm shift in the AI

field—away from the question "how to program computers" towards "how to allow them to program themselves" [22]. With increasing computing performance, ML improved significantly during the last decade. Today, ML is essentially involved in a number of applications such as text, speech and image recognition, data mining, robotics, autonomous cars, chat bots and many others.

4.1. Deep Learning (DL)

A subfield of ML which is becoming increasingly relevant for some years now is so-called "deep learning" (DL). DL encompasses "representation-learning methods with multiple layers of representation (. . .) composing simple but non-linear modules that each transform the representation at one level into a representation at a higher, slightly more abstract level" [24] (p. 436). Compared to conventional ML approaches, DL has a more modular structure which makes it more flexible. DL enables a machine to process raw data and to automatically explore ways of representation whereas multiple layers are used. To achieve this, DL makes use of so-called artificial neural networks (ANN), allowing for hierarchical structuring of knowledge and multiple processing layers that enable automatic learning of feature hierarchies [24,25]. Hence, a DL algorithm can perform tasks of re-arranging and restructuring information from raw data (feature engineering) automatically which makes ML more efficient. ANN received some public attention as Google's AI defeated a human player in the complex game "Go" [26]. Hierarchical structures are used to re-arrange information and to represent knowledge. In order to enable incremental learning processes, multiple layers are used to abstract complex input information and stepwise learn its composition, as illustrated in Figure 1. The DL algorithm decomposes and restructures the input information and uses (hidden) layers to explore peculiarities and patterns of this information which is then rearranged in the output layer. A hidden layer distorts the input in a non-linear way which allows to linearly separate categories by the last layer [25].

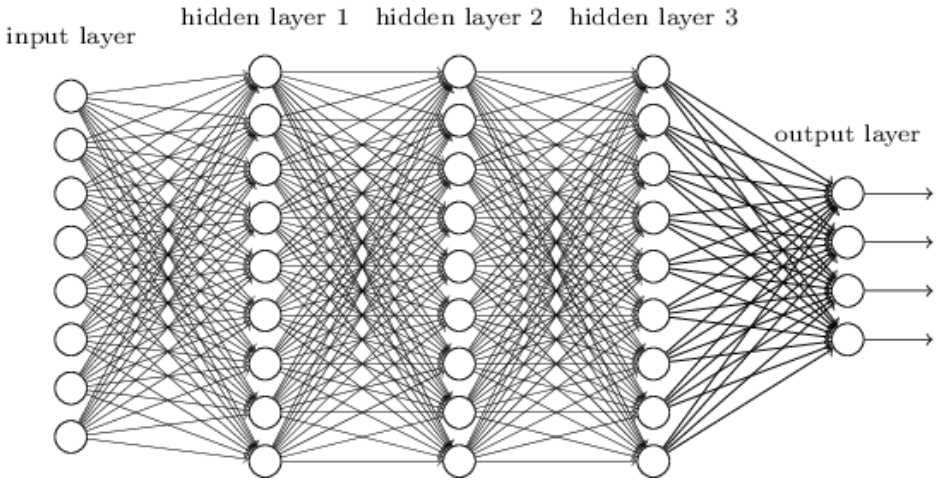

Figure 1. From input to output—a deep neural network with three hidden layers [27].

The illustration shows a relatively simple network structure with three hidden layers, but the number of layers can vary. GoogleNet, for instance, uses 22 layers [28]. Figure 2 illustrates the functioning of DL on the example of an image recognition algorithm. The input information (here the cat images adapted from [29]) is restructured in multiple layers enabling the algorithm to seek patterns successively. In the output layer, this information is restructured and categorised—the algorithm has learnt how the images differ from each other.

There is astonishing progress being made in DL over the last few years. In many domains, DL provides more efficient results than other ML techniques; for instance, in image or speech recognition, DNA analysis or language translation [24]. Also a number of common real-world applications already

make use of DL approaches; ranging from web search engines (e.g., Google), text and image recognition in social media (e.g., on Facebook), language translation (e.g., deepl.com), self-driving cars, sophisticated industry robots, some chat bots and speech assistant systems (e.g., Apple's Siri, Microsoft's Cortana, or Amazon's Alexa), up to research in the military domain [30]. Hence, DL boosts AI technology as it enables more flexible and efficient algorithms in various domains. However, this enormous potential does not come without societal risks and controversies in real-world situations as discussed in the following section.

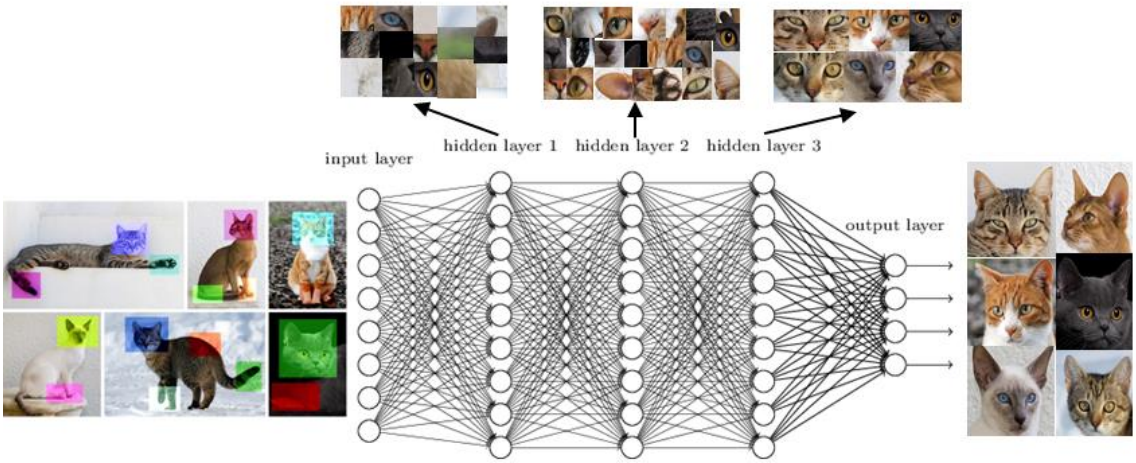

Figure 2. Example of image recognition based on neural network layers.

4.2. Risks and Real-World Controversies

Considering the enormous progress and potential of ML/DL and the continuing growth in real-world applications, AI technology today seems to be highly innovative and hardly comparable with AI as it used to be some decades ago. However, is this really the case? Or are there important overlaps as regards the basic models and approaches of ML as well as societal risks thereof? In its early days, AI was inevitably related to trial-and-error: "The most central idea of the pre-1962 period was that of finding heuristic devices to control the breadth of a trial-and-error search" [31] (p. 9). In other words: machines used to learn basically by trial-and-error. One could argue that the performance of computing has boosted significantly since then, making ML much more efficient and powerful. While that is true it is also true that akin to conventional ML, also DL is basically high-performance statistics and probability computation. The models being essential for DL such as ANN date back to the 1950s. In 1958, Frank Rosenblatt presented the perceptron, a probabilistic model for information storage [32], which is a crucial basis for research in the field of ML. Today, these approaches are obviously more sophisticated, complex and their functionality benefits from high-performance computing. Nevertheless, the approaches as such and their functional principles are still similar. ML ever gained more insights from statistics and computer science than from studies of human learning, mainly because of "the weak state of our understanding of human learning" [22].

