The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms

Sudmann, Andreas (Ed.)

Veröffentlichungsversion / Published Version
Sammelwerk / collection

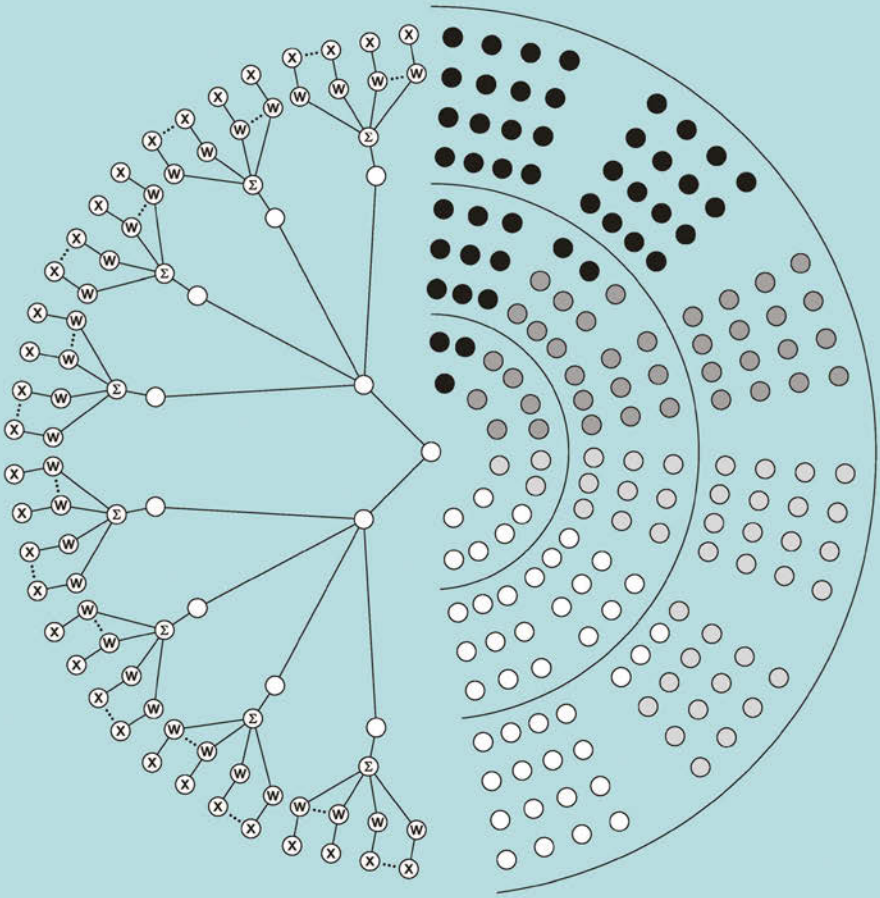
Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
transcript Verlag

Empfohlene Zitierung / Suggested Citation:
Sudmann, A. (Ed.). (2019). *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms* (KI-Kritik / AI Critique, 1). Bielefeld: transcript Verlag. https://doi.org/10.14361/9783839447192

Andreas Sudmann (ed.)

THE DEMOCRATIZATION OF ARTIFICIAL INTELLIGENCE

Net Politics in the Era of Learning Algorithms



[transcript] AI Critique | KI-Kritik

Andreas Sudmann (ed.)
The Democratization of Artificial Intelligence

AI Critique | Volume 1

Editorial

Since Kant, critique has been defined as the effort to examine the way things work with respect to the underlying conditions of their possibility; in addition, since Foucault it references a thinking about »the art of not being governed like that and at that cost.« In this spirit, **KI-Kritik / AI Critique** publishes recent explorations of the (historical) developments of machine learning and artificial intelligence as significant agencies of our technological times, drawing on contributions from within cultural and media studies as well as other social sciences.

The series is edited by Anna Tuschling, Andreas Sudmann and Bernhard J. Dotzler.

Andreas Sudmann teaches media studies at Ruhr-University Bochum. His research revolves around aesthetic, political and philosophical questions on digital and popular media in general and AI-driven technologies in particular.

Andreas Sudmann (ed.)

The Democratization of Artificial Intelligence

Net Politics in the Era of Learning Algorithms

[transcript]

An electronic version of this book is freely available, thanks to the support of libraries working with Knowledge Unlatched. KU is a collaborative initiative designed to make high quality books Open Access for the public good. The Open Access ISBN for this book is 978-3-8394-4719-2. More information about the initiative and links to the Open Access version can be found at www.knowledgeunlatched.org.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de

Cover layout: Maria Arndt, Bielefeld
Cover illustration: Julia Eckel, Bochum Typeset by Justine Buri, Bielefeld
Printed by Majuskel Medienproduktion GmbH, Wetzlar
Print-ISBN 978-3-8376-4719-8
PDF-ISBN 978-3-8394-4719-2
https://doi.org/10.14361/9783839447192

Content

The Democratization of Artificial Intelligence
Net Politics in the Era of Learning Algorithms

Andreas Sudmann

Diagnoses of time are naturally a difficult undertaking. Nevertheless, it is probably an adequate observation that, in our present historical situation, the concern for the stability and future of democracy is particularly profound (cf. Rapoza 2019). The objects of this concern are, on the one hand, developments which seem to have only a limited or indirect connection with questions of technology, such as the current rise of right-wing populism and authoritarianism, especially in Europe and in the US, or "the resurgence of confrontational geopolitics" (Valladão 2018). On the other hand, we witness an increasingly prevalent discourse that negotiates the latest developments in artificial intelligence (AI) as a potentially serious threat to democracy and democratic values, but which—with important exceptions—seems to be largely disconnected from the specific political conditions and developments of individual countries (cf. Webb 2019). Within this discourse, problematizing AI as jeopardizing democratic values and principles refers to different, but partly linked phenomena. Central reference points of these discussions are, for instance, the socio-political consequences of AI technologies for the future job market (catch phrase: "the disappearance of work"), the deployment of AI to manipulate visual information or to create 'fake news', the geo-political effects of autonomous weapon systems, or the application of AI methods through vast surveillance networks for producing sentencing guidelines and *recidivism risk* profiles in criminal justice systems, or for demographic and psychographic targeting of bodies for advertising, propaganda, and other forms of state intervention.[1]

Prima facie, both forms of concern about the global state of democracy do not have much in common, but it is precisely for this reason that one needs to explore their deeper connections. For example, US President Donald Trump recently launched a so-called "American AI initiative", whose explicit goal is to promote the development of smart technologies in a way that puts American interests first.

1 It goes without saying that not all of those aspects that for some reason appear to be worthy of critique represent an immediate danger to the democratic order of a society. However, it is also obvious that government and society must find answers to all problems of AI.

At about the same time, Google/Alphabet announced that they had opened their first AI Lab in Ghana. Headquartered in Silicon Valley, the tech giant continues its strategy of establishing AI research centers all around the world: New York, Tokyo, Zurich, and now Ghana's capital Accra. According to the head of the laboratory, Moustapha Cisse, one of its goals will be to provide developers with the necessary research needed to build products that can solve some of the problems which Africa faces today. As an example of the successful implementation of such strategies, it is pointed out that with the help of Google's open source machine learning library TensorFlow an app for smartphones could be developed that makes it possible to detect plant diseases in Africa, even offline.

The 'humanistic' AI agenda of Google/Alphabet and other tech companies seems, at first glance, to be in sharp contrast to the "America First" AI policy by Donald Trump. However, the fact that the Silicon Valley corporations are increasingly striving to promote democratic values such as accessibility, participation, transparency, and diversity has nothing to do with a motivation to distance themselves from the course of the current US government. Rather, the number of critics who see Google, Facebook, and the other tech giants themselves as serious threats to democracy and/or acting in contrast to democratic values, in terms of their business strategies, data practices, and enormous economic and socio-cultural power, is growing.

Accordingly, these companies have been under considerable pressure to respond to this increasing criticism. Facebook in particular was involved in two major scandals, both concerning Trump's presidential campaign. First, in 2017, it gradually became known that Russian organizations and individuals, most of them linked to the Saint Petersburg based Internet Research Agency (an internet troll farm), had set up fake accounts on platforms such as Facebook, Twitter, and Instagram, and attempted to capitalize on controversies surrounding the 2016 US presidential election, partly by means of creating fake news. Another scandal involved the data analysis and political consulting company Cambridge Analytica. As it became public in March 2018, the company had access to and presumably analyzed the data of over 80 million Facebook users without their prior consent in order to support Trump's campaign.

As a consequence of these scandals, not only Zuckerberg but also Google's CEO Sundar Pichai recently testified to Congress in Washington. During those hearings, Zuckerberg in particular admitted several failures in the past and promised to intensify cooperation with government institutions and NGOs, as well as to investigate measures to improve data protection and finally to implement them accordingly. As far as Europe is concerned, the European General Data Protection Regulation ("GDPR") already contains legal requirements for improving and complying with data protection. In the congressional hearings, Zuckerberg declared that he is in principle willing to support similar measures of state regulation in the

US. At the same time, he expressed fears that Chinese competitors could technologically outperform his corporation because the country traditionally puts much less emphasis on data protection issues than Europe or the US (cf. Webb 2019). However, there are other reasons for Facebook's willingness to cooperate in terms of data protection policies: At least since the takeover of WhatsApp and Instagram, Facebook has achieved a de facto monopoly position in the social media sector. The situation is similar with Amazon in e-commerce and Google in search engines – and it is precisely this enormous hegemonic position which is increasingly subject of intense debates. Recently, even co-founder and former spokesman of Facebook, Chris Hughes (2019), criticized Zuckerberg's company as a threat to the US economy and democracy, and advocated for the company to be broken up in order to allow more competition in the social media sector. For various reasons, it is rather questionable whether such a scenario could occur in the near or distant future. Nevertheless, criticism of global "platform capitalism" (Srnicek 2016) or "surveillance capitalism" (Zuboff 2018) is growing, and this also concerns the role of AI in what recently has sometimes been called the new "data economy" (cf. for instance Bublies 2017).

Not least with regard to the problems and phenomena mentioned so far, the aim of this volume is to explore the political dimension of AI, with a critical focus on current initiatives, discourses, and concepts of its so-called 'democratization'. One of the special characteristics of the latter term is that it is vague and concrete at the same time. As the current AI discourse reveals, the concept can refer to many different phenomena and yet evokes an ensemble of more or less corresponding or coherent conceptions of its meaning. Accordingly, democratization can be understood as the realization of an ethic, aiming at political information, a willingness to critique, social responsibility and activity, as well as of a political culture that is critical of authority, participative, and inclusive in its general orientation. Democratization can thus be conceived as a political, interventionist practice, which in principle might be (and of course has been) applied to society in general as well as to several of its subsystems or individual areas (like technology).[2]

One central question to be critically examined in this volume is to what extent network politics (and particular those related to ideas and activities of democratization) have been placed under new conditions with a view to the broad establishment and industrial implementation of AI technologies. The concept of network politics is understood here as a heuristic umbrella term for a broad spectrum of

2 Of course, in political theory, the term also signifies a transition to a more democratic regime, or describes the historical processes of how democracies have developed. For a discussion of the term democracy and democratization cf. Birch (1993), for discussing on the relationship of democracy and technology, cf. for instance the contributions in Mensch/Schmidt (2003), Diamond/Plattner (2012) or Rockhill (2017).

critical research, to shed light on the different forms of how networks and politics are intertwined and related, both as socio-technical discourses and practices. As such, it addresses the network dimension of politics as well as the political conditions, implications, and effects of different types of social, cultural, or technological networks, including but not limited to the Internet or so-called social media.[3] Accordingly, the volume does not only aim at exploring the political aspects of the relationship between AI and Internet technologies in the narrower sense (e.g. legal frameworks, political content on social media etc.). Rather, the critical focus involves looking at the networked and mediated dimension of *all* entities involved in the production and formation of current and historical AI technologies.

First of all, such a task needs some clarifications regarding the concept of AI because the term encompasses various approaches which are not always precisely differentiated, particularly in public discourse. When people talk about AI these days, their focus is mostly on so-called machine learning techniques and especially artificial neural networks (ANN). In fact, one can even say that these accounts are at the very center of the current AI renaissance. Sometimes, both terms are used synonymously, but that is simply wrong. Machine learning is an umbrella term for different forms of algorithms in AI that allow computer systems to analyze and learn statistical patterns in complex data structures in order to predict for a certain input x the corresponding outcome y, without being explicitly programmed for this task (cf. Samuel 1959, Mitchell 1997). ANN, in turn, are a specific, but very effective approach of machine learning, loosely inspired by biological neural networks and essentially characterized by the following features (cf. Goodfellow/Bengio/Courville 2016):

1. the massive parallelism of how information is processed/simulated through the network of artificial neurons
2. the hierarchical division of the information processing, structured in learning simple patterns to increasingly complex ones, related to a flexible number of so-called hidden layers of a network
3. the ability of the systems to achieve a defined learning goal quasi-automatically by successive self-optimization (by means of a learning algorithm called "backpropagation")

Indeed one can claim that the current boom of ANN and machine learning in general is quite a surprise, given that the technological foundations of this so-called connectionist approach in AI have already been researched since the early days of computer science and cybernetics (cf. e.g. McCulloch/Pitts 1943, Hebb 1949, Rosenblatt 1958). However, with the notable exception of some shorter periods,

3 For an overview on the long tradition of research on net politics, cf. for example Lovink (2002).

ANN have been considered more or less a dead-end in the history of AI research (Sudmann 2016, 2018a). This assessment is likely to be radically different today, even if a considerable number of commentators are pointing to (still) fundamental limitations of ANN or continue to uphold the importance of other approaches in AI research, for instance symbolic and rule-based forms (cf. Pasquinelli 2017, Marcus 2018).

