# On the Media-political Dimension of Artificial Intelligence

## Deep Learning as a Black Box and OpenAI

*Andreas Sudmann*

**Abstract**

*The essay critically investigates the media-political dimension of modern AI technology. Rather than examining the political aspects of certain AI-driven applications, the main focus of the paper is centred around the political implications of AI's technological infrastructure, especially with regard to the machine learning approach that since around 2006 has been called Deep Learning (also known as the simulation of Artificial Neural Networks). Firstly, the paper discusses in how far Deep Learning is a fundamentally opaque black box technology, only partially accessible to human understanding. Secondly, and in relation to the first question, the essay takes a critical look at the agenda and activities of the research company OpenAI that supposedly aims to promote the democratization of AI and tries to make technologies like Deep Learning more accessible and transparent.*

> Neither machines nor programs are black boxes; they are artifacts that have been designed, both hardware and software, and we can open them up and look inside.
>
> (Allen Newell/Herbert A. Simon 1997 [1976], 82)

What does it mean to critically explore the media-political dimension of modern Artificial Intelligence (AI) technology? Rather than examining the political aspects of specific AI-driven applications like image or speech recognition systems, the main focus of this essay is on the political implications of AI's technological infrastructure itself, especially with regard to the machine learning approach that since around 2006 has been called Deep Learning (in short: DL, also known as the simulation of neural networks or Artificial Neural Networks – ANN). Firstly, this essay discusses in how far ANN/DL have to be perceived as a fundamentally opaque black box technology, perhaps not or only partially accessible to human understanding. Secondly, and in relation to the first question, the aim is to take a critical look at the agenda and activities of a research company called OpenAI that

DOI 10.14361/dcs-2018-0111

supposedly promotes the democratization of AI and tries to make technologies like DL more accessible and transparent. Obviously, such idealistic claims should not simply be taken for granted, especially if one takes into account the large amount of money involved in a company like OpenAI. For example, strategies like open-sourcing AI seem more likely to serve the purpose of demonstrating those companies' technological potential, to one-up each other, and/or to attract rare talents. But perhaps even more important than simply questioning the authenticity or ideological implications of such claims, we have to address more fundamental problems here: How can one contribute to the transparency and accessibility of a black box that – perhaps – cannot be opened at all? And can there be a democratization of AI without a democratization of data in general?

## The so-called "AI revolution"

Before addressing these questions, it is important to recapitulate how DL recently managed to become the dominant paradigm of AI. An event of major importance in this respect happened in 2012. Back then, three scholars from the University of Toronto, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, for the first time effectively trained a so-called Convolutional Neural Network, which is a special variant of a traditional Artificial Neural Network (ANN) optimized for image and object recognition, on the basis of the now famous database ImageNet as well as on fast parallel-organized GPU processors (cf. Huang 2016). Especially the substantial employment of GPUs made all the difference: The Toronto team was able to reduce the error rate of previous approaches in image recognition about more than the half.[1] Prima facie, this increase may not sound very impressive, yet it was big enough to attract the attention of leading IT companies like Google and Facebook, which quickly hired leading scientists like Yann LeCun and Geoffrey Hinton and also acquired AI start-up companies such as DNNResearch and DeepMind. The strong interest of these companies in DL technology was no surprise since all of them have already been harnessing big data, and now they had access to a powerful technology to process and harness it intelligently (cf. Reichert 2014). For example, thanks to DL it is possible to automatically tag images uploaded by users on social media-platforms, or to analyse consumer behaviour to generate individualized ads, or make personalized recommendations. Of course, there are many other application areas for which DL/ANN technology is currently used: e. g. to process the sensory data of self-driving cars, to analyse data for stock market predictions, or for optimized machine translations, etc. In general, DL algorithms are an universal instrument for pattern-recognition and prediction tasks, an

1 Over a period of just seven years, the accuracy in classifying objects in the dataset rose from 71.8 % to 97.3 %. Not least due to his high value, 2017 was the last year of this famous competition. Cf. Gershgorn (2017) for the details.

effective tool to manage uncertainty and fuzziness in data content (Goodfellow/Bengio/Courville 2016).

However, it took a while until modern machine learning algorithms were able to unfold their potential. Some of the technological essentials of DL/ANN were already developed in the 1940s and 1950s (cf. Sudmann 2016). Already back then, the basic idea of this AI paradigm was to develop a computer system that should be able to learn by observation and experience to solve specific problems or fulfil certain learning tasks, without having concrete rules or theories guiding the process (Mitchell 1997). This is basically the approach of every existing machine learning system, as opposed to so-called symbolic, rule-based forms of AI systems whose intelligent behaviour typically is more or less hand-coded in advance (Boden 2014). Even though there are many ML approaches out there, it recently turned out that DL methods are the most effective ones, at least with regard to key areas of AI research like voice or image recognition.

Very broadly speaking, one key characteristics of DL is that it is a class of techniques that are loosely inspired by the structure and learning processes of biological neural networks (Alpaydin 2016). As with other machine learning tasks, DL algorithms learn by analysing thousands, if not millions of training data on the basis of thousands or even million iterations up until the very moment the system is also able to predict unseen data correctly. Yet what distinguishes DL from other machine learning approaches is the *hierarchical distribution* of its learning process. DL technology simulates networks typically consisting of millions of artificial neurons that are organized on different layers – an input, an output and a flexible number of intermediate hidden layers (Trask et al. 2015). If a network is called deep, it has at least more than one intermediate layer that processes the information through the network. On the lowest level of layers, the network just analyzes very simple forms of input (for example, lines and edges, in case of visual data) and forwards this information to the next level of layers, which processes more complicated forms (parts of the object like face or legs) and again forwards this information to the next highest level, all the way up to the final layer, the output layer, which than can predict if a certain unknown input correctly matches with a certain output (does the image show a certain object or not?).