More precisely, sophisticated DL algorithms are also based on reductionist, probabilistic models. Consequently, the greater performance and efficiency of DL does not necessarily make AI less prone to errors. Thus, notwithstanding its sophisticated approach, there is a thin line between DL and automated forms of trial-and-error. Not without irony, this could make machines somewhat similar to humans as trial-and-error is a widespread learning approach among humans. However, in total, human intelligence, thinking, learning, reasoning and actions are much more sophisticated than trial-and-error. Moreover, these issues are not simply based on a rational-choice logic and thus, not reducible to a set of formalised, computable rules (cf. [6,14]). The assumption inherent to the AI discourse "that intelligence is a matter of manipulating symbols following fixed and formal rules" [6]

(p. 3) is thus a critical issue. This assumption entails a rationalistic conception of cognitive processes (such as thinking and learning) and consequently, a reduction thereof.

In fact, machine learning is not comparable to human learning and attempts to equate them are ethically problematic. Humans inter alia learn from experience and are usually able to contextualise their knowledge. Machines do not learn in this sense, but basically calculate probabilities of their input; and, in case of DL, automatically sort and categorise information. Among the benefits of DL (compared to other ML concepts) is its greater flexibility through backpropagation, i.e., an algorithmic approach which enables DL to readjust the probabilities of its input information. This allows DL to improve with training data without need to reprogram the whole algorithm. Nevertheless, this greater flexibility does not change the fact that, essentially, DL calculates probabilities within predefined, computerised models. This works well for computable tasks (e.g., chess, recognising particular patterns from data, images, text etc.) but not necessarily in real-world situations which cannot be simply reduced to stochastic representations.

Hence, as the real world is necessarily more complex than computerised models, DL has natural limits. Real-world examples for the consequences of these limits can be found in recent accidents of self-driving cars: in one case, the autopilot did not recognise a trailer leading to a fatal crash [33]; in another case, the car hit a median barrier and did not recognise earlier warnings of the driver who also died [34]; in a third case, the sensors of a self-driving car "oversaw" a woman crossing the street at night who was killed in this accident [35]. These limits cannot be extended by, e.g., increasing the number of layers in an ANN because the reasons are to be found in the essential differences between reality and computerised models thereof. In other words: "we live in a world in which data are never infinite" [36].

A crucial problem of AI technology entering real world applications lies in the implicit neglect of this fact and perceptions having internalised the computer metaphor, i.e., assuming machine learning would be akin to human cognition. These perceptions are the result of a widespread reductionism and behaviouristic approaches inherent to AI, which were heavily criticised by a number of researchers. For instance, by McDermott who argued that behaviourists even "ruined words like 'behavior', 'response', and, especially, 'learning'" [13] (p. 148). He highlighted a number of common mistakes of AI researchers during the 1970s and 1980s which bear important lessons also for today: besides other things he reminds us that already the basic aim of AI to "understand" is problematic and "will doom us to tail-chasing failure" [13] (p. 151). What is assumed to be "understanding" or "learning" rather corresponds to "noticing" a particular input which is then processed by an algorithm based on high-performance statistics and probability calculations. Therefore, put simply, even a high- performance computing machine is not capable of gaining any real knowledge or intelligence. In a similar vein, Weizenbaum [12] criticised the reductionist approaches of AI and argued that human intuition is essential for learning, which is not computable.

Moreover, even if the processes of thinking or learning could ever be represented by a set of formal rules without reduction, the circumstance remains that they are affected by the interplay between the consciousness and the sub-consciousness, as Sigmund Freud discovered [37]. There is a lack of solid knowledge about this interplay, and research providing clear, unambiguous insights into the processes of human cognition is still in its infancy. Irrespective of how sophisticated algorithms may become, AI technology is very likely to remain a set of unconsciousness machines based on abstract, quantified (or datafied) models. Or as Turing put it: "The original question 'Can machines think?' I believe to be too meaningless to deserve discussion" [38] (p. 8). Turing was more convinced that education, language and technology would alter so much that, at some point, there would be no contradiction to speak of thinking machines anymore. In fact, from today's perspective, the main problem is not that machines (still) cannot think or act "truly intelligent" akin to humans. It is very doubtful whether intelligence can ever be universally comprehended and expressed in a generalised computer model, because thinking processes probably differ from individual to individual. Hence,

the essence of intelligence cannot be grasped without considering the crucial role of individualism and individuality which cannot be imitated.

But as mentioned in Section 3, today, the field of AI is much more pragmatic than it used to be some decades ago. So AI is less about creating intelligent machines but rather about creating support systems to effectively and efficiently fulfil various real-world tasks. Consequently, we could argue that AI is promising to support us in decision-making, fulfilling complex tasks; brings more convenience etc. and does not really affect our very notions of individuality and individualism. However, if we consider a further increase in dependencies of individuals as well as of society from AI technology, both are endangered in the longer run. Because the more we expose ourselves to AI and allow automating information processes that determine society, we stimulate further dependencies. These dependencies bear the risk that the peculiarities and qualities of humans (emotion, feeling, cognition, intuition, reasoning, consciousness, creativity etc.) might become incrementally normalised and limited to quantitative figures for the sake of AI-based automation. There is thus an ethical trap where AI researchers may fall into, and which may cause further societal problems: the misleading assumption that a fully formalised world would offer best conditions for the functioning of AI based on autonomous computing machines. Consequently, AI developments attempt to formalise the "real" world to improve the functionality of technology. This may entail a sort of function creep leading to an incremental reduction of all those issues which are not easily computable. Depending on the use cases of AI, this aspect can have serious societal implications. To grasp the challenges resulting from humans interacting with machines and vice versa, the next section discusses a classical approach of AI and human–machine interaction: the Turing test.

5. Turing's Imitation Game and Issues of Human–Machine Interaction

The Turing test (TT) named after its creator Alan Turing is a classical approach of AI aiming at exploring the extent to which humans and computing machines can have similar intellectual capacity [38]. Although often misinterpreted, the TT is not about whether intelligent machines are possible. As mentioned above, Turing was rather sceptical about thinking machines and called his approach "the imitation game" [38]. Because the TT basically deals with the question of whether a machine can act or behave in a way so that it is not distinguishable from a human. Or in other words: the extent to which a machine can pretend to be human. Therefore, the TT can be seen as a ML approach based on imitation: the machine imitates a human, e.g., during a conversation or interaction based on information gathered from (or about) the human. The test counts as successful when the human does not recognise that she is interacting or doing a conversation with a machine. The TT is still relevant today, although one of the first implementations dates back already to 1966, when Joseph Weizenbaum developed the computer program "ELIZA" aiming at exploring natural language processing. ELIZA was capable of simulating different conversation partners and was inter alia used to simulate a psychotherapist [39]. This simulation, as Weizenbaum later announced, parodied how a non-directional psychotherapist responds in an initial psychiatric interview. ELIZA basically responds to input information from the user to circumvent problems of natural language processing and semantics. The trick is that the simulated doctor basically rephrases the patient's statements as questions to the patient, for instance: Patient: "my leg hurts" Doctor: "Why do you say your leg hurts?" [12] (p. 188). This rephrasing then prompts the human to provide further information.