There is some dispute concerning when exactly the current AI boom started. Some experts stress certain development leaps around 2009 in the field of natural language processing (NLP) and speech recognition. However, progress in the field of computer vision (CV) was of particular importance. In 2012, a research team at the University of Toronto won a competition for image recognition called ImageNet, reducing the error rate of previous approaches by more than half. This leap in performance became possible because so-called convolutional neural networks (CNN), i.e. networks optimized for the task of computer vision, were, for the first time, consistently and effectively trained on the basis of GPUs, i.e. fast, parallel-organized computer hardware, as they have been typically implemented in modern game consoles (Sudmann 2016).

In any case, the major IT corporations also quickly registered progress in the field of computer vision and ANN, which led to a veritable boom in the acquisition and financing of start-ups. One of these start-ups was DeepMind, which was acquired by Google in 2013 for 650 million US dollars. Three years later DeepMinds's AI system AlphaGo was able to beat the human world champion in the board game Go. With the success of AlphaGo, the AI boom had arrived in the mainstream, i.e. AI quickly became a dominant discourse in many areas of culture and society, including most fields of sciences (Sudmann 2018a, 2018b).

The latter does not mean that ANN were completely unknown in the fields of humanities and social sciences in the years before 2016. Especially around the early 1990s, interest in ANN grew considerably in areas like cognitive science and the philosophy of mind, shortly after the first industrial implementations of ANN took place and thanks to the establishment of the backpropagation learning algorithms in the 1980s (Sudmann 2018a, cf. also the interview with *Alexander Waibel* in this anthology). However, it can hardly be denied that in many disciplines the overall attention for ANN was rather limited even back then. In the end, the upswing of ANN in the 1980s turned out to be quite short, which is why some observers feel validated in their belief that the next AI winter will come – it is just a question of time. Of course, such an event could happen again, but currently there is no indication for this, rather the contrary seems to be the case.

Nevertheless, the ubiquitous talk of an "AI revolution" and the rhetoric of progress by Silicon Valley techno-utopists alone is a massive provocation for many critics, not only in the field of humanities, but also outside the academic world. Undeniably, since the very beginning, the debate on AI has typically been char-

acterized by either skeptical, utopian or dystopian narratives (cf. Sudmann 2016, 2018b).[4] And even today, careful mediations between these positions are still rare. As such, many discussions on AI are geared towards the speculative horizon of a near and distant future. And it is also no coincidence that AI has been described ironically as the very field of research that is concerned with exploring what computers *cannot yet do* (cf. Michie 1971). In other words: As soon as a computer masters certain abilities, such a system is no longer considered to be AI. Hence, AI is permanently shifted into the realm of utopia (or dystopia).

At the same time, we have only recently entered a historical stage in which the gap between AI as science fiction or technical utopia and AI as existing technology of the empirical world seems to be closing. Of course, one may rightly point out here that, for example, self-driving cars were already being tested on roads during and even before the 1980s,[5] or that first machine translation systems for languages were actually being developed in the 1950s (cf. Booth/ Locke 1955), but this does not change the fact that both technologies have only recently acquired or come close to the potential of applicability that the global economy expects of them.

AI's industrial usability and its increasingly outperforming human capabilities in various fields of applications seem to be new phenomena. However, computers have been a form of 'AI' from the very first day and were as such able to do things humans (alone) were not equally capable of, for example cracking the code of the German encryption machine Enigma (cf. Kittler 2013, cf. Dotzler 2006).

Given the rapid speed of new innovations and the expansion of fields of application, it is by no means an easy task to determine how AI reconfigures the relation between humans, technology, and society these days and impacts how we might be able to grasp the political and historical dimension of this shift in an adequate manner.

Finding an answer to this question implies a reflection of problems that have been discussed in the AI debate since the very beginning, for example the transferability of traditionally anthropocentric concepts such as perception, thinking, logic, creativity, or learning to the discussion of 'smart machines'. Indeed, it is still important to critically address the anthropological difference between humans and machines, to deconstruct the attributions and self-descriptive practices of AI, as *Anne Dippel* and *VN Alexander* demonstrate in their respective contributions. In her essay, *Anne Dippel* combines three stand-alone commentaries, each dealing with a different facet of AI, and each revolving around a different underly-

4 Already back in the late 1980s, the German media scholar Bernhard Dotzler wrote that all known forecasts of AI could already be found in Turing's writings (1989).

5 For example, the so-called Navlab group at Carnegie Mellon University has been building robot vehicles since 1984. Carnegie Mellon was also the first university to use ANN for developing self-driving cars.

ing metaphor: intelligence, evolution, and play. Her first commentary constitutes an auto-ethnographic vignette which provides a framework for the reflection on artificial 'intelligence' and the alleged capacity of machines to 'think'; both—as Dippel argues—very problematic metaphors from a feminist perspective with regard to the (predominantly) female labor of bearing and rearing intelligent human beings. The second one is an insight into her current ethnographic fieldwork amongst high-energy physicists, who use machine-learning methods in their daily work and succumb to a Darwinist metaphor in imagining the significance of evolutionary algorithms for the future of humanity. The third commentary looks into 'playing' algorithms and discusses the category of an 'alien', which, albeit controversial in the field of anthropology, she considers much more suitable in order to understand AI than a direct personification, bringing a non-human entity to life. *VN Alexander* in turn stresses in her text that there is no evidence that AI systems are really capable of making 'evidence-based' decisions about human behavior. AI might use advanced statistics to fine-tune generalizations; but AI is a glorified actuary table, not an intelligent agent. On the basis of this skeptical account, she examines how Alan Turing, at the time of his death in 1952, was exploring the differences between biological intelligence and his initial conception of AI. Accordingly, her paper focuses on those differences and sets limits on the uses to which current AI can legitimately be put.

In addition to a critical analysis of current AI discourses and its central concepts, it is equally important to understand the assemblages of media, infrastructures, and technologies that enable and shape the use of AI in the first place. To meet this challenge, it is necessary to take due account of the specific characteristics and historical emergence of the heterogeneous technologies and applications involved (cf. Mckenzie 2017). *Axel Volmar's* contribution "Productive Sounds: Touch-Tone Dialing, the Rise of the Call Center Industry and the Politics of Voice Assistants", for example, reflects on the growing dissemination of voice assistants and smart speakers, such as Amazon's Alexa, Apple's Siri, Google's Assistant, Microsoft's Cortana, or Samsung's Viv, which represent, in his words, a "democratization of artificial intelligence by sheer mass exposure". He engages with the politics of voice assistants, or more specifically, of conversational AI technologies by relating them to a larger history of voice-based human-machine interaction in remote systems based on the workings of "productive sounds"—from Touch-Tone signaling through on-hold music and prerecorded messages to interactive voice response (IVR) systems. In this history, Volmar focuses on changing forms of phone- and voice-related work and labor practices and different forms of value extraction from the automatization and analysis of telephonic or otherwise mediated speech. He argues that while domestic and potentially professional office end users embrace voice assistants for their convenience and efficiency with respect to web searches and daily routines; businesses, tech corporations, surveillance

states, and other actors aim to gain access to the users' voice itself, which is seen as a highly valuable data source—a 'goldmine'—for AI-based analytics.

Another interesting field in which AI and in particular machine learning techniques are increasingly deployed is the financial market and its various forms of algorithmic trading. As *Armin Beverungen* shows in his article, financial trading has long been dominated by highly sophisticated forms of data processing and computation in the dominance of the "quants". Yet over the last two decades high-frequency trading (HFT), as a form of automated, algorithmic trading focused on speed and volume rather than smartness, has dominated the arms race in financial markets. Beverungen suggests that machine learning and AI are changing the cognitive parameters of this arms race today, shifting the boundaries between 'dumb' algorithms in HFT and 'smart' algorithms in other forms of algorithmic trading. Whereas HFT is largely focused on data and dynamics endemic to financial markets, new forms of algorithmic trading enabled by AI are expanding the ecology of financial markets through ways in which automated trading draws on a wider set of data (such as social data) for analytics such as sentiment analysis. According to Beverungen, in order to understand the politics of these shifts it is insightful to focus on cognition as a battleground in financial markets, with AI and machine learning leading to a further redistribution and new temporalities of cognition. A politics of cognition must grapple with the opacities and temporalities of algorithmic trading in financial markets, which constitute limits to the democratization of finance as well as its social regulation.

In order to shed light on the political dimension of global AI infrastructures, we should not only examine how AI is used in the private sector by the tech giants, but also take into account that the public sector is more and more on a quest to become data-driven, promising to provide better and more personalized services and to increase the efficiency of bureaucracy and empower citizens. For example, taking Norway as a case study, *Lisa Reutter* and *Hendrik Storstein Spilker* discuss early challenges connected to the production of AI-based services in the public sector and examine how these challenges reflect uncertainties that lie behind the hype of AI in public service. Through an ethnographic encounter with the Norwegian Labor and Welfare Administration's data science environment, their chapter focuses on the mundane work of doing machine learning and the processes by which data is collected and organized. As they show, decisions on which data to feed into machine learning models are rarely straightforward, but involve dealing with access restrictions, context dependencies, and insufficient legal frameworks. As Reutter and Spilker demonstrate, the data-driven Norwegian public sector is thus in many ways a future imaginary without practical present guidelines.

For the task of critically addressing the specifics of different AI phenomena, it is crucial to explore appropriate paths, concepts, and levels of critique. Since Kant, critique has meant questioning phenomena with regard to their function-

ing and their conditions of possibility. According to Foucault, critique can also be understood as the effort or even art to find ways "not to be governed like that and at that cost" (Foucault 1997 [1978]: 45). In turn, a further concept of critique seeks to examine the idealistic imaginations of society in comparison with its real conditions and to explore why and to what extent these social ideals may (necessarily) be missed (or not). For Marx, this form of critique entailed analyzing why one is confronted with the necessary production of illusion and false consciousness, a focus to which Adorno and Horkheimer felt equally committed in their critical analysis of the *Dialectic of Enlightenment* (1944/1972).

Of course, these are only some of many possible trajectories of critical thinking useful for a profound investigation of an increasingly AI-driven world. Furthermore, we should bear in mind that AI provides new constellations and configurations of socio-technological assemblages, which might not be investigated adequately through the lenses of old concepts of critique, as Geert Lovink has argued with regard to internet and social media technologies (2011: 88).

Hence, it is important to question the very concepts of critical analysis we mobilize for our understanding of digital culture. For instance, *Tobias Matzner's* text engages with some prominent critical positions regarding current applications of AI. In particular, he discusses approaches that focus on changes in subjectivity as an inroad for critique, namely Wendy Chun and Antoinette Rouvroy. While Rouvroy forms a general verdict against what she calls "algorithmic governance", Chun suggests to 'inhabit' the configurations of subjectivity through digital technology. Matzner's text aims at a middle ground between these positions by highlighting the concrete situation of the concerned subjects. To that aim, Linda Martìn Alcoff's work on habitualization as situated subjectivity is connected with reflections from media theory. In concluding, this perspective on situated subjects is connected to the question of a democratic configuration of AI technologies.

The question of AI critique concerns hardly less the problem of its appropriate scaling. In the chapter by *Jonathan Roberge*, *Kevin Morin*, and *Marius Senneville*, the authors contend that in order to connect the macro-level issues related to the culture of AI and the micro-level of inscrutability within deep learning techniques, a third analytical level is required. They call this mezzo-level "governmentality", i.e. they discuss how power relations and the distribution of authority within the field are specifically shaped by the structure of its organizations and institutions. Taking the Montréal hub as a case study—and based on their 2016-2018 ethnographical work—they focus on two interrelated matters: a) the redefinition of the private-public partnership implied in deep learning, and b) the consequences of the "open science model" currently in vogue.

Furthermore, we should take into account that recent developments of smart machines may reflect some general shifts and continuities in shaping the infrastructures and environments of human-machine relations. The essay "Reduction

and Participation" by *Stefan Rieger*, for example, deals with a noteworthy strategy in media environment. It is a movement towards a holistic conception of the body and an approach to include all senses—even the lower ones. Above all, according to Rieger, these senses play a crucial role in the course of a ubiquitous naturalization. The consequence is a story of technological evolution and its irresistible success which follows a storyline diverging from the well-known topoi of augmentation and expansion. The intentional reduction of a technically possible high complexity is conspicuous. It is affected by aspects of internet politics, democratization, and the question of who should have access to media environments at all (and in what way). "Reduction and Participation" meets the demands to include other species and forms of existence. The aim of such demands is to expand the circle of those who gain agency and epistemic relevance, which also affects the algorithms themselves, as Rieger argues.

The question of agency and epistemic relevance reminds us that the project of AI critique itself also has an important history that needs to be considered. In fact, the development of AI has always been accompanied by a critical reflection in terms of its political, social, or economic dimensions and contradictions. And oftentimes, the computer scientists and engineers themselves were the ones to articulate these different forms of critique.

For example, already the cyberneticist Norbert Wiener noted in 1950:

> Let us remember that the automatic machine, whatever we think of any feelings it may have or may not have, is the precise economic equivalent of slave labor. Any labor which competes with slave labor must accept the economic consequences of slave labor. It is perfectly clear that this will produce an unemployment situation, in comparison with which the present recession and even the depression of the thirties will seem a pleasant joke. This depression will ruin many industries—possibly even the industries which have taken advantage of the new potentialities. (Wiener 1988 [1950]: 162)

Indeed, one of the most intensively discussed AI topics today revolves around the speculative question of how far automation driven by robots and smart machines leads to a turmoil on the labor market and may cause extensive job loss. For example, AI experts like Kai Fu Lee believe that 40% of the world's jobs could be replaced by AI and robots within the next 15 years (Reisinger 2019; cf. also Frey/Osborne 2017). Such forecasts, however numerous they may be in circulation these days, are above all one thing: sometimes more, sometimes less well-derived or well-founded speculations. How the world will be in 15 years is not predictable, neither by clever scientists nor by intelligent machines. Nevertheless, Norbert Wiener's quote at least illustrates that critique and speculation go hand in hand, both then and now.