## The Media-politics of Deep Learning

It is also not very surprising that the current AI boom quickly started to attract the attention of the humanities and cultural sciences, whereas before around 2016 many disciplines outside the natural sciences more or less ignored machine learning technologies or DL/ANN. Of course, there has been a long tradition of an inter- and transdisciplinary debate concerning the potentials and limits of AI (cf. Weizenbaum 1976, Searle 1980), yet those discussions typically did not address the technology of ANN in any great detail. There are some important

exceptions to highlight in this respect, especially the philosophy of mind as well as cognitive psychology, which very early developed an interest for both the symbolic and connectionist forms of AI (cf. Horgan/Tienson 1996, Haugeland 1997). Furthermore, even in the field of media studies, one can find a few cases where scholars have been discussing ANN technology.[2] One example is the introduction of *Computer als Medium* (1994), an anthology co-edited by Friedrich Kittler, Georg Christoph Tholen, and Norbert Bolz. Interestingly enough, the text (written by Bolz only) somehow captures the epistemic relevance of the connectionist paradigm of AI, yet without exploring the details of its media-theoretical or -historical implications.

In the last two years or so (more or less after the success of DeepMind's AlphaGo), the situation has changed significantly. More and more books and articles are published in the area of social and cultural studies that tackle the topic of AI in general and DL technology in particular (Sudmann 2016, Pasquinelli 2017, Finn 2017, McKenzie 2017, Engemann/Sudmann 2018, and of course also this special-issue). For example, Pasquinelli (2017) recently wrote a short essay on ANN from a historical and philosophical perspective, arguing (with reference to Eco and Peirce) that the technology can only manage inductive reasoning, whereas it is incapable to enable forms of what Peirce calls abductive reasoning. Furthermore, there are authors like Nick Bostrom (2014), Ed Finn (2017), or Luciano Floridi (2017) who are already very much engaged in the political and ethical discussion of current AI technologies. For example, Nick Bostrom's recent book (2014) attracted much public attention, partly because of its alarmistic thesis that the technological development of a super machine intelligence is mankind's greatest threat, which was later echoed by a twitter post from Elon Musk. Yet, not everyone concerned with the political and ethical aspects of AI shares these apocalyptical views. Luciano Floridi, for instance, is convinced that mankind is able to handle a AI-driven society as long as society instantiates a "system of design, control, transparency and accountability overseen by humans" (2017: online).

Yet, what is still widely missing in the intellectual debate is a discussion of AI/DL from a decidedly media-political perspective. But what does such a focus involve, and why do we need it in the first place? To begin with, there are – of course – many different ways to think about the media-political dimension of AI in general and DL in particular. For example, one possible approach would be to claim that "media politics" as an analytical agenda is concerned with the mediation of politics and/or the historical relationship of media and politics (cf. Dahlberg/Phelan 2011). Based on such an account, one could ask, for instance, how AI/DL technology inscribes itself in relations of media and politics, or how it participates in the mediation of politics. In both cases, we might assume that a) media/mediation and politics are basically distinct concepts, and that b) possible

2 Of course, there are some more publications from a media studies perspective that deal with AI technology in general, for example: Dotzler 2006.

analytical perspectives are very much shaped and guided by our basic understanding of these terms in the first place (including to perceive AI technology itself as a medium).

Yet another possible approach would be to claim that media have an inherently political dimension (and, similarly, one could claim that nothing political exists outside a medium or certain media). Still, the question remains if this is true for *every* concept of media or medium or just particular ones. But this is a rather theoretical discussion, since most concepts of media politics are more or less based on a traditional understanding of media as mass or popular media (cf. Zaller 1999, Dahlberg/Phelan 2011). Accordingly, one possible way of approaching the media politics of AI and DL would be to examine the politics of representation of different AI technologies in popular media like film and television.

In the context of this essay, my understanding of a media-political account, however, is a broader and in a certain light more basic one. Such a theoretical perspective, as I like to conceptualize it, is not so much concerned with the representations or visible interfaces of AI, but more interested in the political implications and effects of the medial infrastructure and entities that generate and shape the technology (also regardless of particular "use cases").[3] In other words, what I am interested in are – what I like to call – the infra-medial conditions of modern AI technology and their political dimension.[4] For me, *every* entity involved in the process of generating and shaping AI technology can generally be perceived as a mediator of this technology (cf. Latour 2005). And generally, every mediator of technology also matters in political terms. However, not every mediator can be equally conceptualized as a medium, at least not if one applies a more narrow understanding of the term, for example, to regard media as entities or dispositifs that enable communication or that store, process, or transmit information.[5] For this very reason, it generally makes sense to differentiate between the concepts of mediator(s) and medium/media.

3 For a similar account, using the term "media infrastructures" as a critical concept, cf. Parks/Starosielski (2015).

4 Such a perspective is not directly concerned with a specific theoretical framework. Generally, this focus is compatible with many analytical approaches like media archaeology, historical epistemology, or actor-network theory.

5 This is a different account of how media can be conceptualized with reference to Latour's differentiation between "mediators" and "intermediaries". For Latour, an "intermediary [...] is what transports meaning or force without transformations" opposed to "mediators" that "transform, translate, distort, and modify the meaning or the elements they are opposed to carry" (2005: 39). Intermediaries function in a certain sense as black boxes, since their input allows you to predict the respective output (without having knowledge of the object's internal operations). In opposition to that, in case of mediators, despite the specific nature of a certain input, it is never easy to predict the respective output (ibid., cf. also Thielmann 2013).

Yet while I argue that we need such a distinction, I am nevertheless quite sceptical about using a stabile concept of the term "medium" or "media" (even though it would make the task of differentiating both terms much easier). In my mind, in order to make sense of our empirical world's entities (including the immaterial world of our thoughts), the terms media and/or medium are more productive in analytical terms if one regards them as flexible epistemological-heuristic rather than fixed categories.[6] Accordingly, media theory, as I advocate it, can be understood as the general task to explore in what different ways the world is perceivable as a medium or as media (with certain characteristics, functions, inscriptions) rather than simply acknowledging that everything out there in the world depends on media/a medium in some ontological stabile sense (as a precondition of entities to be visible, to be perceivable, or to have a certain form etc.). Hence, even though I opt for a concept of media politics that is focused on the constitutive role of mediators (and – more specifically – as specific media), I still advocate a rather open analytical focus that leaves room for very different perspectives.