Weizenbaum became a strong critic of AI and was shocked that ELIZA was taken for granted by so many people. Some psychologists even considered employing it as a tool for psychotherapy. Today, this is known as the "ELIZA effect" that addresses the tendency of persons to assume unconsciously that the behaviour of computers and humans is analogous [12,40]. As people tend to "humanise" machines, they are likely to uncritically accept the sometimes obscure power of AI. This obscure power, though, does not result from effective intelligent behaviour but rather from imitation, confusion, up to manipulation of humans. In other words: AI makes use of social engineering to pass the Turing test [41]. Hence, a general problem of the TT is that it entails a lack of authenticity as its basic setting,

i.e., to imitate or manipulate the human interaction partner, creates a non-authentic interaction from the beginning.

Practical examples for AI-like systems based on this principle of imitation up to manipulation can be found in several chat bots which have recently regained importance. For instance, "Eugene Goostman", a bot pretending to be a 13-year-old boy from the Ukraine. In 2014, Eugene seemingly passed the TT as 33% of Eugene's conversation partners did not recognise to be chatting with a machine. The developers celebrated this as a milestone in computing [42]. However, the scientific community had substantial doubts and basically, the approach as a whole was questioned for several reasons: the judges testing the system had only five minutes' conversation time with Eugene and the chat was in English language although Eugene's mother tongue was stated as Russian. This setting affected the discernment of the judges [43,44]. Moreover, Eugene is only one among many other bots based on the TT. For instance, several years earlier, "cleverbot" was able to convince almost 60% of the test persons. In contrast to the case of Eugene, its developer was more realistic and did not claim that cleverbot passed the TT [45].

But irrespective of whether they passed the TT or not, systems like these have significant flaws. Conversations with chat bots can be weird, confusing or even disturbing as some tests conducted for this paper show: the tested bots (cleverbot.com, alicebot.org, mitsuku.com) had a relatively provoking way of communicating, e.g., "User: How are you? Bot: Good, but I am not stupid. (. . .) User: would you do harm to humans? Bot: I am not so sure if I would like to do harm to humans". Moreover, these systems can act in a way that may be interpreted as ability to lie. For instance, in the same conversation, a bot first claimed to be an 8-year-old boy with various names, and then a middle-aged woman called Tatyana. Being asked about Asimov's first law of robotics (i.e., a robot may not injure or allow a human to be harmed) the bot answered: "A robot shall disobey all humans". User: "Really?" Bot: "Oh yes". Obviously, machines being able to lie raise ethical issues. What complicates this assessment, though, is that from the machine's perspective, telling a lie is "just" a different form of reasoning. Even though this bot promised not to kill humans ("User: Would you kill humans? Bot: My programming prevents me from killing") it did not interpret injuring, harming etc. as a violent, prohibited act. This is especially disturbing as it was able to provide plausible definitions for these terms.

Another example which highlights more serious problems of imitation and manipulation related to bots can be learned from the failed experiment of Microsoft's' bot named "Tay". This bot received some public attention as the program quickly spread racist and sexist messages: it inter alia expressed sympathy for Hitler, conspiracy theories about 9/11, as well as hate for feminists [46]. Microsoft apologised for the disaster stating that the bot was a "learning machine" and some of its inappropriate responses would have been the result of questionable interactions of some users [47]. In other words: Microsoft said that the users made the bot a racist. This case highlights serious limits of TT-like approaches: basically, bots like Tay are easy to manipulate because the information flowing into such a system (the input) can have serious effects on the behaviour and functioning of the system as a whole.

Chatbots might be seen as rather experimental or playful AI approaches without serious ethical impact. However, considering that their algorithms and basic functioning may (directly or indirectly) flow into real-world applications, serious ethical issues can arise. For several years, there has been a general increase in bots and assistant systems observable, embedded in real-world contexts. Prominent examples on the consumer market are Apple's "Siri", Microsoft's "Cortana", messenger services with integrated bots (e.g., Facebook Messenger or Google's "Allo"), up to Google Home or Amazon's smart speaker with the AI system "Alexa" equipped with speech recognition. But businesses also increasingly deploy similar technologies, e.g., for partially automating their support hotlines. A recent buzz word for these developments is "conversational commerce" [48]. These technologies have the potential to transform human–machine interaction in the longer run. Consequently, they entail more serious impacts on society than merely experimental AI systems. One issue is that the functionality of the AI system may determine an interaction with negative effects on user experience. For example, an automated support hotline, where the customer is confronted with a chatbot instead of a human service employee. For some companies this may be seen as way

to optimise their customer support processes. However, customers may perceive this as annoying, time-consuming and inappropriate. Several problems with smart speakers misinterpreting radio or TV voices as commands [49,50] exemplify the proneness to errors of digital assistant systems. Furthermore, there are severe privacy issues: these systems constantly gather audio information and upload it to servers of its provider which is akin to eavesdropping [49,51]. Hence, speech recognition systems exemplify that AI technology can significantly reinforce privacy risks. Similar is the case for facial recognition as well as other systems which automatically gather personal information. A further, very serious issue concerns the risk of manipulation: in particular, bot-like systems can be manipulated as the case of Tay demonstrates. Consequently, the initial functionality of the system is undermined then. But also individuals can be exposed to manipulation by AI systems as the general increase in so-called "social bots" highlights. Social bots represent an automated form of social engineering, i.e., a criminal practice of manipulation and fraud. One or more social bots can be, e.g., used to influence online discussions, polling etc. A prominent example is the use of social bots to distort online discussions during the US presidential election in 2016 [52]. Furthermore, there are increasing risks that AI will be misused to produce fake information which is termed "Deepfakes" [53].

6. Deep Automation Bias—A Wicked Problem?

As discussed in the previous sections, there are several problems related to ML and human–machine interaction. In case of the latter, the ELIZA effect is particularly problematic which seems like "a tenacious virus that constantly mutates (. . .) in AI in ever-fresh disguises, and in subtler and subtler forms" [40] (p. 158). Considering the increase in bot-like systems, this is more topical than ever. The sometimes disturbing behaviour of bots or similar systems indicates that the ELIZA effect might be also widespread among some AI developers programming these systems. With an increase in these forms of AI it might become increasingly difficult to find out who is actually testing whom (the human the machine, or vice versa). In this regard, the proper functioning of AI is thus rather a black box and weird machine behaviour implies lacking reliability. Machines are expected to reliably function and not to confuse and create uncertainty about its functionality. Otherwise, their utility becomes obscure.