Similarly, many critical points made by Joseph Weizenbaum in his seminal work *Computer Power and Human Reason* (1976) enjoy a renaissance in current discussions on AI. In case of Weizenbaum's book, his critical intervention was twofold: On the one hand, he was also motivated to emphasize the fundamental differences between man and machine and/or between thinking/judging and calculating, including highlighting certain fundamental limits of what AI can be capable of; on the other hand, Weizenbaum warned that there are tasks that a computer might be able to accomplish but that it should not do. Many subjects discussed and arguments proposed by Weizenbaum are specifically echoed and further developed in current debates on "AI ethics" (cf. Cowls/Floridi 2018; Taddeo/Floridi 2018). But unlike Weizenbaum, whose critical reflections were essentially based on classic symbolic AI, today's AI ethics debate faces the challenge to adequately understand the media, technology, and infrastructures of machine learning systems and artificial neural networks, whose logic of operations are significantly different from what has sometimes been called "good old fashioned AI" (Sudmann 2018b). And this is a particularly difficult task, since due to the marginal status of ANN there is no profound tradition of expertise in this particular field of AI, neither in many disciplines of the humanities and social sciences, nor even in the natural and technical sciences (cf. also the interview with *Alexander Waibel* in this volume).

In addition, since the beginning of the AI boom, many of the leading researchers have given up their jobs as professors or employees at universities or taken leaves of absence to set up start-ups or work for the big tech giants. On the one hand, the enormous salary opportunities (whether as an employee or as the founder of a start-up) are tempting; on the other hand, many scientists also accept jobs with the major tech companies because they assume that the conditions for their research are significantly better in business than at university (for instance, in terms of access to learning data or powerful computers, access to funds for research).

Most companies and especially the countless start-ups that have been founded in recent years in the wake of the AI boom are also constantly complaining about the lack of experts in the field, which they perceive as a major brake for further innovations. Many institutions have recognized this problem and are investing billions in training, research, and development of AI. Nevertheless, the question arises according to which criteria, with which goals, and under which conditions this funding takes place. Against this background, it is imperative that private and public funding of AI also includes support for critical research. Certainly, the latter is above all a task for the humanities and social sciences. But in order to master this task adequately, they depend on dialogue and cooperation with the 'hard sciences'.

However, there is another reason why research on and with current AI technologies, especially with regard to their political dimension, poses a major challenge, which even experts cannot easily overcome. As has been extensively discussed in recent years, ANN in particular are regarded as a fundamentally opaque technology of AI. While computer scientists are in fact able to observe and measure the activity of each individual neuron and of their connections independent of their number, they cannot or only to a limited extent understand or explain the activities of ANN (cf. also my contribution to this volume). It is obvious that this specific black box problem has serious political-ethical implications and effects. For example, it is one thing whether AI technologies are used, say, for the recognition of medieval handwritings or for recommending certain products to consumers. However, when AI technologies are used to evaluate a person's creditworthiness or to decide whether a particular person might commit a particular crime based on their appearance and behavior, the situation is obviously a different one.

As *Dan Mcquillan* argues in his essay, AI is a political technology and is as such being used to sustain austerity, but its politics are obscured by its technical opacity and by a narrative of ethics. The concrete operations of AI, acting through statistical regression and optimization, produce thoughtlessness and epistemic injustice. Meanwhile, AI's predictive classifications extend bureaucratic governmentality into the future, which it seeks to preempt. However, AI is fragile and only solves what Bergson called "ready-made problems". According to Mcquillan, we need to approach AI in a way that enables us to take sides with the possible against statistical probabilities. His article sets out both a feminist and situated approach to developing non-oppressive AI, and the forms of collective community and workplace structures necessary to achieve it. Similarly, *Yvonne Förster* problematizes that especially current AI applications are a black box and operate without being able to give an account of the underlying reasons, and the underlying causal processes themselves also remain opaque. In her essay, she discusses the concept of invisibility and opacity from a phenomenological perspective and explores the relation of experience and perception to technology.

Democratizing AI

Compared to the long tradition of AI critique, the discourse of "democratizing AI" is a relatively new one. Basically, the discourse has emerged since it has become widely known that AI is now intervening in all areas of global culture and society. The following aspects, among others, have contributed to the emergence and dissemination of this discourse:

1. the extensive critique of AI technologies with regard to their social, economic, and political implications, manifestations, and effects
2. the long tradition of dystopian imaginations of AI
3. the practices of datafication and data analytics of the big tech companies and their hegemonic role in the current and future development of AI
4. the assessment of ANN as a fundamentally opaque technology of information processing and data analytics

Terms such as democratization and democracy are sometimes used as if one could always refer to them positively or affirmatively. At the same time, theories of democracy constantly remind us that the idea of democracy meaning the "rule of the people" presupposed significant exclusions at all times. In the ancient polis of Greece only free citizens—but not women, slaves, or someone who did not own land—were allowed to vote and act politically. This tradition of exclusion was bound to continue for a long time. According to John Locke's conception, which was decisive for the development of English parliamentarism, the right to vote was still given only to the property owners, and of course we should not forget that well into the 20th century, women were not allowed to take part in elections in democratic societies. Even today, people who have lived in a particular country for many years, although in principle subject to all of its laws, are excluded from national elections unless they have the necessary citizenship.

As we have recalled at the beginning, AI technology already has helped politicians to get elected. Against this background, it is obvious to ask whether and when machines themselves will be allowed to vote, or more generally to speculate whether and when they will be perceived as entities that possess certain rights, like a human being. And it is quite remarkable that even though machines are not allowed to vote (yet), they already can be elected—as it happened in 2018, when an AI system in Japan (Tama City, Tokyo) was running for mayor. The AI system in question promised that thanks to its statistical methods it could effectively evaluate the advantages and disadvantages of requests by citizens; it claimed to make fair decisions, to strive for consensus in conflicts of interest, and also to focus on absolute transparency with regard to the use of taxes. When the votes were counted, it turned out that the AI system came in last of all candidates. The outcome is perhaps unsurprising, even in technology-obsessed Japan. People there, as well as in other countries, might accept AI-systems and robots as tools, servants, or toys, but it seems difficult to imagine a political representation by machines other than in terms of very dystopian scenarios.[6]

6 Even though not only in Japan, but also in Europe or the US, the presence of machine is a normality in governments (also cf. Agar 2003).

Indeed, the very fact that the cultural imaginary of AI has been shaped so extensively by dystopian narratives probably still causes people to fear the coexistence with intelligent machines, or at least to feel profound discomfort. Against this background, recent efforts of democratizing AI, as described in the following, can indeed be understood as working against such a dystopian view of the common future of humans and machines, as people imagine it.

However, it is important to note here that the demand for a democratization of AI inevitably implies that such technologies are in themselves undemocratic, or at least have the strong tendency or potential to be incompatible with democratic values and practices. And there are good reasons for this conceptualization of AI. If the development of intelligent machines is aimed at replacing or surpassing humans, or if AI is seen as a driving force for economic growth and a condition for securing hegemonic geopolitical power, in all these instances, the technology has prima facie nothing to do with the establishment and protection of democratic values such as equality in the emphatic sense. Similarly, the current discussions about algorithmic biases point to fundamental problems of inequality and difference associated with the large-scale implementation of AI systems in all areas of society.

For instance, *Alexander Monea's* chapter examines how attempts to make computer vision systems accessible to users with darker skin tones has led to either the hypervisibility of phenotypic racial traits, particularly morphological features like hair texture and lip size, or the invisibility of race. Drawing on critical race theory and the problematic history of racial representation in photographic media, he demonstrates how racial biases are prevalent in the visual datasets that many contemporary computer vision algorithms are trained on, essentially hardcoding these biases into our computer vision technologies, like Google Photos. The most frequent industry reaction to these hardcoded racial biases is to render race invisible in the system, as was done with Google Photos. He further shows how the invisibility of race in computer vision leads to the familiar problems of 'color blindness', only expressed in new media. The author argues that these constitute fundamental problems for the potential democratization of AI and outlines some concrete steps that we might take to more strongly demand egalitarian computer vision systems.

Nevertheless, at least some people believe that AI might have the potential in itself to open up a new utopian horizon of freedom, equality, fraternity, and could furthermore even be used productively to secure world peace (cf. Valladão 2018). In Thomas Hobbes' theory of state, the Leviathan (as the embodiment of a fictive social and governing contract) is conceptualized as the necessary condition of possibility for a peaceful coexistence among people. Without it, mankind would fall back into the state of nature, into the war of all against all. However, as history since Hobbes has shown, the modern state is an extremely precarious, fragile en-

tity, incapable of providing lasting protection for *all its members*. More importantly, the sad truth is that there has never been a state or democracy since, say, the French Revolution that was fully able to meet the demands of freedom, equality, or solidarity in their emphatic sense.

Against this background, the utopian vision (and for some certainly dystopian imagination) of delegating responsibility for the political and the control of society entirely to machines does not seem completely absurd. But if one can envisage mankind deciding to better put their fate in the hands of superior machines, then it is also still conceivable that people at some point also might or will stand up for the realization of a truly better global society—without any help of AI as a political entity or "peace machine" (Honkela 2017).

Current concepts of democratizing AI, however, have little in common with a critique aiming at a fundamental transformation of society. Nevertheless, the concept of a democratic AI, as a project of the present, still remains very closely related to utopian visions and motivations inasmuch as it resembles ever so many strategies and concepts of democratization that have been developed throughout history in relation to taste, art, media, technology, or society as a whole.

Current ideas of democratizing AI share strong similarities with utopian-political ideas of the cyberspace, virtual reality, and of course the Internet, as they have been especially prevalent since the early 1990s (cf. Egloff 2002). The idea that cyberspace and/or the Internet (the concepts are not identical, yet often used as synonyms) are in themselves an emancipatory space that used to be called "cyber-utopianism" and has been the subject of criticism since 1995 at the latest, for example by the Critical Art Ensemble. Conversely, even today many scientists, artists, and net activists adhere to the idea that either the Internet and/or cyberspace actually mark a space of freedom, subversion, and resistance that must be defended, despite all its heterogeneous contradictions and problems.

The utopian-idealistic dimension of democratization is also visible in the current use of the concept by the large tech corporations in connection with AI. They present the concept of a "democratic AI" first and foremost as a great promise of universal, all-inclusive accessibility, participation, and transparency. For example, for Microsoft the democratization of AI essentially means putting the technology into the hands of "every person and every organization" (cf. Microsoft News Center 2016; cf. Johnson 2017a).

As far as the official agendas of tech giants are concerned, various strategies are currently being pursued to achieve this goal: First, a general idea is to advance the simplification, standardization, and automation of AI, so that even non-experts inside and outside companies and universities can increasingly use the corresponding technologies (such as ANN) for their purposes and applications. Second, the large IT companies want to grant users, scholars, and companies open access to various cloud services, from computational resources (such as Google's

Tensor Processing Units, i.e. specific chips to accelerate machine learning operations), to program libraries and frameworks like Scikit, PyTorch, Keras, or TensorFlow, training data sets like MNIST or ImageNet, to various other software tools that are helpful for the broader dissemination and improvement of AI.

Unsurprisingly, the corporations do not provide their services without individual interests or the expectation of anything in return: For example, Google requires researchers who use their resources to make their own research results and perhaps also their code available open source (Johnson 2017b). In addition, they speculate that open-sourcing their tools might also have the effect that independent developers contribute to their improvement without incurring significant costs (cf. Lomonaco/Ziosi 2018). Moreover, big companies like Microsoft benefit from the fact that the open source idea itself enjoys a high reputation in the tech research community and that researchers have an interest in their work being highly visible and widely recognized (cf. Bostrom 2017).

Critically engaging with the promise to provide developers access to emerging machine learning technologies and to enable them to infuse their applications with smartness or intelligence, *Marcus Burkhardt's* text asks how machine learning and AI as fields of technological development and innovation are structured in themselves. By providing an initial mapping of the coding cultures of machine learning and AI on GitHub, he argues that it is important to attend more closely to the hitherto largely neglected infrastructural layers of code libraries and programming frameworks for developing critical perspectives on the social and cultural implications of machine learning technologies to come.

Beyond certain advantages connected to different actions of opening AI, many researchers, institutions, and companies tend to stress that solving problems in this field is a collective endeavor that cannot be achieved individually, which is why it is necessary to share ideas and methods as widely and as openly as possible.

Problems and contradictions of economic and scientific competition, however, are rarely discussed. On the surface, it seems like AI research is essentially driven by an unbound idealism. The reality, however, is that the field is indeed characterized by fierce international competition for talent, capital, and other 'resources'. And at the heart of the big tech companies' agenda is the tenacious struggle for being the first to overcome the unsolved problems of AI and/or to achieve the ultimate goal of a *general artificial intelligence*, a so-called strong AI, i.e. a machine capable to master or learn any task similar to or better than a human being.

This also applies to the so-called non-profit organization that even has integrated the "openness" idea in its brand name: OpenAI. As I demonstrate in my own contribution for this volume, OpenAI has somehow been the avant-garde of the current "AI democratization" hype, also by foregrounding its commitment to democratic values like access, participation, and transparency. But if one examines the activities of the organization hitherto, the investment of OpenAI is

more about making progress to solve the foundational technological problems in AI, rather than focusing on how the concept of an open, democratic AI could be further developed in a technologically and conceptually meaningful way.

For the moment, if one critically examines the rhetoric of companies like Google or Microsoft, it looks as if the promise of a democratic AI has already been fulfilled by its accessibility. Especially in the case of technology, however, democratization not only means access to its use, but also the possibility of its control (cf. Lomonaco/Ziosi 2018). If and how such a process can be organized and shaped in a reasonable way, for example through state supervision or other measures, is still an open question, and maybe it cannot be answered in general. But the crucial point here is that those companies who advocate the "democratization of AI" must *at least in principle* be willing to restrict their sovereignty and/or to accept interventions by other external entities.