The latter position seems also to be an instructive approach with regard to the political dimension of media politics. For example, we can quite easily claim that almost everything about AI is political, especially if one believes that AI/DL technology affects every aspect of our social and cultural existence. At the same time, the political challenges that AI and DL technology hold for us are very different in nature (the existential threat of AI-driven weapon systems, AI's influence on the future of work, etc.), which is why we cannot simply refer to a master account of political theory suitable for each and every problem.

Such a plea does not basically mean "anything goes" in terms of how we should address the politics of AI/DL. Instead, I argue that one should – first of all – try to explore how contemporary AI technologies emerge as political phenomena (before we apply a certain political theory on AI). This focus entails many relevant aspects, including the analysis of the ways in which computer scientists themselves conceptualize AI technology as a political subject.

In this context, one should also keep mind that the subjects of machine learning in general and ANN/DL in particular are, again, still an unknown territory for most scholars working in the humanities, social, or cultural sciences, even if they have already studied AI. What this basically means is that it might take a while until disciplines like media or cultural studies are really able to evaluate the relevant political and/or ethical aspects of DL's technologies and infrastructures. Obviously, this problem is also a central factor in discussing AI/DL as a black box and in evaluating projects like OpenAI. For many scholars in the field, it is one of media studies' central tasks to focus on processes of knowledge transla-

6 There are perhaps many approaches to justify such a concept of media-thinking. Obviously, we can again refer to Latour's category of "mediators". A similar theoretical reference in this context is also Vogl's term of "becoming-media" (2008).

tion or transformation and to analyse, from a kind of meta-perspective, how the knowledge of one discipline is used, adapted, and reconfigured by another discipline (cf. for example Bergermann/Hanke 2017: 127). But how can media studies provide relevant insights into the black box problem of AI/DL if even computer or data scientists have profound trouble dealing with it? Obviously, media studies has nothing or little to contribute to open this black box in technical terms, yet it can perhaps shed light on different aspects: for example, exploring the problem's complex network of socio-cultural conditions, implications, and effects. Furthermore, media studies can – of course – critically investigate how data scientists treat AI and its black box problem as a political concern. But in order to do so, let's recapitulate what it means – or better – what it could mean to perceive certain entities of our empirical worlds as black boxes.

## Deep Learning: A Black Box that Cannot be Opened?

There are some debates going on about the exact origins of the term black box. Philipp von Hilgers has explored the history of the concept and traced it back to history of World War II, more precisely to the technology of the magnetron (Hilgers 2009). Since then, the concept has been applied and specified in very different contexts with opposed meanings. On the one hand, it can refer to the data-monitoring systems in planes or cars; on the other hand, it encompasses systems whose inner operations are opaque or inaccessible and thus only observable by their inputs and outputs (cf. Pasquale 2015: 3). One early definition of the term black box has been provided by Norbert Wiener, in a footnote of the preface added to the 1961 edition of his famous book *Cybernetics* (1948): "I shall understand by a black box a piece of apparatus [...] which performs a definite operation [...] but for which we do no necessarily have any information of the structure by which this operation is performed" (p. xi). Last but not least, as Latour explains, one has to consider that the operations of science and technology always have a black boxing effect: "When a machine runs efficiently, when a matter of fact is settled, one need focus only on its inputs and outputs and not on its internal complexity. Thus, paradoxically, the more science and technology succeed, the more opaque and obscure they become" (Latour 1999: 99). Prima facie, this also seem to be true for DL. And yet, as opposed to other forms of technology, the case of DL technology seems to be different.

Typically, independent of the black boxing effect just mentioned, many, if not most operations of technology used in practice are in one way or the other accessible to human understanding and explanation. In contrast, DL algorithms seem to be a black box that cannot be opened. At least this is what several experts currently identify as one of AI's biggest problems. But is it actually true that DL is a fundamental opaque technology and if so, to what degree? And even if this is the case, can't we simply accept ANN to be an opaque technology as long as it works

smoothly? The latter question may appear less complicated to answer than the first one. In fact, there already seems to be a large consensus among many scientists, politicians, and leading IT companies to develop a responsible or ethical AI, and making the technology more accountable is one essential part of this endeavor.

This broad consensus is, of course, no surprise. It's one thing if an AI/DL system comes up with a disappointing movie recommendation, but if we use intelligent machines for more serious matters concerning questions of life or death, the story is a completely different one. As Tommi Jaakkola, computer scientist at MIT, recently pointed out: "Whether it's an investment decision, a medical decision, or maybe a military decision, you don't want to just rely on a 'black box' method" (Knight 2017).

For this reason, it might not be enough knowing that the predictions of your AI/DL system are sufficiently accurate. Furthermore, you want to understand why the system comes up with a certain prediction. Both aspects seem highly relevant to secure trust in an AI-driven decision. Yet, to grasp the meaning of AI's prediction models seems to be rather challenging. To illustrate this last point: Recently, researchers at the Icahn School of Medicine at Mount Sinai developed an AI program called "Deep Patient". The system is able to identify different forms of diseases and even early indications of psychiatric disorders like schizophrenia astonishingly well, yet they still do not have a clue how this is possible. Of course, Deep Patient can be of great help for doctors, but they need the system to provide a rationale for its predictions, so that they have a solid reference for the medical treatment of their patients. "We can build these models," as Joel Dudley, the director of biomedical informatics at the Icahn School of Medicine, explains, "but we don't know how they work" (Knight 2017).