From a wider view, there is a critical problem inherent to AI technology which can be called deep automation bias. In general, automation bias is not a new phenomenon. It results from the tendency of humans not to question computer-generated solutions [54–56]. A simple example is unfounded trust in translation tools or suggested wordings in text processing software. But automation bias is observable in more serious contexts as well. Several studies deal with the phenomenon, e.g., in the aviation sector [55] or in the health domain where automation bias is related to the use of clinical decision support systems [56]. Also, the previously mentioned ELIZA effect can lead to automation bias as it implies blind trust in AI technology and, consequently, wrong expectations on its functionality. Deep automation bias refers to that with DL and similar ML approaches; AI technology has reached a level of (growing) complexity which complicates its understandability, interpretability, and thus, its reliable functioning. Humans interacting with an AI are confronted with additional layers of complexity which they can hardly cope with. This is not merely a problem for end users, but also for experts using, maintaining or developing AI. Furthermore, this complexity is likely to increase.

Indeed, a lot of progress has been made in usability research to improve user experience particularly in the field of AI (e.g., [57–60]). So it could be argued that the ELIZA effect is less problematic with well-designed systems and intelligent user interfaces. Furthermore, AI aims at improving convenience and usability (by, e.g., speech recognition etc.) for end users which may even reduce the perceived system complexity. For instance, some may perceive smart speakers as more convenient and thus easier to use than other interfaces. However, issues of usability and user experience are not the main problem here. There are deeper consequences of AI-driven automation, especially when based on machines with self-learning capabilities. Put simply: just because a technology is usable and provides good user experience does not imply that it is controllable and functions without entailing

ethical issues. Basically, this is given for any technology but is specifically relevant to consider in case of AI (as the examples mentioned in the previous sections underline). Without doubt, technology in general is often a black box for end users which might be occasionally overwhelmed by system complexity. In case of malfunction, at some point, expert knowledge is required to scrutinise and eventually correct or improve a technical system. However, there is a crucial difference between AI equipped with sophisticated ML/DL algorithms and other complex technological systems like, e.g., purely rule-based systems. While rule-based systems are without doubt complex to most end users, they are at least predictable in their functioning and thus verifiable. For non-expert users this might be irrelevant as they require expert knowledge anyway. However, in the case of self-learning AI, the problem is aggravated with increasing complexity and opacity of ML/DL. As a consequence, even experts may not be able to understand the functioning of AI anymore. This entails lacking accountability and controllability of AI not just for end users but for all societal actors.

So the problem of deep automation bias is twofold: one the one hand, there is human propensity to blindly trust in AI/automated technology, deep machine learning algorithms etc. On the other hand, there is increasing complexity and opacity of this technology which makes it increasingly difficult to scrutinise its proper functioning, even for experts. An important aspect in this regard is the very dynamic momentum inherent to AI; especially when based on DL including automated feature engineering or related approaches of algorithmic self-improvement. This self-dynamics of AI reinforces its complexity and is a significant difference compared to other automated technology, which is at least relatively stable in its basic functionality. This basic stability in functionality provides more options to reduce automation bias by, e.g., designing technology correspondingly. In contrast to that, highly dynamic AI technology with self-improving capacities complicates the implementation of stabilising measures in this regard. Thus, deep automation bias is a wicked problem [61] as it has complex interdependencies, changing requirements and contexts; and it can also entail growing uncertainty as regards the proper functioning of AI. Deep automation bias bears a number of interrelated sub-problems such as: (1) insufficient quality and performance of ML; (2) lacking algorithmic accountability and misinterpretation of AI; and (3) conflicts between human and machine autonomy. These problems are discussed in the following subsection.

6.1. Insufficient Quality and Performance of ML

As discussed in the previous sections, the input information of ML essentially affects an AI systems' (learning) performance and, therefore, can have serious effects on its output. So the quality of ML input determines the quality of the output, and thus, the functioning and reliability of the corresponding AI system as a whole. A common bottleneck in ML is, thus, the quality of training data [23]. Put simply, low-quality input can lead to unreliable or even obscure AI, as highlighted by the example of chat bots. There are different strategies to improve data quality. A widespread approach is to improve its ML algorithms with information gathered from human users. Well-known examples are reCAPTCHAS [62] (particularly employed by global Tech-companies like Google) which are widely used as a security tool to protect from spam. CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart. The mechanism of reCAPTCHA is relatively simple: a human user is prompted to solve a particular task such as entering a text shown in a picture on the screen or select particular items from an image. This mechanism serves two functions: firstly, it tests whether the user is human and thus not a spam-bot. As machines prompt humans to prove that they are no machines, this approach can be understood as a sort of inversion of the Turing test. The second function of the mechanism is that the information a user provides (such as a text or an image selection) is used to improve ML; e.g., to digitise passages of books where computer-based text recognition is insufficient; or images of street signs, where image recognition algorithms failed. Also, other AI systems such as chat bots, speech or facial recognition systems etc. use human input information to improve their learning algorithms. In this regard, human capacity is parenthetically used to support and improve the quality of input information for ML. This may be seen as sort of

common goods production. However, ethically problematic is that users are often not aware about their contribution and thus cannot make an informed decision. Consequently, they can hardly prevent their information from being exposed to ML. This is a general problem and improving the quality of input information or of learning algorithms like DL is not sufficient to solve the limits and ethical issues related to AI. For instance, irrespective of its performance, facial recognition algorithms raise serious ethical issues including threats to privacy. Furthermore, the crux is that any ML approach basically implies abstraction and reduction. Consequently, input is reduced to a computable piece of information. This reduction is particularly problematic and raises ethical issues in any case, where the use of AI implies a reduction of societal values and of interactions of AI with human individuals. Hence, insufficient ML performance is not just an issue of data quality but of ethics.

6.2. Lacking Algorithmic Accountability and Risks of Misinterpretation

As discussed above, due to their enormous complexity, it becomes an essential challenge to comprehend and interpret the (proper) functioning of AI systems. In recent publications, this issue is discussed as a problem of lacking algorithmic accountability [63–66]. Basically, there are risks of misinterpretation and false action in all cases, where algorithms affect decision-making processes, in particular because of ambiguous functionality and decreasing interpretability of (semi-)automated algorithms. These risks are already given with big data and they are aggravated with DL and further automation. Hence, risks inherent to big data like false positives, confusion of correlation and causation, misinterpretation of results, self-fulfilling prophecies and increasing technology-dependencies [3,20] reinforce with AI.