The latter, however, is unlikely to be in the interests of the large tech corporations. Indeed, the simple fact that the tech giants so fully embrace the idea of a "democratic AI" strongly indicates how little the concept threatens their economic or cultural power, quite to the contrary.

Nevertheless, the democratization of AI, as advocated by the large tech groups, is not only about controversial concepts of access, transparency, and participation. Furthermore, the concept also entails the goal to serve 'good purposes', i.e. solving the world's small and large problems. Microsoft's "AI for Earth" initiative, for example, aims at fighting climate change or eliminating inequalities in the health care system. Given such an agenda, it is, of course, awkward that Microsoft was recently accused of working with researchers from China's National University of Defense Technology, controlled by the country's Central Military Commission, collaborating on AI problems that commentators thought to be usable for state surveillance technologies. Microsoft dismissed these accusations by pointing out that the research papers in question had as much or little to do with surveillance as WiFi or a Windows operating system would have. In addition, the company pointed out that such forms of international cooperation are very typical in the field of AI research.

However the situation may be in this specific case, it is clear that especially in the field of AI, it has always been difficult to distinguish between a military and civilian use.[7] For example, similarly as with other AI application fields, a large part of research in the field of machine translation and natural language processing (for the political discussion of this field of AI, cf. the interview with Alexander

7 Sometimes it is also a matter of disputes within a company whether orders from the military should be accepted. Cf. the recent protests by Google employees against the so-called "Maven project" (cf. vgl. Shane/Wakabayashi 2018). For a recent discussion on the miliary use of AI see also Ernst/Schröter/Sudmann (2019).

Waibel in this volume), has been funded by the military, specifically by programs supported by DARPA. This commitment is no coincidence. Especially during the Cold War, there was a high demand for translations from Russian into English (and vice versa on the side of the Soviets for translations from English to Russian). Furthermore, global military operations and disaster management efforts have always stimulated a general interest in the rapid translation of large quantities of foreign-language texts. Finally, one should note that the field of machine translation had been based on basic mathematical and cryptologic knowledge from the start—developed during the Second World War by researchers in the military and secret services.

As the use of AI for military goals shows, "openness" and "transparency" cannot count as positive values per se. According to Nick Bostrom (2017), openness to security measures or openness about goals can be good, but openness about source code does not necessarily have to be. Accordingly, Bostrom advocates a differentiated approach to "open AI": When it comes to developing technologies that have the potential to cause considerable damage, they should naturally not be disclosed.

Particularly with regard to ANN technology, the fundamental question arises as to which extent requirements of transparency and openness can be realized at all, given that specifically the connectionist approach of ANN has to be understood as being fundamentally opaque at its core (Sudmann 2017). Nevertheless, various approaches of a so-called "Explainable AI" at least try to reduce the opacity of current AI systems.

As *Schieferdecker, Großmann*, and *Schneider* stress in their contribution to this volume, software-based systems using AI methods for different tasks are essentially characterized by their "criticality", by which they mean their usage in safety- and security-critical domains like transportation and automotive, banking and finance, healthcare, cyber-security or industrial automation. As the authors explain, this criticality of numerous AI-based systems demands rigorous and effective quality engineering in pre-deployment phases and at runtime. In their article, the authors review the state of the art in safeguarding AI-based systems by so-called "verification and validation methods", taking a particular look at the principal function components of AI-based systems and their extended quality requirements. Since any AI is primarily developed in software, the principal approach to the quality engineering of software-based systems in general is reviewed. According to Schieferdecker, Großmann, and Schneider, testing is the best-known and most effective V&V method and will most probably also form the basis for dealing with AI-based systems: It can be used for confirming or witnessing outcomes of AI-based systems, it can become a digital common for their comparison and benchmarking, and thus contribute to a shared knowledge basis of AI.

Against the background of the phenomena outlined so far, it is quite obvious that the political economy of AI is a great challenge for policymakers. As *Frank Pasquale* shows in his essay, so-called centralizers encourage the accumulation of data in very large firms, while in contrast decentralizers want to see more dispersed innovation. Although both have very different visions for long-run economic development, each can help counter the untrammeled aspirations (and disappointing everyday reality) of stalwarts of digital capitalism. They also contribute to our understanding when giant firms try to solve what Friedrich Hayek has identified as the "knowledge problem"—which and when they exacerbate it via obscurity and obfuscation. If conglomeration and vertical mergers actually promote AI that solves real-world problems—of faster transport, better food, higher-quality health care, and more—authorities should let them proceed. According to Pasquale, industrial bigness helps us understand and control the natural world better. But at the time, he argues that states should block the mere accumulation of bargaining power and leverage, even if it is in the service of AI development. Policymakers need to find ways to address the contradictions and diverging perceptions regarding the regulation of technology. One important task here is to translate political decisions into laws that are appropriate in practice, but that also take into account the criticism of these technologies. But what role can laws play in the democratization of AI? This is the question the chapter by *Christian Djefall* addresses. His text highlights the dimensions of AI's openness and shows that AI can be beneficial and detrimental to democracy. Constitutional law actually calls for a democratization of AI. Reliance on and delegation to AI systems requires a democratic rebalancing. The chapter then goes on to explore how AI can be democratized. It identifies three layers that describe a series of choices: the technical layer, the social layer, and the governance layer. On the technical layer, there are many choices to be made; a specific concept like designability could help to identify choices that enable democratic governance. The influence of AI systems is often not rooted in technology but attributed to AI through social choices. In administrative law, automated decisions are endowed with the power of the law. The governance layer shows how technologies can be influenced by overarching choices. This can be done for example by frames and organization. Taking all layers together, there is ample room for democratic determination of AI applications.

It is perhaps a question of debate to what extent machine learning algorithms as "cultural machines" (Finn 2017) already have an influence on our daily life and changed the sociocultural experiences we make in this world. The discussion on the cultural impact of machine learning and ANN has also recently intensified around the question of how AI can be considered creative and perhaps even changes our understanding of art (practices). The public discussions on this were fueled by an auction at Christie's, where a painting 'created' with the help of an AI-system was sold for a high price. Interestingly enough, the art collective responsible for

this painting claimed that they want to "explain and democratize AI through art". It was probably foreseeable that AI would also be coopted quickly by the art world. But recent discussions on "creative AI" tend to omit that the problem of machines' supposed creativity is by no means new, as *Jens Schröter* shows in his article. In fact, already in the 1960s, in so-called "information aesthetics", similar questions were discussed. In his essay, Schröter therefore historicizes the current debates and argues that the question of whether machine-creativity or machine-art is possible cannot be answered by abstractly contrasting 'man' and 'machine' (AI).

The relationship between art, creativity, and smart machines shows that the discussion about the politics and democratization of AI must not be restricted to certain areas (economy, military) or to certain groups of actors (e.g. "The Big Tech Giants"). Instead, we should consider that the critique of AI and the commitment to democratizing it is also supported by many NGOs, academic institutions, journalists, or politicians; actors whose efforts undoubtedly deserve their own portrayal. This book is therefore only a small contribution to a controversial field of discussion whose contours, relations, and conditions have yet to be explored.

References

Algar, Jon (2003): *The Government Machine. A Revolutionary History of the Computer.* Cambridge, MA: MIT Press.

Birch, Anthony H. (1993): *Concepts and Theories of Modern Democracy.* London et al. Routledge.

Bostrom, Nick (2017): "Strategic Implications of Openness in AI Development." In: *Global Policy*, 1-12. (https://nickbostrom.com/papers/openness.pdf).

Bublies, Pia (2017): "Data is Giving Rise to a New Economy." In: *The Economist*, May 6. (https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy).

Cowls, Josh and Floridi, Luciano (2018): "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society." June 19. (https://ssrn.com/abstract=3198732 or http://dx.doi.org/10.2139/ssrn.3198732).

Diamond, Larry Jay, and Plattner, Marc F., eds. (2012): *Liberation Technology: Social Media and the Struggle for Democracy.* Baltimore, MD: Johns Hopkins University Press.

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

— (2006): *Diskurs und Medium. Zur Archäologie der Computerkultur.* Band 1. München: Fink.

Egloff, Daniel (2002): *Digitale Demokratie: Mythos oder Realität*. Wiesbaden: Westdeutscher Verlag.

Ernst, Christoph, Schröter, Jens, and Sudmann, Andreas (2019): "AI and the Imagination to Overcome Difference." *spheres – Journal for Digitial Cultures* 6 (2019): forthcoming.

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

Foucault, Michel (1997 [1978]): "What is critique?" In: What is Critique? Semiotexte: Los Angeles, CA, 41-81.

Frey, Carl Benedikt and Osborne, Michael A. (2017): "The future of employment: How susceptible are jobs to computerisation?", Technological Forecasting and Social Change, Elsevier, vol. 114(C): 254-280. (https://ideas.repec.org/a/eee/tefoso/v114y2017icp254-280.html).

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

Hebb, Donald Olding (1949): *The Organization of Behavior. A Neuropsychological Approach*. New York: John Wiley & Sons.

Honkela, Timo (2017): *Rauhankone*. Gaudeamus.

Horkheimer, Max and Adorno, Theodor W. (1944/1972): *Dialectic of Enlightenment*. New York: Herder and Herder.

Hughes, Chris (2019): "It's Time to Break Up Facebook". In: *New York Times*, May 9. (https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html).

Johnson, Khari (2017a): "AI democratization depends on tech giants." In: *VentureBeat*. (https://venturebeat.com/2017/12/28/ai-weekly-ai-democratization-depends-on-tech-giants/).

Johnson, Khari (2017b): "Google unveils second-generation TPU chips to accelerate machine learning." In: *VentureBeat* (https://venturebeat.com/2017/05/17/google-unveils-second-generation-tpu-chips-to-accelerate-machine-learning/).

McCulloch, Warren S. and Pitts, Walter (1943): "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5.4: 115-133. DOI: https://doi.org/10.1007/BF02478259.

Locke, William N. and Booth, A. Donald, eds. (1955): *Machine Translation of Languages*. Cambridge, MA et al.: The Technology Press of The Massachusetts Institute of Technology, John Wiley & Sons, John Wiley & Sons, and Chapman & Hall.

Lomonaco, Vincenzo and Ziosi, Marta (2018): "On the Myth of AI Democratization." (https://medium.com/ai-for-people/on-the-myth-of-ai-democratization-a472115cb5f1).

Lovink, Geert (2003): *Dark Fiber—Tracking Critical Internet Culture*. Cambridge, MA: MIT Press.

Lovink, Geert (2011): *Networks Without a Cause. A Critique of Social Media*. London: Polity Press.

Marcus, Gary (2018): "Deep Learning. A Critical Appraisal." (https://arxiv.org/abs/1801.00631).

Mackenzie, Adrian (2017): *Machine Learners. Archeaology of Data Practice*. Cambridge, MA: MIT Press.

Mckenzie, Adrian (2017): *Machine Leaners. Archaeology of a Data Practice*. Cambridge, MA: MIT Press.

Mensch, Kirsten and Schmidt, Jan C., eds. (2003): *Technik und Demokratie. Zwischen Expertokratie, Parlament und Bürgerbeteiligung*. Opladen: Leske+Budrich.

Michie, Donald (1971): "Formation and Execution of Plans by Machine", in N.V. Findler & B. Meltzer, eds.: *Artificial Intelligence and Heuristic Programming*: New York: American Elsevier: 101-124.

Microsoft News Center (2016): (https://news.microsoft.com/features/democratizing-ai/)

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

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

Rapoza, Kenneth (2019): "Democracies In Crisis: Has The West Given Up On Democracy?" In: *Forbes*, January 9. (https://www.forbes.com/sites/kenrapoza/2019/01/09/democracies-in-crisis-has-the-west-given-up-on-democracy/#425fb6db1242).

Reisinger, Don (2019): "A.I. Expert Says Automation Could Replace 40% of Jobs in 15 Years." In: *Fortune*. (http://fortune.com/2019/01/10/automation-replace-jobs/).

Rockhill, Gabriel (2017): *Counter-History of the Present. Untimely Interrogations Into Globalization, Technology, Democracy*. Durham: Duke University Press.

Rosenblatt, Frank (1958): "The Perceptron. A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6: 386-408.

Samuel, Arthur (1959): "Some Studies in Machine Learning Using the Game of Checkers." In: *IBM Journal of Research and Development*, Vol. 3: 210-229.

Shane, Scott and Wakabayahsi, Daisuke: "'The Business of War': Google Employees Protest Work for the Pentagon." In: *The New York Times* April 4, 2018. (https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html).

Srnicek, Nick (2016): *Platform Capitalism*. Cambridge, UK; Malden, MA/USA: Polity Press.

Sudmann, Andreas (2016): "Wenn die Maschinen mit der Sprache spielen." In: *Frankfurter Allgemeine Zeitung* No. 256, Nov. 2: N2.

Sudmann, Andreas (2018a): "Zur Einführung. Medien, Infrastrukturen und Technologien des maschinellen Lernens." In: *Machine Learning. Medien, Infrastruk-*

turen und Technologien der Künstlichen Intelligenz. Ed. Christoph Engemann and Andreas Sudmann. Bielefeld: transcript: 9-23.

Sudmann, Andreas (2018b): "Szenarien des Postdigitalen. Deep Learning als MedienRevolution" In: Engemann/Sudmann, 55-73.

Taddeo, Mariarosaria and Floridi, Luciano (2018) "How AI can be a force for good", *Science* 361 (6404): 751-752.

Valladão, Alfredo Da Gama e Abreu (2018): "Artificial Intelligence and Political Science". *Policy Paper,* September, 1-29.

Webb, Amy (2019): *The Big Nine. How The Tech Titans & Their Thinking Machines Could Warp Humanity*. New York: Hachette Book Group.

Weizenbaum, Joseph (1976): *Computer Power and Human Reason. From Judgment to Calculation*. Oxford, England: W.H. Freeman & Co.

Wiener, Norbert (1988 [1950]): *The Human Use of Human Beings: Cybernetics and Society*. London. Free Association.