In the following, I discuss in how far this assumption, which regularly appears in current AI discourses, is somehow misleading and needs some clarifications. Firstly, one should keep in mind that the math behind current DL technology is pretty much straight forward (cf. Goodfellow/Bengio/Courville 2016). Ultimately, it is a matter of statistics, albeit an advanced form of it. This aspect is important to highlight since we can observe a general tendency of mystifying DL that is counterproductive and needs to be contained. Secondly, many experts stress that ANN *are in fact* an accessible technology, especially if one compares them with biological neural networks. For example, Roland Memisevic, chief scientist of the Toronto-Berlin-based DL company TwentyBN, points out that "DL algorithms are at least way more accessible than the human brain, where the neuronal activity patterns as well as the transformations effected by learning are, even today, still very much opaque. In contrast, if one looks at an ANN model, you can record, observe, measure everything, down to the smallest detail. For example, it is easy to find out which features have falsely resulted in a dog being labelled as a cat, because certain ear shapes might again and again lead to certain misclassifications" (Memisevic 2018, *my own translation*). However, what indeed is difficult to understand is the interplay of the artificial neurons, as Memisevic agrees, "since

having such a great number of neurons that are active in parallel, one is confronted with emergent phenomena, whereby the whole encompasses more than the sum of its parts" (ibid.).

Thus, while it is certainly true that computer scientists have to deal with what is commonly labelled the interpretability problem of DL, it is not as fundamental as it is often described in the current discourse (cf. Knight 2017). And, not surprisingly, computer scientists inside and outside the tech industry are currently very busy to come to terms with this interpretability problem. In fact, researchers have already developed a number of approaches to better understand and reverse-engineer DL's prediction models.

## Strategies of Explainable AI (XAI)

One example to make not only ANN but machine learning technologies in general more accessible is the program Local Interpretable Model-Agnostic Explanations (LIME), developed by Marco Tulio Riberio, Sameer Singh, and Carlos Guestrin. The authors describe it as "a technique to explain the predictions of *any* machine learning classifier, and evaluate its usefulness in various tasks related to trust" (Riberio/Singh/Guestrin 2016). The basic idea behind LIME is to change different forms of inputs (e. g. texts or images) to the AI system in such a way that one can observe if and how these variations of the input have an impact on the output. A recent article in the journal *Science* explains how LIME works in practice, with reference to an ANN that is fed with movie reviews:

> [A neural network] ingests the words of movie reviews and flags those that are positive. Ribeiro's program, called Local Interpretable Model-Agnostic Explanations (LIME), would take a review flagged as positive and create subtle variations by deleting or replacing words. Those variants would then be run through the black box to see whether it still considered them to be positive. On the basis of thousands of tests, LIME can identify the words – or parts of an image or molecular structure, or any other kind of data – most important in the AI's original judgment. The tests might reveal that the word 'horrible' was vital to a panning or that 'Daniel Day Lewis' led to a positive review. (Voosen 2017)

As one already can deduct from this short description, it seems to be an exaggeration to claim that this model indeed provides an explanation in any profound sense. Basically, it is a 'experimental system' that simply highlights those elements that play an important role in the system's decision-making process, without actually revealing the reasoning implicit in this prediction model.

Another interesting tool that – in a way – helped to make visible how ANN work, is a now famous program called "DeepDream", introduced by engineers and scientists at Google in 2015. DeepDream is a special DL-based image recognition algorithm, yet it operates a little bit different from a typical CNN. First, the

algorithm is trained with millions of images that show a particular object (for example, a cat) so that, at some point, the NN is able to predict or classify objects in images as cat which it hasn't been trained for. After the initial training, the network can operate in reverse. Instead of adjusting the weights of a networks, as would be the standard procedure with the back prop algorithm, the weights remain unchanged, and only the input (the original image of a cat) is minimally adjusted. This technique has very interesting results if you apply it to images that do not contain any cats but are labelled as if they would. In this case, the software begins to modify and enhance certain patterns of images, so that they start to look more and more like a cat, yet not similar to any particular existing one in our empirical world, but like a cat the way a NN has learned to perceive, if not to say: dream it. As a result of this process, the system produces images that have a surreal and grotesque quality: for example, a photograph of a pizza can entail many little dog faces or you can also turn the Mona Lisa into a LSD-like hallucinatory nightmare.[7] The generated images reveal at least two interesting aspects: On the one hand, they show that DL is not an entirely mysterious technology in so far as the algorithm enhances familiar visual features. On the other hand, the images illustrate how differently the algorithm works in comparison to human perception, foregrounding, in other words, that it might focus on aspects of an image to which we usually, as humans, do not pay attention (cf. Knight 2017).

A third potential approach to expose the working mechanisms of a DL system is the so-called "Pointing and Justification (PJ-X)" model developed at the University of California, Berkeley, and the Max Planck Institute for Informatics (see Park et al. [2016]). The model is able to justify its prediction or classification tasks by highlighting and documenting the evidence for the algorithmic decision using an attention mechanism combined with a natural language explanation. A key element of the system is that it is trained with two different data sets. The first one is meant to determine what an image shows, while the second one has the function to reveal *why* something (i. e., a certain human activity or object) appears in a particular image. Thus, the idea is to correlate images that show objects or human activities not only with their description (by labelling them), but also with their respective explanation. For the latter purpose, each image of the training data is associated with three questions as well as with ten answers for each of them. On this basis, the system can answer questions like "Is the person in the image flying?" And the answer might be: "No, because the person's feet are still standing on the ground" (cf. Gershgorn 2016).

Again, this model – like all of the above – is still far away from being able to explain its own internal operations or those of different machine (or of another ANN, if you will). Perhaps, this specific capability would require that machines develop some kind of self-consciousness, or even a meta-consciousness. Before

7 For a critical perspective on DeepDream from a media-theoretical and psychoanalytical perspective, cf. Apprich (2017).

this happens (if this ever going to happen), DL technology need to understand reasoning, planning, causal relationships, and so on. For the moment, the technology of DL or ANN only provides correlations, but no profound causal explanations. In that regard, DL is still in its fledgling stage. Hence, one could argue that – in a certain sense – the label "Explainable AI" is misleading or perhaps at least an exaggeration.