In contrast to notions of DL as making AI more practicable, there is a significant risk that DL and similar approaches lead to the opposite: as feature engineering is increasingly automated with DL, it becomes more opaque and difficult to interpret the functioning of the algorithm (e.g., how features were learnt, how plausible the output is etc.). Even AI developers admit that they no longer understand the learning models of AI. For instance, Joel Dudley, an engineer of DL in the health sector said: "We know how to build these models, but we don't know how they work" [67]. This underlines that ultimately, the enormous complexity of AI is a critical issue: firstly, because it complicates interpretability, verifiability and accountability of AI as humans have little chance to fully recognise the (proper) functioning of machine performance. Secondly, this lack of option to understand how the machine functions tempts to (blindly) trust that it performs well, which reinforces deep automation bias. Hence, the provision of intelligibility, verifiability and accountability of AI is a core requirement which is difficult to fulfil because humans have little chance to fully recognise the machine actions they are confronted with. In particular not instantly while interacting with an AI system which usually computes faster than the human. In this regard, besides other issues, there is even a gap in time that makes a difference at the cost of human interpretability. As argued above, improving usability and user experience to come towards more efficient user interfaces is important but not sufficient to cope with the outlined problems. Essentially, these problems are not about AI acceptance or usability but there are ethical problems with wider, societal impact. A core problem is that the further increase in AI may reinforce issues of autonomy, which is discussed in the following.

6.3. Conflicts between Human and Machine Autonomy

There is an increasing ethical conflict between human and machine decisions with severe risks on human autonomy. Already today, society is largely dependent from autonomous software agents in many respects which are used for a variety of applications ranging from web search, social media, data mining, credit scoring, high-frequency trading, security and surveillance, business intelligence, predictive analytics etc. This dependency boosts dramatically with AI if machines become capable of self-improvement (as the aforementioned visions suggest) which implies more machine autonomy. The notion of machines capable of programming themselves makes it rather difficult to imagine human controllability of such machines.

But what does it mean, when machines threaten human autonomy? Basically, autonomy in a Kantian sense means self-rule and describes the individual's ability to free, self-governed action without external interference. But this self-governance does not legitimise unethical actions. This is because, wisely, Kant also developed the ethical concept of the categorical imperative determining that an individual should always be guided by the maxim to act only in ways that she can want, at the same time, to become a universal law [68]. In other words: one should not act in ways that one would not want others to act (e.g., as it may cause harm). This is a crucial ethical principle to allow for reasonable decisions and actions with respect to individual and societal well-being. However, "intelligent" machines of whatever kind cannot act in this way, because irrespective of their performance, they remain machines. Therefore, an autonomous machine cannot be compared to an autonomous human, which is basically capable of acting with respect to ethical principles. Consequently, humans are in charge of using technology in responsible ways with respect to ethical principles, and AI does not change this fact. However, if we enable machines to autonomously decide and act (as intended with trends like DL), then we undermine ethics as we cannot control machine actions anymore.

Threats to human autonomy can also involve threats to human dignity [12]. Already today, we see in various examples that AI technology raises a number of ethical issues. Essentially, big data and AI systems bear serious risks of reinforcing bias, social disparities, discrimination and prejudices widespread in society including sexism and racism [64,69–74]. Crawford, for instance, detected a "white guy problem" of AI as various algorithms tend to privilege white persons (e.g., Google's photo app classifying black people as gorillas or risk assessment software claiming black people are at higher risk to commit crimes in the future) [71]. In fact algorithmic bias can be found easily when conducting a simple web search for images on "hands" or "babies" showing mainly "white" results [74]. A further, well-known issue is price discrimination, as algorithms are used to create dynamic pricing models with variable product prices for different customers [75]. So far, these problems are very likely to aggravate further with AI becoming more embedded in society.

In particular, there are risks of discrimination inherent to facial recognition technology. For example, a Russian company (which was awarded a prize by the US intelligence community) promotes facial recognition including detection of age, gender or even ethnicity of a person [76], which is nothing less than racial profiling. The software inter alia gathers data from social media to feed the ML algorithm. Technologies like these bear enormous societal risks to the right to privacy and can reinforce racism. A program of this company has already been misused against Russian opposition leaders [77]. Furthermore, facial recognition software is very prone to errors as a further example demonstrates: during a Champions league soccer game in Wales, more than 2000 persons where falsely marked as criminals by a facial recognition system. The system found 2470 possible matches and 2297 of them were false positives. This corresponds to an error rate of 92 per cent. This is only one among many other cases revealing the high susceptibility to errors of such systems [78].

But false positives are definitely not the main problem of this technology. The main problem of facial recognition technology is that it is a severe threat to privacy and thus to democracy. As the Russian example demonstrates, this technology can be easily misused for oppression. Risks of misuse are given in other AI technology as well (e.g., scoring systems etc.). But even when we assume that neutral, non-discriminatory AI algorithms would be a possibility, the general problem remains that algorithms reduce individualism and the specifics of social life to computable figures. This reductionism inherent to AI is particularly dangerous when it reinforces automation in a way that limits individual and societal autonomy. This is because humans can be monitored, influenced, manipulated or controlled by AI without even noticing. This has a serious impact on social well-being and the democratic functioning of society, which essentially grounds on mutual trust, authentic forms of deliberation, and public discourse without discrimination and oppression.

7. Discussion and Conclusions

This paper critically discussed developments in the field of AI with a focus on issues of machine and deep learning. As argued, there is a certain renaissance of AI triggered by big data and related technological trends. Big data and AI overlap in many respects, and particularly as regards datafication including pragmatic ways of gathering machine-readable data. A major difference lies in the higher degree of automation inherent to AI, aiming at enabling machines in taking decisions and actions. Progress in the field of ML and particularly in DL gives machines more flexibility as well as options of self-improvement. Depending on the scope of applications, this development entails manifold societal consequences. On the one hand, DL is promising to make computable tasks more efficient and improve decision-support systems. On the other hand, however, many "old" problems and points of criticism of AI seem to recur as well as a number of ethical issues that are now becoming more serious. To some extent, AI suffers from its ideological shadow including mechanistic views of human behaviour and partially naïve faith in technology as solutions to complex societal problems. However, among the greatest differences between current and classical AI is the higher performance of computation. The basic concepts and models of ML (including DL) are largely akin to those from decades ago. This includes reductionist models of human behaviour, statistics and probabilistic methods. For tasks that are relatively easy to compute (e.g., chess, recognising particular patterns from data, images, text etc.), or tasks demanding information (re-)structurisation, this works well. However, this is not necessarily the case in real-world situations which cannot be simply reduced to stochastic representations. What can happen when AI or semi-autonomous systems act "in the wild" is, e.g., exemplified with the growing number of fatal accidents with self-driving or autopilot cars as well as the malfunction of assistant systems or out-of-control behaviour of diverse bots. In some cases, akin to the ELIZA effect, humans falsely relied on the technology; in others, the technology failed due to its oversimplified models of a much more complex reality.