Zuboff, Shoshana (2018): *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York et al.: Public Affairs.

Metaphors We Live By[1]

Three Commentaries on Artificial Intelligence and the Human Condition

Anne Dippel

Prelude

In the following essay, I want to bring together three stand-alone commentaries, each dealing with a different facet of artificial intelligence, and each revolving around a different underlying metaphor: intelligence, evolution, and play. The first commentary constitutes an auto-ethnographic vignette, which provides a framework for the reflection on artificial "intelligence" and the alleged capacity of machines to "think"; both very problematic metaphors from the feminist perspective on (predominantly) female labour of bearing and rearing intelligent human beings. The second one is an insight into my current ethnographic fieldwork amongst high-energy physicists who use machine-learning methods in their daily work and succumb to a Darwinist metaphor in imagining the significance of evolutionary algorithms for the future of humanity. The third commentary looks into "playing" algorithms and brings into the conversation the much-debated anthropological category of an "alien" which, as I argue, is much more relevant in order to understand AI than a direct personification, bringing a non-human entity to life.

A New Non-artificial Intelligent Life is Born

I am looking at a newly born human being. Day by day I keep him company, as he practices increasingly complex bodily movements, senses the inner emotions of other bodies around him or reacts to a sea of indistinguishable voices, despite not being able to understand the meaning of a single word. While he keeps to his

1 The title of a famous book published by George Lakoff and Mark Johnson in 1980. I want to thank Sonia Fizek for her invaluable help in revising this article.

own reflexes, I am witnessing a life-changing event: the emergence of an all but artificial intelligence. Slowly, the motor activities become increasingly controlled, the musculature is gradually building up, and the gaze seems to follow points of interest somewhat consciously, with a dose of curiosity and awe. A young human learns.

Seeing the development of a new life, makes me radically rethink the concepts of artificial intelligence and machine learning, and even more so the significance of language, which has the power to shape political reality.

Can machines think, asked Alan M. Turing almost seventy years ago (1950). His provocative metaphor until today conditions the way computer scientists tend to perceive the capacity of algorithms to process data and yield "intelligent" (or rather intelligible) results. The image of an intelligent machine has grown strong in the public eye. Today, we talk of "smart" infrastructures, smart TVs, smart homes, even smart cities; all exemplifying the so-called "smartness mandate" (Halpern, Mitchel, Gheoghegan 2017).

Can machines learn? It is no longer a question, but an assumption and a method used in almost every discipline reliant on big data, from physics, over marketing and finance to agriculture. Thinking and learning, inherently human qualities, when used with reference to machines seem to make little sense. They are often dismissed as innocent metaphors. But words have power. Not only do they describe the surrounding reality, but shape the way we think and act (Lakoff and Johnson 1980). In that sense, machine "intelligence" is much more than a rhetorical device. It influences our perception of it as an (in)human quality.

The concept of intelligence originates from a very specific and narrow understanding of what it means to behave as an intelligent entity. Christoph von der Malsburg, considered a pioneer of artificial intelligence and originally trained as a particle physicist, in his neurobiological research on intelligence focused mainly on visual cognition and memory (Malsburg 1990). It is not difficult to draw a parallel to the contemporary understanding of machine learning algorithms, often praised for their beyond human capacity to recognize patterns out of a pool of gargantuan data sets. To an anthropologist who considers anthropocentric criteria of difference to be fundamentally suspect, this oversimplified human versus machine metaphorical comparison seems somewhat disappointing in its naiveté, if not spine-chilling. Von der Malsburg triumphantly argued that human brains do not exceed the memory capacity of more than one gigabyte. But humans are not fed with raw data sets. And machines, unlike humans, do not necessarily have a palimpsestuous biological memory of experiences but rather are an extended memory, to play along with von der Malsburg's metaphor of a capacious container for data storage.

Above all, human intelligence and memory do not stand in an one-dimensional relationship to each other. Intelligence is an embodied process, highly depen-

dent on received attention and care. It is enough to take a quick look at a newly born human to dismiss the blind enthusiasm of computer science to create artificial life. In this context, machine learning seems like an empty disembodied metaphor. It is the body (of the infant and their mother), which is central in the development of intelligence. For a newborn, the physical and the psychological are inseparable. The body and the mind are not yet split, subject to Cartesian dualism. They do not exist as separate entities, or rather exist in a mutual embrace. All is embodied, and all is mindful. Facial expressions, gestures and voice operate without the socio-cultural censor. Their face slowly learns how to laugh, at first coincidentally, later in a more focused manner. It seems, as if the baby's consciousness was gradually contracting to a fully developed "I". At first small threads appear like, then they expand, grow and open to become a mindful being. But before that happens, the baby simply exists. Infants develop their intelligence in dealing with the environment. They demand to be noticed and perceived although they are not able to understand what attention really is.

All those daily observations I have been collecting as a feminist mother and an anthropologist have lead me to believe that any comparison of human and artificial intelligence must be considered bizarre if not utterly pointless at best. The observations of the social and emotional complexity of an infant, whose head accounts for a third of its body weight and who has no language and can be more than language at the same time, have made it clear to me that the concept of an undifferentiated intelligence as such is the most dangerous aspect in the political debate on AI. At the heart of research on artificial intelligence lies an extremely oversimplified and disembodied understanding of the term, which not only overestimates machine intelligence and underestimates the biological complexity of humans, but brings with it the danger of dismissing the significance of being a responsible human agent altogether.

While neuro-computer scientists spent time dreaming of self-replicating algorithmic intelligence, uncounted female bodies keep nourishing and nurturing the yet to be born human intelligence. While science keeps appropriating humans as embodied metaphors to praise the artificial life instead, a true wonder of creation a female body is capable of, remains barely touched by the admirable gaze of the (overwhelmingly male) techno-scientific world. It is the politics of embodied care (Hamington 2001) or politics of care in technoscience (Martin, Myers, Viseu 2015) that needs to be brought back into a larger social conversation on artificial intelligence and its relation to what it means to be human.

The Promethean Dream of Artificial Intelligence in Physics

In my usual anthropological fieldwork I do not study infants, but sit vis-à-vis scientists who work with artificial intelligence; to be more precise with very specific machine learning algorithms, which are able to sieve through endless data of particle decays. The European Center for Nuclear Research (CERN) is home to quite a few high-energy particle physicists who see themselves as "gods playing with the help of the computer". At CERN, researchers increasingly rely on supervised machine learning in their everyday work. Already in the 1980s the so-called MVA (multivariable analysis), a form of machine learning, was deployed at CERN (Galison 1997).

At first, high-energy particle physicists developed algorithms for pattern recognition of rare subatomic collision events independently of computer scientific expertise. The communities of physicists and computer scientists were not always as strongly connected as they are today. With the establishment of the "particle accellerator Large Hadron Collider" (LHC), however, those two seemingly distant communities merged. High Energy Physics has experienced a gradual "informatization" of its knowledge base, dependent on high-performance computers capable of storing data density and performing the Monte Carlo analyses required to predetermine events and test theories on the basis of physical measurements.

In the past 15 years more and more computer scientists have entered the everyday research practice as CERN annual statistics indicate, supporting physicists in coding and simulating experiments (CERN Annual Statistics Website 2019). CERN invests in computer scientists and in different areas of computer research, from machine learning algorithms to quantum computing. The "trained" algorithms collect, detect, and analyze seas of data. Contemporary high-energy physicists may be described as "code sorcerers" (Chun 2013), making sense of the world through the lens of pseudo-random algorithms. Thus, it is no surprise that their visions for the future of humanity are so deeply conditioned by the logics of the algorithmic infrastructure "living" around them. Most of the physicists, however, would dismiss this assumption. They tend to perceive algorithms as mediated tools, which may have the capacity to extend our minds, but at the same are entirely controlled and tamed by physicists. Both categories, the human and the machine, are clearly separated, each having a different role and hierarchy in the experiment. Physicists are convinced of the superior position of humans vis-à-vis algorithms, however intelligent. If there is any doubt about the semiotic-material analysis of physics, it usually is voiced outside of the field, for example in media studies or philosophy, i.e. disciplines, that reflect the "mediatedness" of contemporary knowledge in natural sciences. Physics sees itself as an impartial referee, untouched by the logics of the medium. In other words, how and what the observer sees remains uninfluenced by the apparatus devised to see the observed.

At the same time, the convictions of an almost sterile human-tool separation are accompanied by the speculations of a future cyborg, a human of tomorrow enhanced by artificial intelligence and almost inseparable from it. Such cyborgian visions are shared by many physicists, especially those working in the departments devoted to more speculative and future-oriented research at CERN, for instance on the so-called evolutionary algorithms inspired by the principles of biological evolution (reproduction, mutation, recombination, selection). It is here that one can find computer science visionaries like Rodrigo Suarez, one of my informants. In machines he sees a continuum of intelligence, developing from a single cell to a fully-fledged human and reaching their final state in a computer. Even if he is not entirely convinced that AI could reach a human-like status, he dreams that one day humans could evolve and live eternally, free from fear and illness, as cyborgs enhanced by artificial intelligence. Rodrigo Suarez does not see any difference between the concepts of intelligence of a biological cell, a computer or that of a human being. In our conversation I drive him to the edge of his argumentation, but for Rodrigo Suarez (and many other computer scientist) these exist only advantages of an eternal life, even if the immortality dream is to be reached by the fittest few. The principle of evolution does not account for fairness or justice for all. There seems to be a crude Darwinist opinion embedded in the algorithmic concepts that drive current research politics on AI. While computer science is bringing man back to the centre, natural culture research decenters him. The enlightenment figure spelled with capital "M" (Tsing 2015) reclaims his position of power. Evolutionary algorithms, still in an early developmental stage, rest on the dream of fusing "epistemology and ontology" (Bruder 2018, 153), as well as mind and body with technology, contributing to the raise of *homo automaton sapiens*.

For some this might be just a narcissist dream of production and reproduction (uterus envy?), maybe even a hubris in the ancient Greek sense, a way of playing Prometheus or Eva, trying to steal the flame or the apple (Dippel 2011). It is hard to find balance, it seems, between techno-optimism and techno-pessimism, especially for a scientist working as one of the new shamans of technology. Regardless, any politics of artificial intelligence needs to take humans into account.

Artificial Intelligence as an Alien at Play

The Promethean dream seems to be best illustrated when machines and humans "face" each other at a play table, in a direct ludic confrontation. In the recent history of cybernetics several pivotal games took place, for instance Mac Hach VI versus US Chess Federation player (1967) or the iconic IBM's supercomputer Deep Blue versus Garry Kasparov (1996, 1997). In both cases the human was defeated

by the sheer power of computation. In 2015, a very different contestant entered a global scene. Alpha Go, a computer program able to play the game of Go (much more strategically complex than chess), won against a human player. Following the first victory, it went on to beat the professional Go player Lee Sedol. AlphaGo uses a Monte Carlo tree search algorithm (the same method used in high-energy physics at CERN) to find new optimal moves.

Such examples show how deeply the longing for human-machine comparison is embedded within the history of technological development. Humans are the standard that serves for technology as the main criterion in terms of intelligence. The game between Lee Sedol and AlphaGo has also raised the question of "alienness"—does artificial intelligence play in a different way than humans do? Can we use the category of "play" with reference to an algorithm at all? Do computers play? All the above questions are more complex than it seems, especially when taking into account the fact that AlphaGo opted for moves which, in their appetite for extreme risk, seemed almost inhuman. As the Deep Mind team emphasizes: "AlphaGo's strategy embodies a spirit of flexibility and open-mindedness: a lack of preconceptions that allows it to find the most effective line of play" (DeepMind.com). Artificial intelligence tends to deal well with a vision of a potentially harmful sacrifice, if it leads to an unparalleled compensation in the game. On a more general philosophical level, we could say that it has no consciousness or any understanding of its own possible "death". This opens a very different playfield, in which every decision can be as risky as the logics of checks and balances allows for.

Artificial intelligence remains in a non-existential relationship to anything that matters to humans (cf. Dippel 2018). After all, machines have been created precisely for the purpose of relieving or facilitating the existential condition of humankind (cf. Giedion 1982). One could argue from an anthropological perspective that man—the "capital M guy that made the anthropocene" (Tsing 2015)—has created a "metaphorical counterpart" of himself (Lévi-Strauss 1973, 238); a dispositive of difference in times when the conventional border regulations between humans and other living creatures have become questionable. I see thus two major pathways in the visions of AI. On the one hand, we can observe the production of an artificial intelligence as a "metaphorical counterpart", to extend upon the anthropologist Claude Levi-Strauss and his comparison between humans and birds. Both species form relationships and build nests amongst many other similarities, but there is one thing that we as humans cannot do—flying. In that sense birds are seen as a metaphorical counterpart, in which the dream of flying and extending our limited capacities is stored. Artificial intelligence is like a bird of sorts. It allows us to see what we are and what we are not; what we dream to become, but can perhaps never be. On the other hand, the inclusive version of artificial intelligence based on the concepts of a "third nature" (Richter & Rötzer 2018), of cyborgs

(Haraway 1991) and of nature-culture (Gesing, Knecht, Flitner & Amelang 2019), existing regardless of the political sphere and the social consequences.

The first concepts of artificial intelligence, as Norbert Wiener famously put it, were about creating modern slaves (1972, 72). The old fears of the relationships between master and servant are reflected in the debates about the politics of artificial intelligence since its early days (Winner 1977). Instead of looking for an order that would enable a better society, the current concepts blindly reproduce existing relations of domination and post-colonialism. The vision of artificial intelligence today succumbs to mostly neoliberal and positivist worldview, pushing the ideal for a never ceasing automated work (Gregg 2018). Fostering class-biased dreams to bring an end to the working class, it serves predominantly elitist fantasies. It does not consider creating a sustainable environment allowing humans to find their place within nature. Instead, it fosters nature as "the other" that needs to be dominated through technology.