## The Politics of OpenAI

As I indicated earlier, providing models of an explainable and – more generally – a responsible AI has some obvious motivations and reasons. First and foremost, those who currently develop DL systems have a strong economic interest to counter the social fears and scepticism related to a profoundly opaque AI technology. Nonetheless, many scientists and industrial actors underscore the political and ethical importance of developing an explainable AI beyond the commercial aspects connected to the interpretability problem described above. One of the most visible and powerful actors among those highlighting this agenda is OpenAI, a "non-profit research company" (self-description), also specialized on DL technology. Here is how the company outlined its mission goal, shortly after it has been founded in October 2015:

> Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact.
> We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as possible. The outcome of this venture is uncertain and the work is difficult, but we believe the goal and the structure are right. We hope this is what matters most to the best in the field. ("Introducing OpenAI")

To make sure that OpenAI is "unconstrained by a need to generate financial return," the founders of the company, among them, most prominently, Elon Musk and Sam Altman, invested more than US$1 billion in this venture. Interestingly enough, this initial launch posting does not explicitly or directly refer to what Elon Musk has named one of his key motivation for his initial investment in OpenAI, namely, that he regards (general) AI to be humanity's biggest existential threat.[8] This apocalyptic view has been around since the very beginning of AI research and even before. In fact, as media scholar Bernhard Dotzler already pointed at the end of the 1980s, you can find all well-established projections of the future of AI already in the work of Alan Turing (cf. Dotzler 1989). And yet, since very recently,

8 In February 2018, Musk announced that he is leaving the board of OpenAI due to a potential conflict of interest with his (future) work at Tesla (Vincent 2018).

the development of AI has given us little reason to expect any dystopian 'Terminator' reality to be just around the corner.

For the first time in the history of mankind, the current situation might indeed be a different one vis-à-vis the undeniable fast progress of current DL technology. At least this is what many experts beyond Musk believe to be the case. OpenAI's agenda acknowledges this new situation, but in a more nuanced, less dramatic manner:

> AI systems today have impressive but narrow capabilities. It seems that we'll keep whittling away at their constraints, and in the extreme case they will reach human performance on virtually every intellectual task. It's hard to fathom how much human-level AI could benefit society, and it's equally hard to imagine how much it could damage society if built or used incorrectly. ("About OpenAI")

Indeed, no one is able to foresee the future of AI or can evaluate whether it will more likely have a positive or negative effect on society and culture. We might also tell ourselves that technology is never inherently good or bad, hence what matters only is its specific use. This argument, however, has always been a rather problematic one, since, in fact, it makes a big difference if we deal with nuclear technology or, say, wind power. Furthermore, even though it is rather a truism that the future is uncertain, we should also not forget that we can never be sure at which concrete point we might take the wrong path towards it.

It is particularly this latter argument that seems to correspond with how OpenAI is linking its current agenda to the problem of an unforeseeable future:

> Because of AI's surprising history, it's hard to predict when human-level AI might come within reach. When it does, it'll be important to have a leading research institution which can prioritize a good outcome for all over its own self-interest. ("About OpenAI")

What is interesting about this passage is the implicit assumption that the whole question concerning the drastic negative or positive effects of AI is still a rather speculative matter and not so much one that concerns the current state of technology ("... when the human-level AI might come within reach"). While OpenAI is right about avoiding any speculative discussion, it seems important to realize that DL *already has* both positive and problematic implications. The technology can do many astonishing good things, as it *already has become* a very powerful and also dangerous surveillance technology that expands the possibilities not only to (semi-automatically) observe the world (after being trained to do so), but to be able to make sense of it.

Very recently, it turned out that ANN/DL are not only able to identify objects, people, landscapes, and animals (again, after being trained to do so), but that they have started to understand quite complex actions and gestures. In other words: DL systems have begun to understand a basic form of common-sense knowledge

of the world. The interesting aspect: In order to achieve this, the ANN has been trained with hundreds of thousands of short video-clips (showing certain activities and hand gestures). Hence, the specificity of media is essential for developing advanced forms of AI. At the time of this writing, this particular DL system has not yet been implemented in industrial applications. But the technology is out there and ready to be used (cf. Sudmann 2016).

Ed Finn has recently argued that today's algorithmic culture is more than ever driven by the "the desire to make the world effectively calculable" (2017: 26). Without specifically distinguishing them from learning algorithms, he regards algorithms in general as "cultural machines" (54) whose operations are very much determined by an "ideology of universal computation" (23). Indeed, one could argue that especially modern DL technology fuels a phantasmatic version of instrumental reason, precisely because it reawakens the old dream Leibnizian dream of a *mathesis universalis*, capable of perfectly understanding every aspect of our world. But even more: The great promise of DL is not only to make machines understand the world, but to make it predictable in ever so many ways: how the stock market develops, what people want to buy, if a person is going to die or not, and so on. Already at this particular moment in history, we can regard DL as the very technology that is capable with complexities humans aren't able to handle. The algorithmic power of DL lies in its potential to identify patterns by learning from the past to evaluate the present in order to master an uncertain future. And all of this happens in an ever faster way. DeepMind just presented a new version of its Go-program "AlphaGo Zero" that was able to learn the ancient board game in only three days from scratch (without implementing any rules how the game works or how it might be played successfully) and managed to win against the older system of 2015/16 (that beat the human world champion Lee Sedol) with 100 to 0 (Perez 2017).

The rapid speed of innovations in the field of DL should also remind us to be careful about quickly jumping at conclusions about what AI technology is or is not able to achieve. Hence, we should not only stop speculating about a distant future of AI, but we should also be careful about our sceptic views on what AI systems are capable of (or not). In general, we should acknowledge there is still a lot of work for us to do if we are trying to come to terms with the current development of AI and machine learning technology. Maybe companies like OpenAI succeed in making AI technology more accessible. But how exactly do they justify their central claim of democratizing AI? If we take another look at the company's official website, we will realize that it provides very little information: "We publish at top machine learning conferences, open-source software tools for accelerating AI research, and release blog posts to communicate our research" ("About OpenAI"). This is basically all the company has to say about its agenda of democratization AI, at least if we just consider the website's official mission statement. One thing that is very remarkable about this passage is the fact that there is nothing special about it. Facebook, Microsoft, and many other IT companies basically have the same agenda (cf. Boyd 2017).