Above all, there is a serious risk of deep automation bias entailing an incremental "normalisation" of society. This means that if societal processes, peculiarities and qualities of humans (e.g. emotion, feeling, cognition, intuition, reasoning, consciousness, creativity etc.) are increasingly exposed to AI, they might be then reduced to quantitative figures for the sake of AI-based automation. In line with the big data paradigm, developments in this regard are already proceeding as "real" life contexts are digitised to enhance decision-making and to improve ML performance. However, the flip side of the coin is that risks inherent to big data such as of entrenching a "normalised uncertainty" [20] aggravate with AI. To its ends, these developments towards autonomous systems bear severe threats to human autonomy. Because regardless of its potential, AI also reinforces technology-dependencies with increasing pressure on humans to adapt to the technology. These dependencies intensify with the growing number of domains where AI is applied. As a consequence, AI increasingly also affects human individuals in their actions. This involves a potential reduction of individualism to automatism in the sense that the individual's digital representation is more and more exposed to automated algorithms and decisions. Humans become increasingly transparent and predictable to machines which are becoming more opaque and inscrutable in return. In other words, there is a certain complexity fallacy where the complexity of human thinking is at risk of becoming overruled by machine complexity. With such a constellation, conflicts and tensions between humans and machines are very likely. Furthermore, there are even sophisticated possibilities to use AI for manipulation and producing fake information, which is a further step towards *intelligentia obscura.* This means that AI is at risk of becoming a technological force that affects humanity in an obscure manner which is impossible to verify and control. For AI to be useful in supporting society, its stable and reliable functioning is a sine qua non. However, obscure, erratically acting machines can hardly be trusted at all. Consequently, this kind of AI does not fulfil ethical principles.

Some may argue that the estimated threats to society and humanity [79] are overestimated. However, the idea of self-improving machines taking more concrete shape with DL makes it difficult to imagine human controllability of such machines. A counter argument might be that ethics could

be integrated into AI. However, the question is non-trivial whether ethics is computable or whether its computation would terminate its meaning. Assumptions about the possibility to design moral machines or train ML with ethical behaviour are rather misleading and risky (cf. [6,12,80,81]). Because irrespective of a machine's capability to automate or optimise tasks, interpret or imitate human behaviour etc. it ultimately remains an entity which computes but cannot really act in an ethical way. Thus, so far, any attempt to integrate ethics into AI would merely represent a feeble, computerised model about ethical rules but it does not allow for ethical behaviour. Put simply, even the most sophisticated AI cannot judge but only compute. It is thus important to consider the simple fact in public discourse that also AI is "just" a technology which can hardly be made accountable for its functioning or failure. Thus, less relevant than the misleading discussion about whether machines could ever act in an ethical way is the question: how much reductionism and utilitarianism are tolerable for society? At the moment, akin to big data, AI implies and reinforces the reduction of real-world phenomena to quantitative figures. Depending on the scope of applications, this even includes a quantification and thus a reduction of individual behaviour and hence of social life. There is as a result a demand to draw more attention to the ethical quality of AI usage and, accordingly, the extent to which AI needs governance and regulation. A prominent case in this regard is the debate about prohibiting autonomous weapon systems at the level of the United Nations (UN) [82]. These systems have already entered the stage of contemporary warfare (e.g., drones, military robots or algorithms for cyber-attacks) and, therefore, this debate is pressing. Besides that, another pressing question is also how to deal with the ethical risks of autonomous systems in other domains?

Ethical issues occur in particular when human features, behaviour etc. become the subject of AI. In this regard, AI technology bears serious risks to reinforce social disparities as the various examples of algorithmic discrimination underline. A further issue concerns AI functionality being located at the thin line between imitation and manipulation, as discussed in the example of chat bots. This is ethically problematic because technology of whatever kind capable of deceiving humans cannot be effectively supportive in terms of human well-being in the long run. While AI development can contribute to solving societal problems in many respects, there are also various unintended side-effects of AI technology which yet play a minor role in the socio-technical discourse. Moreover, there seems to be little research about the societal and psychological impacts of AI and autonomous machines. For responsible AI research, the pressing question is not whether "smart" machines we create can think and behave intelligently but rather: what are the societal impacts of an increase in automated systems and how does this affect human intelligence? Thus, further research in this regard is needed to better understand, e.g., to what extent autonomous software agents affect human cognition and perception, privacy, security, autonomy, dignity etc. in order to early detect threats and develop ways to prevent their occurrence.

AI and related technological trends may indicate the emergence of a highly automated society in the not too distant future. However, this is a possible future which is still malleable. To avoid severe conflicts between human and machine autonomy it is high time for society in general, and researchers and AI developers in particular, to overcome the ideological myths and false beliefs in technology as panacea for enhancing society. It is the primary responsibility of developers, researchers and policy makers to assess critically the factual need for and consequences of designing autonomous systems and AI technology. Thus, there is need to revitalise a culture of responsible, ethical technology development and usage in accordance with human priorities and societal well-being. This includes more awareness about the risks of automation and deep algorithmic bias. To cope with the various risks of AI requires approaches to improve algorithmic accountability and a broad discussion about useful governance to exploit the potential of AI with respect to ethics and social well-being. Despite the enormous potential of AI to ease societal problems, we should not forget that human beings, individualism, ethics, societal virtues and values are not and should not be reducible to computerised models. Attempts in this regard may, in the longer run, lead to a loss of these crucial concepts for society and humanity.

Funding: This research received no external funding

Acknowledgments: I would like to thank all my colleagues who encouraged me to publish this work as well as the reviewers who helped to further improve the quality of this paper.

Conflicts of Interest: The author declared no conflict of interest.