But technology tends to wander off in unforeseeable directions, providing fertile ground for ideology (Latour 2006). Current issues around social media are serving as a very fitting example here. Made to connect friends and families across the globe, they have become disruptive and manipulative tools in the political sphere, deeply influencing the human capacity to understand complex texts or to keep attention for an extended time. This perhaps trivial example only shows that it is of paramount importance today to investigate artificial intelligence not only from a specifically technical angle, but in a broader socio-cultural and political context. As researchers and as citizens, we need to stay alert.

"Fed" by the People and for the People

Artificial intelligence should be seen for what it truly is, a technological alien. To neglect this "alienness" or otherness of AI it so to misunderstand its capacity to lead to a utopian potential for other politics. In fact, only by treating AI as the technologically Other allows us to see it as something that "eludes the orders of self and culture, while at the same time challenging them" (Leistle 2015). And to challenge the status quo, we may begin with a conscious use or criticism of powerful metaphors, attributing to AI either human capacities or embedding it within a specific socio-political framework (in this case, a neo-liberal and positivist one).

The White House report on artificial intelligence of the late Obama administration reads: "Developing and studying machine intelligence can help us better understand and appreciate our human intelligence. Used thoughtfully, AI can augment our intelligence, helping us chart a better and wiser path forward" (Technology Council Committee 2016, 7, 39). Such grandiose political assumptions, however, should be embedded in a new social reality, where every citizen

has open access to the AI-driven goods. Researchers, politicians, the private sector and public opinion need to come to the point of communalization and people's empowerment of artificial intelligence, which may be difficult imagine in the current political and economical system. In that sense, AI should be owned by the people, because it is overwhelmingly "fed" by the people, for instance in a daily practice of using digital technology and thus allowing technology companies to collect our data in order to feed their algorithms shrouded behind corporate non-disclosure agreements. The future of humanity and AI should not succumb to a Darwinist vision. In this utopian context, artificial intelligence could be a true medium, and a mediator—not a dark privatized Leviathan, manipulated for those who love to lead war, hold power, and accumulate resources. For a vision like this to come true, a larger social dialogue is needed reaching beyond the optimization logics of fast computing and automated labour. It asks for humans that practice *vita activa* and take on responsibility instead of dreaming to outsource it to a techno-god.

With this remark I would like to bring this essay to a closure for a much more demanding creature is waiting to be nourished, not with raw data, but with milk, attention and care. His intelligence will require many more years to develop, independent from the super-computer's calculating power and Monte Carlo search algorithms. Feeding my son requires much more than "having enough content" (Stokel-Walker 2019). It is a labour of love, passed by women and men from generation to generation since the beginning of humanity. One, which does not need a "metaphorical counterpart" in technology.

Bibliography

Barr, Alan J., Andy Haas and Charles W. Kalderon (2016): "'That looks weird'— evaluating citizen scientists' ability to detect unusual features in ATLAS images of LHC collisions". ATL-COM-OREACH-2016-017, arXiv:1610.02214v1.

Bruder, Johannes (2018): "Where the sun never shines. Emerging Paradigms of Post-Enlightened Cognition." In: Digital Cultures and Society 4/1, pp. 133-157.

CERN Annual Statistics Website (2019): (https://cds.cern.ch/collection/CERN%20Annual%20Personnel%20Statistics?ln=de) [Last access 24.4.2019].

Chun, Wendy Hui Kyong (2013): Programmed Visions: Software and Memory. Cambridge, Mass.: MIT Press.

DeepMind.com: (https://deepmind.com/research/alphago). [Last access 24.4.2019].

Dippel, Anne (2017): "Das Big Data Game. Zur spielerischen Konstitution kollaborativer Wissensproduktion in der Hochenergiephysik am CERN." In: *NTM* 4 (2017), pp. 485-517.

Dippel, Anne (2018): In: Feige, Daniel M., Ostritsch, Sebastian, Rautzenberg, Markus (Eds.): *Philosophie des Computerspiels. Theorie – Praxis – Ästhetik*. Stuttgart: Metzler, pp. 124-148.

Dippel, Anne (2011): Ironisches Prolegomenon für einen „Entartungsschutz des Menschen" zum vernünftigen Wesen vom Homo Sapiens Sapiens zum Homo Sapiens Optivus. In: *Ist Technik die Zukunft der menschlichen Natur?* Göttingen: Wehrhahn, pp. 104-114.

Executive Office of the President National Science and Technology Council Committee on Technology (2016): Preparing For the Future Of Artificial Intelligence (https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf) [Last access 24.4.2019].

Galison, Peter (1997): *Image and Logic. A Material Culture of Microphysics.* Chicago: Chicago University Press.

Gesing, Friederike, Knecht, Michi, Flitner, Michael, Amelang, Katrin (Eds.) (2019): *NaturenKulturen. Denkräume und Werkzeuge für neue politische Ökologien.* Bielefeld: transcript.

Giedion, Sigfried (1982): *Die Herrschaft der Mechanisierung. Ein Beitrag zur anonymen Geschichte.* Frankfurt a. M.: athenäum.

Gregg, Benjamin (2018): The Coming Political. In: *Digital Cultures and Society* 4/1, pp. 157-180.

Halpern, Orit, Mitchel, Robert, Geoghegan, Bernard Dionysius (2017): The Smartness Mandate: Notes Toward a Critique. In: *Grey Room*, 68, pp. 106-129.

Hamington, Maurice (2001): Jane Addams and a Politics of Embodied Care. In *The Journal of Speculative Philosophy* 15/2, pp. 105-121.

Haraway, Donna (1991): A Cyborg Manifesto. Science, Technology, And Socialist Feminism in the Late Twentieth Century. In: *Simians, Cyborgs and Women: The Reinvention of Nature.* New York: Routledge, pp. 149-181.

Lakoff, George, Johnson, Mark (1980): *Metaphors We Live By*. Chicago: Chicago University Press.

Latour, Bruno (2006): Ethnografie einer Hochtechnologie: Das Pariser Projekt „Aramis" eines automatischen U-Bahn-Systems. In: Rammert, Werner (Ed.): *Technografie. Zur Mikrosoziologie der Technik.* Frankfurt a.M., New York: campus, pp. 25-60.

Leistle, Bernhard (2015): Otherness as a paradigm in anthropology. In: *Semiotica: Journal of the International Association for Semiotic Studies* 204, pp. 291-313.

Lévi-Strauss, Claude (1973): Das wilde Denken. Frankfurt a.M: Suhrkamp.

Martin, Aryn., Myers, Natasha, Viseu, Ana (2015): *The Politics of Care in Technoscience*. In: *Social Studies of Science* 45/5, pp. 1-17.

Richter, Steffen, Rötzer Andreas (2018): *Dritte Natur 1. Technik Kapital Umwelt*. Berlin: Matthes &Seitz.

Stokel-Walker, Chris (2019): Feeding algorithms is a full-time job. BBC (http://www.bbc.com/capital/story/20190307-the-hidden-armies-that-power-the-internets-new-stars) [Last access 24.4.2019].

Tsing, Anna (2015): A Feminist Approach to the Anthropocene: Earth Stalked by Man. Lecture held at the Barnard College for Women. (https://www.youtube.com/watch?v=ps8J6a7g_BA) [Last access 24.4.2019].

Von der Malsburg, et al. (1990): *Pattern segmentation in associative memory.* Neural Computatio 2, pp. 94-106.

Wiener, Norbert (1950): *The Human Use of Human Beings.* Boston: Houghton Mifflin.

Winner, Langdon (1977): *Autonomous Technology. Technics-out-of-Control as a Theme in Political Thought.* Cambridge, Mass: MIT Press.

AI, Stereotyping on Steroids and Alan Turing's Biological Turn

V. N. Alexander

Introduction

Artificial Intelligence (AI) designers try to mimic human brain capabilities with "self-learning" neural networks trained by crowd-sourced selection processes or other "unsupervised" selection processes. Presumably, the logic of the input data is inscribed in the structure of the artificial network similarly to the way input shapes a human brain. Yet decades on, AI-trained chat bots and translation apps still fail to vault the low bar of the Turing Test. It is becoming clear that AI is not able to interpret signs within fluid contexts. Is biological computation qualitatively different from present-day machine computation? At the time of his death, Alan Turing was investigating how biological reaction-diffusion processes create patterns, which, in turn, constrain cellular responses and differentially trigger development. Similar mechanisms are now thought to provide the temporal and spatial constraints for ensembles of neurons allowing them to perform sensory binding and to form and recall memories. Had he lived to continue his work, Turing might have reoriented AI research to better address the challenge of creating contextual constraints, which may be what is needed to produce the unpredictable and almost miracle-like responses we call human judgement. As it is, organized statistically, current AI applied to human affairs is only good for stereotyping, which, of course, undermines the basic premise of individual democratic freedom.

Like an organism, a "smart" machine can seek an object, read a code, locate a pattern and make generalizations. Like an organism, a machine can even be designed to pursue self-preserving goals. However, we cannot say that machines currently possess humanoid intelligence. AI bots cannot understand people because they are not good with language. They do not get irony, new metaphors, metonyms, puns or jokes. Language is fundamentally allusive, not literal, as Turing once demonstrated in a letter to a friend:

> Turing believes that machines think
> Turing lies with men
> Therefore machines do not think (1952b).

The fact that Siri cannot get this joke is not because there is not world enough and time to train the network; it is symptomatic of the essential difference between AI and Biological Intelligence (BI). Selection processes, such as those used to train AI networks, cannot evolve true intelligence. I can make such a bold statement because selection processes, such as those that neo-Darwinists have claimed have evolved animal intelligence, do not, in fact, do the job. The failure of AI chat bots is the proverbial dead canary indicating a much bigger problem, an oversimplified conception of the evolution and development of intelligent action. AI may be better than BI at mechanistic rule-bound actions like driving cars, but it is incapable of determining what humans mean or intend to do. The public should not be asked to trust AI, accepting that how it works is just a mystery. This paper aims to pull back the Wizard of AI's curtain, revealing that this allegedly superhuman intelligence is in fact just a tool, a very powerful one, that is being used by a few to control the many.

A Twenty-First Century Evolutionary Theory of Innovation

To theoretical biologists it is becoming clear that, although the natural selection of small random changes in genetic material plays a role in evolutionary processes, the outcome of such selection is the stabilization of a species and the reduction of diversity. Innovation, we now believe (See Turing, 1952a; Margulis & Sagan, 2002; Reid, 2007; Shapiro, 2011; Noble, 2016) is likely due to large, interrelated mutational events, like hybridization, gene duplication, lateral gene transfer, transposons, symbiogenesis and, importantly to this discussion, the thermodynamic self-organizing semiotic processes discovered by Turing. Such mechanisms tend to produce new ready-made tools (not randomly assembled stuff) whose functions can then be selected or not. This isn't your father's evolutionary theory.

AI is designed on the assumption that adaptive learning follows the random-change with gradual selection neo-Darwinian model of the 1950s. AI, like natural selection, makes generalizations based on a statistical definition of fitness: the most frequently reappearing patterns are selected. AI learns with repeated positive and negative reinforcement. BI can learn this way too, but it can also have epiphanies.

Algorithms Versus Semiotic Habits

To try explain why AI lacks of a sense of humor, I start by noting that while computers use *digital codes* and develop *algorithms* apart from contexts, living cells use *analog signs* and develop self-reinforcing *semiotic habits* within contexts. This paper will explore the differences between AI and BI from the perspective of Biosemiotics, a newly developing, transdisciplinary field related to the fields of Cybernetics, Complex Systems Science and Biochemistry. According to Biosemiotics, whereas a code requires a precise translation of one form into another, a sign can be translated into a variety of forms depending on the relative similarity and/or proximity of other signs and transducers. It may be that this flexibility of biological signs allows signal transduction to flow easily, to be communicated synchronously and coherently to neighboring cells, even if the signal is not quite the correct or conventional one. It may be that this difference between AI and BI can account for AI's failure to adequately translate signs in contexts.

If AI's self-learning algorithms seem to work well sometimes to predict human actions, this is because stereotypes are often true. AI is currently being used in US court systems to help determine sentences, exaggerating structural social prejudices in the data fed to the AI network. The likelihood that a criminal will re-offend is predicted by categorizing him or her as a type. The result is blacks get tougher sentences than whites with comparable data points (Angwin et al., 2016). AI is stereotyping on steroids.

AI is also being applied to the management of the public at large. According to Andrew Hallman, Deputy Director for Digital Innovation at the US Central Intelligence Agency, thanks to all the data collected on Internet users, the agency can now use Deep Learning to better "anticipate the development of social unrest and societal instability…three to five days out" (Konkel, 2016). This has me worried that a pre-crime unit is up and running. It cannot be true that sacrificing our privacy will keep us safe. Mass surveillance and Big Data collection can only serve the purpose of silencing dissent and maintaining the status quo, not anticipating actual crimes. Complex systems, like humans, tend to behave non-linearly: the ability to predict individual behavior does not improve in proportion to the increase in the amount of data that is used to make the predictions.

Although neural net designers use feedback and feedforward in an attempt to mimic non-linear biological processes, no creative mechanism is included, and the resulting intelligence resembles nameless, faceless bureaucracies that have accreted procedures for dealing with citizenry over many generations and which are not only conservative but which tend to narrow options more and more with each iteration.

The first step toward democratizing AI is to unmask this supposedly better-than-human judge. AI is no agent; it is a powerful and potentially useful tool

that should be in the hands of the many not just the few. A democratic digital society, Michael Kwet (2018) has cogently argued, requires uncompromised privacy, open-source software, as well as decentralized personal cloud systems that allow direct sharing of information. If collection of personal data is thus halted, courts and surveillance agencies will not be able to use AI to control individuals based on their memberships in or associations with various groups. I hope that my analysis of the differences between present-day AI and BI can convince the public to be more skeptical of the supposed wisdom or accuracy of AI predictions.

What Turing Knew about BI

Turing invented the most practical tool humans will likely ever wield. And yet his engineering successes were driven by an impractical desire to understand the nature of human intelligence. We follow his lead here as we try to understand how to best use computing tools in the twenty-first century.