Of course, one could argue that OpenAI at least started the current wave of developing a responsible and safe AI. The more important point, however, is: How can OpenAI legitimate its special status as a non-profit research company when it essentially does what all other big players in the AI game are doing: improving existing technology and/or finding the right path to develop an AGI – artificial general intelligence? Concerning this matter, and very similar to the situation at DeepMind, OpenAI's research is focused on strategies of reinforcement learning in connection with simulations (like games) instead of using the common approach of supervised learning that depends on correctly labelled data from the empirical world (cf. Rodriguez 2017). Within the specific limits of their approach, both OpenAI's and DeepMind's agenda have been quite successful. Yet, as of now, simulations are still not a suitable substitute for empirical learning data. If this turns out to be a permanent problem, it will have tremendous implications for how we conceive the epistemological status of simulations (in many theories and histories of digital and visual culture), but this remains to be seen. The reason why I have highlighted this point is a different one: As we just saw, there are many facets to the black box problem of DL. It is not my aim to get into every detail of how leading IT companies currently try to develop highly efficient AGI system at some point. Instead, what we can learn by taking a closer look at those different research agendas is the simple fact that DL is not a homogenous approach, but an umbrella term for very different accounts.

Furthermore, referring to the heterogeneity of DL is not only important in terms of how we address the black box problem of AI, but also for how we can develop a critical perspective on intelligent machines. To provide just one example: A few years ago, Alexander Galloway wrote a very interesting article in which he somehow politicized the black box by arguing that it is no longer a cipher like the magnetron technology during the Second World War, but instead has become a function that is more or less completely defined by its inputs and outputs (cf. Galloway 2011: 273). By using the term, he does not exclusively mean technical devices but refers to all networks and infrastructures of humans, objects, etc. that may interact with each other, yet thereby only articulating their external grammar. Obviously, Galloway's concept of the black box shares some similarities with how the term is used in the actor-network theory, though with an important difference: According to Galloway, the elements of a network that constitute a black box are no longer able to reveal anything about themselves. In other words: He believes that those networks have become a black box that *cannot be* opened (this is also how Hilgers defines a black box – as system whose inner processes remain constantly inaccessible; cf. Hilgers 2009). Opposed to that, for example, Michel Callon has argued that any black box whose actor-network operations do not adequately model the working of a system not only can be, but must be cracked open, thereby producing a "swarm of new actors" (Callon 1986). At first glance, it seems that that Galloway's concept of black box could be useful to describe the infrastructures and technological networks mediated by modern DL/ANN algo-

rithms. But this is not as easy as it might seem in the first place. Galloway's model is based on the existence of given inputs and outputs. Yet, ANN technology does not always operate with both inputs and outputs available. For example, in case of what is called unsupervised machine learning, the algorithm is trained without given outputs. Hence, as this simple example shows, if we want to understand the nuances of a DL/ML infrastructure as a black box, Galloway's intervention might be of limited use. At the same time – and this is the aspect where the actor-network theory comes into play again – we cannot simply assume that the black box problem as a political (or ethical) issue only concerns the algorithm itself. Instead, the question encompasses many different things: legal aspects, institutional procedures, environmental issues, existing political as well as legal regulations, and so on.

These aspects are also important to consider if we only think about how DL programs exhibit racial or gender biases. There was great turmoil when Microsoft's chatbot "Tay" was trained by Twitter users to learn racist, sexist, as well as anti-Semitic statements (Vincent 2016). This particular scandal is very insightful, since it demonstrates how much the operations of learning algorithms actually depend on the data and – even more importantly – on the people who label the data, at least in the case of supervised learning tasks. In other words: It is not the algorithms that produce prejudices or political problematic outcome, but in fact the human actors who design and generate the learning data, among them the hundreds or thousands of crowd-workers hired and organized through platforms like Amazon Mechanical Turk or CrowdFlower. Thus, if we want to talk about a bias problem of AI, we should also address the general structures of prejudices and ideology that still inform our society and thus the experts and workers who design the AI systems. Furthermore, this example clearly shows why it matters to take a closer look at the way certain forms of media act as key mediators of modern AI technology.

Without doubt, it is somehow short-sighted that the discussion on AI as a black box so far has focused almost exclusively on the technological aspects in the narrower sense. This also concerns the critique of a "democratic AI". For example, philosopher Nick Bostrom recently questioned the whole logic of making AI safer by making it more transparent: "If you have a button that could do bad things to the world, you don't want to give it to everyone" (quoted after: Metz 2016). Prima facie, this argument may sound convincing, but at the same time, it seems a little bit odd. If we think about nuclear weapons, for example, one can easily observe how complicated it is to keep a possibly "dangerous button" just for yourself. (We might also point to recent discussions here about the US president's right to decide if he uses nuclear weapons as a first strike or not). I do not want to argue that the concept of a balance of deterrence during the Cold War actually had a peace-securing effect, nor do I want to put the specific technology of nuclear weapons on the same level as AI. I just want to illustrate why the whole practice and discourse of a responsible or transparent AI maybe more complicated than

Bostrom's statement suggests. Neither it is true that the only alternative to the idea of a transparent AI would be to keep all the relevant knowledge about AI secret. At least, the latter strategy cannot be an option for OpenAI, since it would destroy the company's very identity.