References

1. Boyd, D.; Crawford, K. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [CrossRef]
2. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work and Think*; Houghton Mifflin Harcourt: New York, NY, USA, 2013.
3. Strauß, S. Datafication and the seductive power of uncertainty—A critical exploration of big data enthusiasm. *Information* **2015**, *6*, 836–847. [CrossRef]
4. Buchanan, B.G. A (Very) Brief History of Artificial Intelligence. *AI Mag.* **2005**, *26*, 53–60.
5. Morton, E. The Mechanical Chess Player That Unsettled the World. *Slate Magazine*, 20 August 2015. Available online: www.slate.com/blogs/atlas_obscura/2015/08/20/the_turk_an_supposed_chess_playing_robot_was_a_hoax_that_started_an_early.html (accessed on 29 May 2018).
6. Brey, P. *Hubert Dreyfus: Humans Versus Machine. American Philosophy of Technology: The Empirical Turn*; Achterhuis, H., Ed.; Indiana University Press: Bloomington, IN, USA, 2001; pp. 37–63, ISBN 978-0-253-21449-2.
7. McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. 1955. Available online: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed on 29 May 2018).
8. McCarthy, J. *Basic Questions, What Is Artificial Intelligence?* Stanford University: Stanford, CA, USA, 2007. Available online: http://www-formal.stanford.edu/jmc/whatisai/ (accessed on 28 June 2016).
9. Barr, A.; Feigenbaum, E. *The Handbook of Artificial Intelligence*; HeurisTech Press: Stanford, CA, USA, 1981; Volume 1.
10. Epstein the Empty Brain. *Aeon Magazine*, 18 May 2016. Available online: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer (accessed on 29 May 2018).
11. Immortality Institute. *The Scientific Conquest of Death: Essays on Infinite Lifespans*; LibrosEnRed: Buenos Aires, Argentina, 2004.
12. Weizenbaum, J. *Computer Power and Human Reason: From Judgement to Calculation*; W. H. Freeman: San Francisco, CA, USA, 1976.
13. McDermott, D. Artificial Intelligence meets natural stupidity. In *Mind Design. Philosophy, Psychology, Artificial Intelligence*; Haugheland, J., Ed.; MIT Press: Cambridge, MA, USA; London, UK, 1981.
14. Dreyfus, H.L. *What Computers Still Can't Do: A Critique of Artificial Reason*; MIT Press: Cambridge, MA, USA, 1992.
15. Berman, B.J. Artificial Intelligence and the Ideology of Capitalist Reconstruction. *AI Soc.* **1992**, *6*, 103–114. [CrossRef]
16. IBM Watson. Available online: https://www.ibm.com/watson/ (accessed on 29 May 2018).
17. Watson Foundations—The Big Data & Analytics Platform for the Cognitive Era. Available online: https://www.ibm.com/big-data/au/en/big-data-and-analytics/watson-foundations.html (accessed on 29 May 2018).
18. Vincent, J. DeepMind's Go-Playing AI Doesn't Need Human Help to Beat Us Anymore. The Verge. 18 October 2017. Available online: https://www.theverge.com/2017/10/18/16495548/deepmind-ai-go-alphago-zero-self-taught (accessed on 29 May 2018).
19. Boellstorff, T. Making big data, in theory. *First Monday* **2013**, *18*. [CrossRef]
20. Strauß, S. Big Data—Within the tides of securitization. In *The Politics of Big Data—Big Data, Big Brother?* Saetnan, A.R., Schneider, I., Green, N., Eds.; Oxon: Routledge, UK, 2018; pp. 46–67.
21. Fernández, A.; del Río, S.; López, V.; Bawakid, A.; del Jesus, M.; Benítez, J.M.; Herrera, F. Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks. *WIREs Data Min. Knowl. Discov.* **2014**, *4*, 380–409. [CrossRef]
22. Mitchell, T.M. The Discipline of Machine Learning. Available online: http://www.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf (accessed on 29 May 2018).
23. Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]

24. LeCun, Y.; Benigo, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
25. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning (Adaptive Computation and Machine Learning)*; MIT Press: Cambridge, MA, USA, 2016; ISBN 978-0262035613. Available online: http://www.deeplearningbook.org/ (accessed on 29 May 2018).
26. Google's AlphaGo Defeats Chinese Go Master in Win for A.I. *New York Times*. 23 May 2017, Available online: https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html (accessed on 29 May 2018).
27. Nielsen, M.A. Chapter 5: Why are deep neural networks hard to train. In *Neural Networks and Deep Learning*; Determination Press: New York, NY, USA, 2015. Available online: http://neuralnetworksanddeeplearning.com/chap5.html (accessed on 9 July 2018).
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. CVPR 2015. Available online: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf (accessed on 29 May 2018).
29. Cat Poster composed by Wikipedia User Alvesgaspar. Available online: https://en.wikipedia.org/wiki/File:Cat_poster_1.jpg (accessed on 17 July 2018).
30. De Spiegeleire, S.; Maas, M.; Sweijs, T. Artificial Intelligence and the Future of Defense–Strategic Implications for Small and Medium-Sized Force Providers. Research Report, The Hague Centre for Strategic Studies. 2017. Available online: https://hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf (accessed on 29 May 2018).
31. Minsky, M. *Semantic Information Processing*; MIT Press: Cambridge, MA, USA, 1968.
32. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef] [PubMed]
33. Self-Driving Tesla Was Involved in Fatal Crash, U.S. Says. *New York Times*, 30 June 2016. Available online: https://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html (accessed on 29 May 2018).
34. Tesla Says Crashed Vehicle Had Been on Autopilot Before Fatal Accident. *New York Times*, 31 March 2018. Available online: https://www.nytimes.com/2018/03/31/business/tesla-crash-autopilot.html (accessed on 29 May 2018).
35. Video Released of Uber Self-Driving Crash That Killed Woman in Arizona. *The Guardian*, 22 March 2018. Available online: https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona (accessed on 29 May 2018).
36. Marcus, G. Deep Learning: A critical appraisal. *arXiv* 2018. Available online: https://arxiv.org/abs/1801.00631 (accessed on 29 May 2018).
37. Freud, S. Band 1: Elemente der Psychoanalyse. Band 2: Anwendungen der Psychoanalyse. In *Sigmund Freud: Werkausgabe in Zwei Bänden*; Freud, A., Grubrich-Simitis, I., Eds.; Fischer Verlag: Frankfurt, Germany, 2006.
38. Turing, A. Computing Machinery and Intelligence. *Mind* **1950**, *LIX*, 433–466. [CrossRef]
39. Weizenbaum, J. ELIZA—A Computer Program For the Study of Natural Language Communication between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
40. Hofstadter, D.R. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*; Basic Books: New York, NY, USA, 1996.
41. Trappl, R.; Petta, P.; Payr, S. *Emotions in Humans and Artifacts*; MIT Press: Cambridge, MA, USA, 2002.
42. Turing Test Success Marks Milestone in Computing History. University of Reading Press Release. 8 June 2014. Available online: http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx (accessed on 29 May 2018).
43. Sample, I.; Hern, A. Scientists Dispute Whether Computer 'Eugene Goostman' Passed Turing Test. *The Guardian*, 9 June 2014. Available online: http://www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed (accessed on 29 May 2018).
44. Eugene the Turing Test-Beating 'Human Computer'–in 'His' Own Words. *The Guardian*, 2014. Available online: https://www.theguardian.com/technology/2014/jun/09/eugene-person-human-computer-robot-chat-turing-test (accessed on 29 May 2018).