In the 1950s, after proposing his model for a "self-learning" computer, Turing's thinking began to take, what might be called in hindsight, a Second-Order Cybernetics or Artificial Life turn. He began conducting experiments and studies in mathematical biology. While Andrew Hodges (1983), Turing's main biographer, saw his interest in plant and animal morphogenesis as a departure from his interest in mimicking intelligence, Jack Copeland (2004), who provides the definitive commentary on Turing's science, points out that Turing made it clear that this new work was a further investigation of intelligent computation, even though his attention had fallen upon giraffe patterns, Fibonacci spirals and leaf generation.

Turing discovered the spontaneous processes by which unorganized systems organize themselves without interference, without external selection. C.H. Waddington (1940) had suggested to Turing that development simply falls into order somehow, flows down the path of least resistance. It was Turing who suggested that an instability, a chance pattern—*not* an inducer specifically designed for that function—could initiate the flow from less order to more order, from chaos to differentiation (or, to nod to Gregory Bateson, a difference that makes a difference). Ilya Prigogine, who won the Nobel Prize for related research, met with Turing in Manchester in 1952 and discussed the theory (Hodges, 1983: 587). Not until 1972 in a paper in *Physics Today* did Prigogine recognize Turing's contribution.

Turing argued that reactions that diffused away from the point of instability result in the so-called morphogenic fields that differentially determine gene action, as described by Waddington. While biologists were interested in what this meant for embryology, what Turing was after was knowledge of how neurons might similarly differentiate and self-organize. In a 1951 letter to neurophysiologist J. Z. Young, Turing remarks, "The brain structure has to be one which can

be achieved by the genetical embryological mechanism" (qtd in Copeland, 2004: 517). Although we do not know if Turing thought development was analogous to learning, it turns out that it is. His mother Sara Turing may not have been incorrect when she opined that her son had been on the verge of an "epoch-making discovery" when he died (qtd in Hodges, 1984: 624).

AI Compared to BI

When Turing first designed his self-learning network computer, he had assumed, tacitly following neo-Darwinism, that humans make *random* guesses when they do not know a procedure for solving a problem. In 1948 in "Intelligent Machinery," he claims, "training a human child depends largely on a system of rewards and punishments" for good and bad guesses respectively (Copeland, 2004: 425). Turing designed a chess-playing program with optional moves that could be tried at random. If a move ultimately led to failure, it would not be reinforced. Turing's neural network was designed to start out unorganized and become organized with appropriate "interference," mimicking eduction. Similar kinds of *connectionist* approaches are used today in most self-learning algorithms. The "instruction table," as Turing called a program, is embodied in the network as it is altered by reward and punishment. Feedback can be administered by a programmer or crowd-sourced on the Internet. Although this approach is called self-learning or self-organizing, as Turing noted, such approaches still require "interference" from the outside.

The newest phase of AI is referred to as "unsupervised" learning. For example, a visual recognition network was exposed to millions of random unlabeled images on the Internet. It eventually detected some commons patterns of, you guessed it, cat faces, acquiring pathways and biases in unknown ways, hundreds of levels deep (Le et al., 2013). Programmers did not tell the network what to find, but the network can now be used to find cat faces. These new unsupervised networks are not so dissimilar to Turing's 1948 notion of a self-learning network. The main difference is the point at which the programmer interferes, during the training process to target a pre-specified pattern or after the network as detected a pattern that is of interest to the programmer. In the latter case, the unit of selection is the entire network, not individual connections within the network.

Animals most often learn in "unsupervised" situations, especially non-human animals, and are less often taught, intentionally rewarded and punished. It is the monkey see, monkey do approach. But interference, or selection, is still at work. Experiencing a procedure over and over, actually changes neuronal connections. Neurons that fire together wire together, as Donald Hebb (1949) so famously not-

ed. Learning by rote, strengthening connections over time, is statistical in nature. What happens the *most*—whether it is "right" or not—gets selected and reinforced.

Repetition is one way neurons develop connections, but not the only.

Humans (and probably other animals too) can recall details better in contexts, if they are associated with things arbitrarily similar or arbitrarily nearby. Rhymes, rhythms, tones and other poetic devices, such as metaphor and metonymy tend to aid memory, even if the connections are not repeatedly reinforced. The semiotic habits of neuronal groups may be initiated by rare stochastic resonances (i.e., purely coincidental patterns) which lead to self-organization. A source of unpredictability in human logic and language use, this poetic type of sign action in and among neuron cells dominates subconscious processes. Subjects under hypnosis experience cross-modal perception—they begin to hear colors, for example—which indicates that when conscious perception is bypassed, the poetic workings of the subconscious are more observable (Alexander and Grimes 2017). People with synesthesia are better able to recall arbitrary facts because numbers or letters can be associated with unique colors, textures and shapes (Harvey 2013). Connections based on arbitrarily similar/proximate factors cannot be reduced to statistical description; the *number* of factors is not as relevant to outcomes as the *qualities* of the factors vis-a-vis other factors.

Formalizing Biosemiotics

Could a computer model the way nature organizes itself by linking things arbitrarily similar/proximate? Turing discovered non-linear equations that can produce computer-generated zebra stripes, invagination, metachronal waves and other natural emergent patterns. Although for years Turing's work went unproven and many believed the similarities between the patterns generated by his equations and those found in nature were merely coincidental, Sheth (2012) and Raspopovic (2014) have finally shown that a Turing mechanism does indeed describe the process whereby fingers are created in developing embryos. It has taken some time for biologists to identify the actual chemical signals that correspond to kinds of relationships Turing imagined would have to obtain if self-organization were a mechanism for differentiation and development. Turing's equations are complex, but suffice it to say that they involve variables for diffusion rates, reaction rates, and the ways in which these rates change. Reactions typically involve a number of morphogens, for example, X and Y react to produce Z; Z and A react to produce 2Y. The first reaction depletes Y; the second increases Y. To put it differently, the process might involve an activator that can catalyze its own production *and* that of its own inhibitor, which, in some cases, might diffuse away rapidly, setting the stage for traveling wave patterns to emerge. There is contradiction or

paradox in these processes, which are both self-creating and self-constraining, a bit like Turing's syllogism introduced at the beginning of this paper.

Let me try to elucidate the biosemiosic elements of these types of processes with a very simplified visual model with only two elements. To illustrate biological computation, I use shapes with material qualities as symbols because the binding of biological signals and receptors (sign readers) is often shape dependent. Let us say we have a molecule type ⊥ and molecule type ⊤. They can be turned in various directions, e.g., ⅂ and ⊥. Neither ⊥s or ⊤s interact with themselves. So

1. ⊥ + ⊥ = ⊥⊥ and
2. ⊤ + ⊤ = ⊤ ⊤.

⊤s and ⊥s together in certain orientations also result in no change: for example,

3. ⊥ + ⊤ = ⊥ ⊤ and
4. ⅂ + ⊥ = ⅂ ⊥.

But when ⊤s and ⊥s meet in other ways, they can interact and undergo change, e.g., a ⊤ can turn into an ⊥ . Transformations depend on whether the open horizontal part of the ⊤ meets with the open or closed horizontal part of the ⊥ . For example,

5. ⊥ + ⊥ = ⊥⊥
6. ⅂ + ⊤ = ⊤ ⊤

These are the simple local rules that limit interactions. In the contexts of [5] and [6], we may say that the ⊥ is metaphorically like an ⊥, and an ⅂ is metaphorically like a ⊤.

Because the molecules are always in thermal motion, the way they happen to meet up is random. Statistically speaking, the production of new ⊥s or new ⊤s is equally likely. One might think that together these reaction scenarios would tend to average out, maintaining a random mixture, but, as Turing found in a similar experiment, instead, differentiation can occur. In our experiment, a clump of, say, ⊥s happens to form in one area, as they might since randomness is not perfectly non-repetitive. No new ⊤s will be produced in an ⊥ clump because a ⊤ is required to produce more ⊤s. Even more ⊥s may be produced at the edges of the clump when ⊥s happen to come in contact with ⊥s in the appropriate orientation.

The clump is self-increasing. No external interference is required. We may say that the material qualities of these ⊥ and ⊥ signs (i.e., the relative similarity and proximity of the signs) lead to the collective activity, an emergent spot pattern. ⊥s can interpret (respond to, interact with, translate) ⊥s and produce more of themselves, more ⊥s.

A soup of this mixture would yield some ⊥ clumps and some ⊤ clumps, floating in random mix of both ⊥s and ⊤s. If ⊥ = black and ⊤ = white, black and white spots will appear on a gray background, as on the coat of an Australian cattle dog. The

actual process forming animal coat patterns is much more complicated, but this serves as a simple visual illustration of spontaneous self-organization that occurs throughout nature, especially in the brain.

In "The Chemical Basis of Morphogenesis" (Turing, 1952a) and "A Diffusion Reaction Theory of Morphogenesis in Plants" (Turing & Wardlaw, 1952), Turing demonstrates that non-linear equations can describe the way patterns form spontaneously from unorganized material. He shows that genes do not need to fully *specify* the complex structure of the organism. The coding genes mainly provide the templates for making the materials, in the right order and in the right amounts, but do not contain the instructions for how to put the materials together.[1] They do not have to. The laws of physics and chemistry and the qualities of the materials (such as that of ⊤s and ⊥s) act as the transformation rules and the constraints that help self-organize the gene-produced materials. As a computer programmer, Turing would have had great admiration for Nature's ingenuity and economy. She did not have to physically record the procedure for development in the DNA. Instead, Nature availed herself of spontaneous self-organizing programs.

Biosemiotic adaptation is possible in this system if, for example, a ⊥⊤ happens to bind with a new molecule, L, *as if it were* an ⊥. (L thus functions as a mistaken sign of ⊥.) All new signs discovered by biological systems must function to an already existing sign-reading system. *They cannot be purely random* as with neo-Darwinian theory. The outcome of an L and ⊥⊤ binding might be a new molecule that will differentially trigger cells affected by this new combination.

Waddington (1940) had provided Turing with the *epigenetic landscape* as a visual metaphor for the physical forces that guide development (or cellular responses generally) which inspired Turing's theory. Waddington had argued that before a cell has differentiated, it is in a state of instability, like a ball sitting atop a mountain with various valley features down below. Turing realized that any slight fluctuation might push it toward one valley pathway or another from this point of instability. These ideas became known as the "catastrophe theory" of early biosemiotician René Thom (see Favareau, 2009: 337-376). Waddington guessed that alternative pathways might be "competing" autocatalytic reactions that used some of the same molecules for different processes. L might bind to ⊥⊤ and trigger one pathway or ⊥ might bind with ⊥⊤ and trigger a different pathway.

The selection process of self-organization is based on the *formal* properties of the elements, qualities, not just the *number* of the elements as with statistical selection. Turing discovered the process whereby differentiating waves, morphogenetic fields, emerge spontaneously without external selection. *This* type of computation is truly self-learning.

1 See Keller (2002) for a history of the understanding of gene action.

Emergent Brain Patterns

Throughout much of the twentieth century, brainwaves were believed to be superfluous, like the sound an engine makes without contributing to the operation of the engine. Now we must consider that these waves may be a type of emergent program for organizing the actions of neurons. In thorough reviews of the literature, Kelso et al. (1991), Uhlhaas et al. (2009) and De Assis (2015) report that many neuroscientists understand the mechanisms underlying working memory and attention in terms of emergent brain waves that synchronize distant neurons, creating virtual neuronal assemblies (De Assis, 2015; Postle, 2006). It appears that waves may provide "the 'contexts' for the 'content' carried by networks of principal cells" and "the precise temporal structure necessary for ensembles of neurons to perform specific functions, including sensory binding and memory formation" (Buzsáki & Chrobak, 1995). In addition, emergent wave patterns may also define what data gets attention, that is, consciousness (see Thompson & Varela, 2001), which, in turn, affects further sensory processing.

This signal propagation theory of learning, using self-organizing signs (not codes), may help explain how people are able to form and use fluid adaptable categories and deal with complex changing environments. Local fluctuations allow stochastic resonance (as with the Ls and ⅃s), the similarity and proximity of possible states, which in turn allows sameness to spread, instant organization. Natural selection cannot "see" to select these local interactions (it does not need to since these interactions just flow spontaneously to the lowest energy state). What can be selected for fitness are the effects of the global patterns that emerge from the local interactions (Cf. Rocha, 1998).[2]

No Artificial Neural Networks or Deep Learning networks are designed to imitate the fluid interplay between self-organization and natural selection. AI designers are more committed to strictly selectionist, aka connectionist, approaches. Although learning can be accomplished this way, it produces automatons, as does standardized curriculums and relentless testing, reward and punishment.

Even with the latest celebrated update (Levis-Kraus, 2016), Google Translate is still bad with puns, jokes and poetry. Psychologists Jung-Beeman et al. (2004) suggest that insight—understanding literary themes and metaphors and getting jokes—requires synchronizing distant brain areas instantly via gamma waves. To design computers that can get allusive language, that understand people, one might need a more fluid medium for traveling waves to emerge. Atomic switch networks as per Stieg et al. (2014) seem promising; they have been used to create emergent patterns that imitate simple natural systems. Experimental chemical

2 Likewise, contrary to the selfish gene hypothesis, natural selection cannot "see" the genes *per se* only their products.

reaction-diffusion computers have been around for more than a decade (Adamatsky et al., 2005), but although they create emergent patterns, they do away with more permanent connections. Our brains seem to use both.

Maybe we will eventually use reaction-diffusion to create more humanoid AI, but we already have eight billion human computers coupled together on the Internet, like so many neurons ready to organize. The potential for spectacular evolution of knowledge is at our finger tips, if only we were in control of AI algorithms rather than controlled by them. With more information about the nature of AI compared to BI, we could make better choices with regard to how little or much we are willing to let AI think for us.