Furthermore, it is important to highlight that as much as the black box problem does not only concern the technology itself, we also have to acknowledge that any attempt to democratize AI cannot just be reduced to activities of open-sourcing the tools and knowledge around its technology (cf. for a further critical view on AI as a black box beyond issues of transparency and accountability, see Matzner 2017). It is not a dystopian position to argue that we already live in a post-privacy age where people have very little control about the processes of data collection, storage, processing, and transmission related to their personal lives and activities. Especially the revelations of Edward Snowden confirmed the worst conspiracy theories about surveillance to be true (Sprenger 2015). The problem here is not only that companies or secret services, or governments in general, collect and analyse private data against our will. All too often, many people simply do not care enough about the data they generate and circulate while being online or using this or that application. And even if they individually try to protect their private information, there is no guarantee that their friends, family or colleagues will do the same. These aspects have been the subject of cultural critique long before the current AI boom took off. We should therefore not simply discuss how to democratize AI but continue our efforts to secure democratic standards for our data-driven world in general. To achieve this goal, linking the political analysis of AI with a broader discussion about datafication is nothing more than a first step, but arguably a very important one.

Currently, it is hard to think of any institution or law, globally or locally, that can prevent us from the dangers of AI as well as the misuse of big data. Neither do we have any profound reason to believe that companies like OpenAI, Facebook, or Google will achieve this goal. At the same time, it is perhaps short-sighted to think of these tech companies as the enemies of a democratic digital culture just because they are the hegemonic forces to control both the data as well as the intelligent algorithms to make sense of it. Obviously, there are dangers of AI that are more urgent, for example, if non-democratic states use AI or DL technology to oppress their political opposition or terrorists for their illegal activities. This threat is not a scenario of a big data paranoia: As experts have recently demonstrated, by only having access to a so-called API, you are able to reverse-engineer machine learning algorithms close to 100 % accuracy. Hackers are able to steal AI technology from IT companies like IBM or Microsoft for whatever their specific goals might be (for the technical details, see Claburn 2016). Of course, having a truly open AI might solve this particular problem in the first place. But then again, one has to ask how we can make sure that an open AI is not used for harmful purposes.

As of now, OpenAI seems less concerned with any concrete political vision of AI and more keen on participating in the competitive race towards developing an artificial general intelligence. Hence, it is quite seductive to believe OpenAI's political or ethical agenda is basically a PR stunt and nothing else. But instead of simply questioning if OpenAI's concrete practices matches their agenda or not, it might be more productive for a media-political account to discuss the political implications and effects of a transparent or responsible AI in the context of a broader focus: How the technology of learning algorithms reshapes the conditions of an instrumental rationality so deeply connected with every aspect of our digital culture and society. And this important project just has started.

## References

"About OpenAI." *OpenAI* Website. https://openai.com/about/#mission [Last access: 2018/03/06].

Alpaydin, Ethem (2016): *Machine Learning. The New AI*. Cambridge, MA: MIT P.

Apprich, Clemens (2017): "Daten, Wahn, Sinn." *Zeitschrift für Medienwissenschaft* 17, 54–62.

Bergermann, Ulrike and Christine Hanke (2017): "Boundary Objects, Boundary Media. Von Grenzobjekten und Medien bei Susan Leigh Star und James R. Griesemer." In: *Grenzobjekte und Medienforschung*. Eds. Sebastian Gießmann and Nadine Taha. Bielefeld: transcript, 117–130.

Boden, Margaret A. (2014): "GOFAI." In: *The Cambridge Handbook of Artificial Intelligence*. Eds. Keith Frankish and William M. Ramsey. Cambridge, UK: Cambridige UP, 89–107.

Bolz, Norbert (1994): "Computer als Medium – Einleitung." In: *Computer als Medium*. Eds. Norbert Bolz, Friedrich Kittler, and Christoph Tholen. München: Fink, 9–16.

Bostrom, Nick (2014): *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford UP.

Boyd, Eric (2017): "Microsoft and Facebook create open ecosystem for AI model interoperability." *Microsoft.com*. September 7. Online: https://www.microsoft.com/en-us/cognitive-toolkit/blog/2017/09/microsoft-facebook-create-open-ecosystem-ai-model-interoperability/ [Last access: 2018/03/06].

Callon, Michel (1986): "The Sociology of an Actor-Network: The Case of the Electric Vehicle." In: *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Eds. Michel Callon, John Law, John, and Arie Rip. Sheridan House Inc. 29–30.

Claburn, Thomas (2016): "How to steal the mind of an AI: Machine-learning models vulnerable to reverse engineering." *The Register*. Online: https://www.theregister.co.uk/2016/10/01/steal_this_brain/ [Last access: 2018/03/06].

Dahlberg, Lincoln, and Sean Phelan, eds. (2011): *Discourse Theory and Critical Media Politics*. Basingstoke: Palgrave Macmillan.

Dotzler, Bernhard (1989): "Know/Ledge: Versuch über die Verortung der Künstlichen Intelligenz" *MaschinenMenschen. Katalog zur Ausstellung des Neuen Berliner Kunstvereins*, 17.–23.07. Berlin: NBK. 127–132.

Dotzler, Bernhard (2006): *Diskurs und Medium. Zur Archäologie der Computerkultur*. Bd. 1. München: Fink.

Engemann, Christoph and Andreas Sudmann, eds. (2018): *Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*. Bielefeld: transcript. (Forthcoming, pre-publication version)

Floridi, Luciano (2017): "The rise of the algorithm need not be bad news for humans." *Financial Times*, May 4. Online: https://www.ft.com/content/ac9e1 0ce-30b2-11e7-9555-23ef563ecf9a [Last access: 2018/03/06].

Finn, Ed (2017): *What Algorithms Want. Imagination in the Age of Computing*. Cambridge, MA. MIT P.

Galloway, Alexander R. (2004): *Protocol. How Control Exists After Decentralization*. Cambridge, MA: MIT P.

Galloway, Alexander R. (2011): "Black Box, Schwarzer Block." *Die technologische Bedingung*. Ed. Hörl, Erich. Frankfurt/M.: Suhrkamp, 267–280.

Gershgorn, Dave (2016): "We don't understand how AI make most decisions, so now algorithms are explaining themselves." *Quartz*. December 20. Online: https://qz.com/865357/we-dont-understand-how-ai-make-most-decisions-so-now-algorithms-are-explaining-themselves/ [Last access: 2018/03/06].