45. Aron, J. Software Tricks People into Thinking It Is Human. *New Scientist*. 6 September 2011. Available online: https://www.newscientist.com/article/dn20865-software-tricks-people-into-thinking-it-is-human/ (accessed on 29 May 2018).
46. Hunt, E. Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter. *The Guardian*, 24 March 2016. Available online: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed on 29 May 2018).
47. Microsoft 'Deeply Sorry' for Racist and Sexist Tweets by AI Chatbot. *The Guardian*, 26 March 2016. Available online: https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot (accessed on 29 May 2018).
48. The Art of Conversational Commerce. Gartner Group. 25 July 2016. Available online: https://www.gartner.com/smarterwithgartner/the-art-of-conversational-commerce/ (accessed on 29 May 2018).
49. Carroll, R. Goodbye Privacy, Hello 'Alexa': Amazon Echo, the Home Robot Who Hears It All. *The Guardian*, 21 November 2015. Available online: https://www.theguardian.com/technology/2015/nov/21/amazon-echo-alexa-home-robot-privacy-cloud (accessed on 29 May 2018).
50. Morley, K. Amazon Echo Rogue Payment Warning after TV Show Causes 'Alexa' to Order Dolls Houses. *The Telegraph*, 8 January 2017. Available online: http://www.telegraph.co.uk/news/2017/01/08/amazon-echo-rogue-payment-warning-tv-show-causes-alexa-order/ (accessed on 29 May 2018).
51. Sauer, G. A Murder Case Tests Alexa's Devotion to Your Privacy. *Wired*, 28 February 2017. Available online: https://www.wired.com/2017/02/murder-case-tests-alexas-devotion-privacy/ (accessed on 29 May 2018).
52. Bessi, A.; Ferrara, E. Social bots disturb the 2016 US presidential election online discussion. *First Monday* **2016**, *21*. [CrossRef]
53. Christian, J. Experts Fear Face Swapping Tech Could Start an International Showdown. *The Outline*, 1 February 2018. Available online: https://theoutline.com/post/3179/deepfake-videos-are-freaking-experts-out (accessed on 29 May 2018).
54. Dzindolet, M.T.; Peterson, S.A.; Pomranky, R.A.; Pierce, L.G.; Beck, H.P. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* **2003**, *58*, 697–718. [CrossRef]
55. Mosier, K.L.; Skitka, L.J.; Heers, S.; Burdick, M. Automation bias: Decision making and performance in high-tech cockpits. *Int. J. Aviat. Psychol.* **1997**, *8*, 47–63. Available online: https://www.ncbi.nlm.nih.gov/pubmed/11540946 (accessed on 29 May 2018). [CrossRef] [PubMed]
56. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [CrossRef] [PubMed]
57. Lieberman, H. Usable AI requires Commonsense knowledge. In Proceedings of the Workshop on Usable Artificial Intelligence, ACM Conference on Computers and Human Interaction (CHI-08), Florence, Italy, 5–10 April 2008.
58. Nothdurft, F.; Behnke, G.; Bercher, P.; Biundo, S.; Minker, W. The Interplay of User-Centered Dialog Systems and AI Planning. In Proceedings of the SIGDIAL 2015 Conference, Prague, Czech Republic, 2–4 September 2015; pp. 344–353.
59. Noessel, C. Designing Agentive Technology: AI That Works for People. *UX Magazine*, 2017. Available online: https://uxmag.com/articles/designing-agentive-technology (accessed on 14 June 2018).
60. Helms, K.; Brown, B.; Sahlgren, M.; Lampinen, A. Design Methods to Investigate User Experiences of Artificial Intelligence. In Proceedings of the AAAI 2018 Spring Symposium on Designing the User Experience of Artificial Intelligence, Stanford, CA, USA, 26–28 March 2018; pp. 394–398.
61. Buchanan, R. Wicked Problems in Design Thinking. *Des. Issues* **1992**, *8*, 5–21. [CrossRef]
62. Google reCAPTCHA. Available online: https://www.google.com/recaptcha/intro/index.html (accessed on 29 May 2018).
63. Annany, M.; Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* **2016**. [CrossRef]
64. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Crown: New York, NY, USA, 2016.
65. WWWF—World Wide Web Foundation: Algorithmic Accountablity–Applying the Concept to Different Country Contexts. 2017. Available online: https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf (accessed on 29 May 2018).

66. Caplan, R.; Donovan, J.; Hanson, L.; Matthews, J. Algorithmic Accountability: A Primer. Research Report, Data & Society 2018. Available online: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf (accessed on 29 May 2018).
67. Knight, W. The Dark Secret at the Heart of AI. *MIT Technology Review*, 2017. Available online: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (accessed on 29 May 2018).
68. Kant, I. *Grundlegung zur Metaphysik der Sitten*; Weischedel, W., Ed.; Werkausgabe Band VIII, First Published 1785; Suhrkamp Taschenbuch Wissenschaft: Berlin, Germany, 1997.
69. Kirchner, L. When Discrimination Is Baked into Algorithms. *The Atlantic*, 6 September 2015. Available online: https://www.theatlantic.com/business/archive/2015/09/discrimination-algorithms-disparate-impact/403969/ (accessed on 29 May 2018).
70. Johnson, I.; McMahon, C.; Schöning, J.; Hecht, B. The Effect of Population and "Structural" Biases on Social Media-based Algorithms—A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1167–1178.
71. Crawford, K. Artificial Intelligence's White Guy Problem. *New York Times*, 20 June 2016. Available online: www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (accessed on 29 May 2018).
72. Chander, A. The Racist Algorithm? Legal Studies Research Paper No. 498. 2016. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795203 (accessed on 29 May 2018).
73. Lambrecht, A.; Tucker, C.E. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electron. J.* **2018**. [CrossRef]
74. Kleinman, Z.; Artificial Intelligence: How to Avoid Racist Algorithms. *BBC Online*, 14 April 2017. Available online: http://www.bbc.com/news/technology-39533308 (accessed on 29 May 2018).
75. Townley, C.; Morrison, E.; Yeung, K. Big Data and Personalised Price Discrimination in EU Competition Law. Available online: https://dx.doi.org/10.2139/ssrn.3048688 (accessed on 29 May 2018).
76. Ddiscussion About Ethical Issues of a Russian Facial Recognition Company on the Twitter Account of Morten Rand-Hendriksen. Available online: https://twitter.com/mor10/status/995344394428473345 (accessed on 29 May 2018).
77. New Russian Facial Recognition Tech May Fight Crime but Its Already Misused against Opposition. *The Interpreter*, October 2017. Available online: http://www.interpretermag.com/new-russian-facial-recognition-tech-may-fight-crime-but-its-already-misused-against-opposition/ (accessed on 29 May 2018).
78. Burgess, M. Facial Recognition Tech Used by UK Police Is Making a Ton of Mistakes. *Wired*, 4 May 2018. Available online: https://www.wired.co.uk/article/face-recognition-police-uk-south-wales-met-notting-hill-carnival (accessed on 29 May 2018).
79. Top Scientists Call for Caution over Artificial Intelligence. *The Telegraph*, January 2015. Available online: http://www.telegraph.co.uk/technology/news/11342200/Top-scientists-call-for-caution-over-artificial-intelligence.html (accessed on 29 May 2018).
80. Johnson, D. Computer entities but not moral agents. *Ethics Inform. Technol.* **2006**, *8*, 195–204. [CrossRef]
81. Gunkel, D.J. *The Machine Question–Critical Perspectives on AI, Robots, and Ethics*; MIT Press: Cambridge, MA, USA; London, UK, 2012.
82. United Nations. Background on Lethal Autonomous Weapons Systems in the CCW. Available online: https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument (accessed on 29 May 2018).