Gershgorn, Dave (2017): "The data that transformed AI research – and possibly the world." *Quartz*. July 26. Online: http://www.notey.com/@qz_unofficial/external/17246232/the-data-that-transformed-ai-research%E2%80%94and-possibly-the-world.html [Last access: 2018/03/06].

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016): *Deep Learning*. Cambridge, MA; London: MIT P.

Hilgers, Philip von (2009): "Ursprünge der Black Box." *Rekursionen. Von Faltungen des Wissens*. Eds. Ana Ofak and Philipp von Hilgers. München: Fink, 281–324.

Horgan, Terence und John Tienson (1996): *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT P.

Huang, Jensen (2016): "Accelerating AI with GPUs: A New Computing Model." *Nvidia*. Online: https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/.

"Introducing OpenAI." *OpenAI Blog*. Online: https://blog.openai.com/introducing-openai/ [Last access: 2018/03/06]

Knight, Will (2017): "The Dark Secret at the Heart of AI". *MIT Technology Review*. 4. April 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [Last access: 2018/03/06].

Latour, Bruno (1999): *Pandora's Hope. Essays on the Reality of Science Studies*. Cambridge, MA: Harvard UP.

Latour, Bruno (2005): *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford: Oxford UP.

Matzner, Tobias (2017). "Opening Black Boxes Is Not Enough – Data-based Surveillance In Discipline and Punish And Today." *Foucault Studies* 23: 27–45.

Memisevic, Roland (2018): "Wunderwerke der Parallelisierung." In: Sudmann/Engemann, a. a. O., o. S. (Pre-publication version)

Metz, Cade (2016): "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free." *Wired*. April 28. Online: https://www.wired.com/2016/04/openai-elon-musk-sam-altman-plan-to-set-artificial-intelligence-free/. [Last access: 2018/03/06]

Mitchell, Thomas (1997): *Machine Learning*. New York: McGraw-Hill.

Newell, Allan and Herbert A. Simon (1997 [1976]): "Computer Science as Empirical Inquiry. Symbols and Search." *Mind Design II: Philosophy, Psychology, Artificial Intelligence*. Ed. John Haugeland, Cambridge, MA: MIT P, 81–110.

Nott, George (2017b): "Google's research chief questions value of 'Explainable AI'" *Computerworld*. 23. Juni 2017. Online: https://www.computerworld.com.au/article/621059/google-research-chief-questions-value-explainable-ai/ [Last access: 2018/06/03].

Park, Dong Huk et al. (2016): "Attentive Explanations: Justifying Decisions and Pointing to the Evidence." 14. December. Online: https://arxiv.org/pdf/1612.04757v1.pdf [Last access: 2018/03/06].

Parks, Lisa and Nicole Starosielski, eds. (2015): *Signal Traffic. Critical Studies of Media Infrastructures*. Chicago: U of Illinois P.

Pasquale, Frank (2015): *The Black Box Society. The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard UP.

Pasquinelli, Matteo (2017): "Machines that Morph Logic: Neural Networks and the Distorted Automation of Intelligence as Statistical Inference," *Glass Bead journal*, Site 1, "Logic Gate: The Politics of the Artifactual Mind."

Perez, Carlos E (2017): "Why AlphaGo Zero is a Quantum Leap Forward in Deep Learning." *Medium.com*. Online: https://medium.com/intuitionmachine/the-strange-loop-in-alphago-zeros-self-play-6e3274fcdd9f [Last access: 2018/03/06].

Reichert, Ramón, ed. (2014): *Big Data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*, Bielefeld: transcript.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016): "Introduction to Local Interpretable Model-Agnostic Explanations (LIME)A technique to explain the predictions of any machine learning classifier." *O'Reilly*. August 12. Online: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime [Last access: 2018/03/06].

Rodriguez, Jesus (2017): "Technology Fridays: OpenAI Gym Makes Reinforcement Learning Real." *Medium.com*. Online: https://medium.com/@jrodthoughts/

technology-fridays-openai-gym-makes-reinforcement-learning-real-bcf762c16774 [Last access: 2018/03/06].

Searle, John. R. (1980): "Minds, brains, and programs." *Behavioral and Brain Sciences* 3 (3): 417–457.

Sprenger, Florian (2015): *The Politics of Micro-Decisions: Edward Snowden, Net Neutrality, and the Architectures of the Internet*. Lüneburg: Meson P.

Sudmann, Andreas (2016): "Wenn die Maschinen mit der Sprache spielen." *Frankfurter Allgemeine Zeitung* Nr. 256, 2.11., N2.

Thielmann, Tristian (2013): "Jedes Medium braucht ein Modicum." *ZMK Zeitschrift für Medien- und Kulturforschung* 4/2: "ANT und die Medien," 111–127.

Trask Andrew, David Gilmore, Matthew Russell (2015): "Modeling order in neural word embeddings at scale." *Proceedings of the 32nd International Conference on Machine Learning* (ICML-15); Lille, France. July 6–11.

Vincent, James (2016): "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day." *The Verge*. March 24. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist [Last access: 2018/03/06].

Vincent, James (2018): "Elon Musk leaves board of AI safety group to avoid conflict of interest with Tesla." *The Verge*. Feb 21. Online: https://www.theverge.com/2018/2/21/17036214/elon-musk-openai-ai-safety-leaves-board [Last access: 2018/03/06].

Vogl, Joseph (2008): "Becoming-media: Galileo's Telescope." *Grey Room* 29 (Winter): 14–25.

Voosen, Paul (2017): "How AI detectives are cracking open the black box of deep learning." *Science Mag*. July 6. Online: http://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning [Last access: 2018/03/06].

Weizenbaum, Joseph (1976): *Computer Power and Human Reason. From Judgment to Calculation*. New York: W. H. Freeman.

Wiener, Norbert (1961 [1948]): *Cybernetics: or the Control and Communication in the Animal and the Machine*, Cambridge, Mass.: MIT P.

Zaller, John (1999): *A Theory of Media Politics. How the Interests of Politicians, Journalists, and Citizens Shape the News*. Chicago, IL: Chicago UP.