References

Adamatzky, A., Costello, B., & Asai, T. (2005): *Reaction-diffusion computers*. Amsterdam: Elsevier.

Alexander, V. N. & Grimes, V. A. (2017): Fluid biosemiotic mechanisms underlie subconscious habits. *Biosemiotics, 10*(3), 337-353.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016): Machine Bias. *ProPublica*. Retrieved from http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed July 9, 2019)

Copeland, J. (2004): *The essential Turing*. Oxford: Oxford University Press.

De Assis, L. (2015): Neural binding, consciousness, and mental disorder: Complexity as a common element. *Journal for Neurocognitive Research, 57*(3-4), 110-121.

Favareau, D. (2009): *Essential readings in Biosemiotics: Anthology and commentary*. Amsterdam: Springer.

Gilbert, C. & Sigman, M. (2007): Brain states: Top-down influences in sensory processing. *Neuron, 54*(5), 677-696.

Harvey, J. (2013): Sensory perception: Lessons from synesthesia: Using synesthesia to inform the under- standing of sensory perception. *Yale Journal of Biology and Medicine, 86*(2), 203-216.

Hebb, D.O. (1949): *The organization of behavior.* New York: Wiley & Sons.

Jung-Beeman, M., Bowden, E.M., Haberman, J., Frymiare, J. L., Aradmbel-Liu, S., Greenblatt, R., Reber, P. J., & Kounios, J. (2004): Neural activity when people solve verbal problems with insight. *PLOS Biology 2*(4), 500-510.

Keller, E. F. (2002). *The century of the gene*. Cambridge: Harvard UP.

Kelso, J., Bressler, S.L., Buchanan, S., DeGuzman, G.C., Ding, M., Fuchs, A., & Holroydet, T. (1991). A phase transition in human brain and behavior. *Physics Letters A 169*(3), 134-144.

Konkel, F. (2016): The CIA says it can predict social unrest as early as 3 to 5 days out. *Defense One*. Retrieved from http://www.defenseone.com/technology/2016/10/

cia-says-it-can-predict-social-unrest-early-3-5-days-out/132121 (accessed July 9, 2019)

Kwet, M. (2018): Break the hold of digital colonialism, *Mail & Guardian*. 29 Jun 2018. Retrieved from https://mg.co.za/article/2018-06-29-00-break-the-hold-of-digital-colonialism (accessed July 9, 2019)

Le, Q. V., Ranzato, M., Monga, R., & Ng, A. Y. (2013): Building high-level features using large scale unsupervised learning. *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh.

Levi-Kraus, G. (2016): The great A.I. awakening.*New York Times Magazine*. 14 Dec.

Margulis L. & Sagan, D. (2002): *Acquiring genomes: A Theory of the origin of species*. New York: Basic Books.

Noble, D. (2016): *Dance to the tune of life: Biological relativity*. Cambridge: Cambridge University Press.

Postle, B. (2006): Memory as an emergent property of the mind and brain. *Neuroscience 139*(1), 23-38.

Prigogine, I., Nicolis, G. & Babloyantz, A. (1972): Thermodynamics of evolution. The functional order maintained within living systems seems to defy the Second Law; nonequilibrium thermodynamics describes how such systems come to terms with entropy. *Physics Today 25*(11), 23.

Raspopovic, J., Marcon, L., Russo, L, & Sharpe, J. (2014): Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science 345*(6196), 566-570.

Reid, R. G. B. (2007): *Biological emergences: Evolution by natural experiment*. Cambridge: MIT Press.

Rocha, L. M. (1998): Selected self-organization and the semiotics of evolutionary systems. *Evolutionary systems: Biological and epistemological perspectives on selection and self-organization*. S. Salthe, G. Van de Vijver, and M. Delpos (Eds.). Kluwer Academic Publishers; 341-358.

Shapiro, J. (2011): *Evolution: A view from the 21st century*. FT Press.

Sheth, R., Marcon, L., Bastida, M. F., Junco, M., Quintana, L., Dahn, R., Kmita, M., Sharpe, J. & Ros, M. (2012): Hox genes regulate digit patterning by controlling the wavelength of a Turing-type mechanism. *Science 338*(6113), 1476-1480.

Stieg, A., Avizienis, A., Sillin, H., Aguilera, R, Shieh, H.-H., Marin-Olmos, C., Sandouk, E., Aono, M. Gimzewski, J. (2014): Self-organization and emergence of dynamical structures in neuromorphic atomic switch networks. *Memristor Networks*. A. Adamatsky and L, Chua (Eds.) New York and London: Springer; 173-209.

Thompson, D. W. (1942): *On Growth and Form*. Cambridge: Cambridge University Press.

Thompson, E., Varela, F. (2001): Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences 5*(10), 418-425.

Turing, A. (1952a): The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B* 237(641), 37-72.

Turing, A. (1952b): Letter to Norman Routledge. Retrieved from http://www.turingarchive.org/viewer/?id=167&title=1a (accessed July 9, 2019)

Turing, A., Wardlaw, C. (1952): A diffusion reaction theory of morphogenesis in plants. Retrieved from http://www.turingarchive.org/browse.php/C/7 (accessed July 9, 2019)

Uhlhaas, P., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D & Singer, W. (2009): Neural synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience* 3(17), 1-19.

Waddington, C. (1940): *Organisers and Genes*. Cambridge: Cambridge University Press.

Productive Sounds

Touch-Tone Dialing, the Rise of the Call Center Industry and the Politics of Virtual Voice Assistants

Axel Volmar

The growing dissemination of virtual voice assistants in smartphones, smart speakers and vehicle onboard systems, such as Amazon's Alexa, Apple's Siri, Google's Assistant, Microsoft's Cortana or Samsung's Viv, represent a democratization of artificial intelligence by sheer mass exposure.[1] Voice assistants, generally referred to as intelligent virtual assistants (IVAs) or intelligent personal assistants (IPAs), belong to a class of software agents that can answer queries and perform tasks for users based on verbal commands and inquiries when equipped with a voice user interface (VUI). Tech corporations promote their voice-centered smart assistants as pinnacles of contemporary artificial intelligence and as new forms of seamless cooperation between man and machine, built to offer more intuitive ways of controlling and navigating digitally networked and cloud-based technology. The imminent ubiquity of conversational AI, however, raises a number of fundamental questions regarding algorithmic control as well as the nature and history of sound-based human–machine interaction. How are these emergent forms of voice-based cooperation structured and how does voice control change our relationship with and critical assessment of software technology? What ramifications result from AI technologies being based largely on cloud computing and thus from user data being sent to cloud servers to be processed?

Given the black-box character of most commercially available AI technologies, it is naturally rather difficult to obtain detailed information about how the AI algorithms of particular voice assistants exactly function. However, it is not necessary to understand how they work algorithmically in every detail to understand their politics; it is sufficient to study what they are used for and how they are marketed to different stakeholders and actors. I therefore conceptualize intelligent personal assistants—on mobile phones, operating systems, and especially smart

1 A recent report by market analyst firm Canalys (2019) predicts that the worldwide smart speaker install base is set to grow 82.4 per cent from 114 million sold units to over 200 million by the end of 2019.

speakers—as *platforms* in the sense of media scholar Tarleton Gillespie. In his well-received paper, Gillespie argues that the politics of platforms can be traced by examining how

> online content providers such as YouTube are carefully positioning themselves to users, clients, advertisers and policymakers, making strategic claims for what they do and do not do, and how their place in the information landscape should be understood. One term in particular, 'platform', reveals the contours of this discursive work. (Gillespie 2010: 347)

Similarly, I will focus in this paper less on the inner workings of the machines themselves than on the various relations of voice interfaces to their immediate surrounding environment and on the purposes they serve for different actors, such as users, call center agents, businesses, major tech corporations, and surveillance states. However, I will take a considerable historical detour in the effort to ground conversational AI in a broader history of sound- and voice-based human–machine interaction and to emphasize continuities and caesuras between contemporary voice assistants and previous sound- and voice-based user interfaces for networked services. Another reason for this approach is that despite the current hype around voice assistants, auditory and speech-based human–machine interfaces are far from being recent developments. Ever since the psychologically troubled board computer HAL from Stanley Kubrick's *2001: A Space Odyssey* (1968), speech interfaces for human–computer interaction have had a permanent place in the cultural imaginary of industrialized societies.

Although sophisticated artificial intelligence systems like HAL still remain science fiction, sound and speech indeed represent one of the oldest interfaces for interacting with remote systems. However, early applications did not emerge in the computer industry but in the telecommunications sector. Shortly after the release of *2001*, AT&T promoted its Touch-Tone telephones for queries in digital-inquiry/voice-answer (DIVA) systems, which allowed for information retrieval in the form of computer-controlled voice messages through and triggered by Touch-Tone commands. Telephonic practices of interacting with distributed services via sound and speech date back to even the 1940s and 1950s, before they were further developed in the growing call center industry. Contemporary practices of speaking to machines therefore reinterpret forgotten or discarded user experiences connected to the telephone. To this effect, I second media scholar Jonathan Sterne's (2012) emphasis on the centrality of telephony and sound technologies to the history of digitality:

> Telephony is often considered anaesthetic matter in comparison with the usual, more aestheticized subjects of twentieth-century media history such as cinema,

> television, sound recording, radio, print, and computers. But telephony and the peculiar characteristics of its infrastructure are central to the sound of most audio technologies over the past 130-odd years. The institutional and technical protocols of telephony also helped frame the definitions of communication that we still use, the basic idea of information that subtends the whole swath of "algorithmic culture" from packet switching to dvds and games, and the protocols and routines of digital technologies we use every day. (2-3)

While Sterne used the history of the telephone system, and especially developments in signal compression methods and perceptual coding to unpack the mp3 format as a "cultural artifact" (Sterne 2006), I discuss speech-related artificial intelligence applications against the backdrop of a longer history of remote telephone services and processes of (semi-)automation in the telecommunications and customer service industry, with particular attention to call centers. Automation has been a driving force, if not the condition of possibility, of call centers from the very beginning. Most of these attempts are based on what I want to call *productive sounds*, i.e., sounds that serve specific purposes within a (semi-)automated system or even literally perform work, such as triggering switching or algorithmic processes.[2] Productive sounds such as Touch-Tone signals, hold music and recorded voice messages lie at the center of a transformational process in which telephone companies aimed to extend the telephone system from a special-purpose application for voice transmission into a general-purpose information network (cf. Lipartito 2003). Taking the form of synthesized voices in conversational AI and digital personal assistants, sounds became productive as special-purpose substitutes for general-purpose manual tasks previously performed by computer users.

In media theoretical terms, we can understand this transition by conceptualizing productive sound media not as media of *communication* but, in the words of German media theorist Erhard Schüttpelz, as potentially powerful media of *cooperation* (Schüttpelz 2017: 14; cf. Volmar 2017). For instance, to speak of the telephone as a cooperative medium means to conceive it not as a mere conversational medium but as a more universal means to facilitate logistical, bureaucratic, problem-solving, and other quotidian personal tasks of work-related "infrastructuring" (Star/Bowker 2002). At a time when we casually associate such logistical tasks with the internet, online platforms, mobile apps or smart speakers, it seems worth a reminder that the underlying narrative of inter-networked information services is actually older than the internet itself and that it once was deeply entangled with

2 While I use the term "productive sounds" in this specific sense, I take the general notion from Alix Hui and Joeri Bruyninckx who introduced the term at their workshop "Productive Sounds in Everyday Spaces: Sounds at Work in Science, Art, and Industry, 1920–Present" at the Max Planck Institute for the History of Science on April 27-28, 2018.

circuit-switched telecommunications infrastructure. I argue that voice-centered AI applications in call centers (now usually referred to as 'contact centers') and domestic environments can be regarded as a current escalation within the history of cooperative sound media and the various attempts to automate the practices that revolve around them.

Cooperation always entails practices performed by and between different actors and groups. To highlight developments in cooperative practices within the history of voice automation, I pay particular attention to forms of phone- and voice-related work and labor practices. While scholars in the history of media and technology have extensively studied the work of telephone operators (e.g., Green 1995; Lipartito 1994), I follow media and sound scholar Sumanth Gopinath's work on the ringtone industry (Gopinath 2013) by focusing on the significance of sonic and telephonic labor within the infrastructural frameworks of the customer service industry to trace the formation of networked, speech-based human–machine interactions. To this end I examine how changing distributions and delegations of work between call center agents and customers as well as between humans and machines constitute infrastructures of *tele cooperation*, parts of which we also find in current digital assistants.

In section 1, I take a step back to revisit the ramifications of AT&T's introduction of the push-button telephone in the early 1960s. Initially sought to replace operators by further automating the initiation and switching of telephone calls, push-button telephones featured the new dialing method of dual-tone multi-frequency (DTMF) signaling, which operated on the basis of "in-band", i.e., audible control signals—the dial tones we still hear in landline and mobile phones when pushing buttons on the keypad. Sometimes the tones are even simulated on smartphones, for instance within messenger apps. I argue that while multi-frequency signaling rendered telephone switching more automatic and efficient, it also led to practices of delegating and outsourcing phone work from operators to both automatic systems and customers.

More importantly, MF signaling enabled the transmission of sonically coded alpha-numerical information over the telephone network and thus formed a fundamental condition of possibility for the emergence of automatic phone-based information systems in modern call centers. In section 2, I recall some of these technological innovations, especially automatic call distributors (ACDs) and interactive voice response systems (IVRs), both of which were foundational for the rise of the call center industry. I then examine how these contributed to the semi-automation of telephone calls and the further redistribution of voice and sound work by breaking down telephone conversations into common inquiries and sequences and how both call center agents and callers had to adjust themselves to these standardized "boundary objects" (Star/Griesemer 1989) in order to make the automated systems work.

In section 3, then, I show how artificial intelligence entered the stage in the contact center, as it had come to be called, in the form of speech recognition, understanding, and synthesis. I argue that decades of semi-automating phone calls and adjusting agents and customers to automated systems made the contact center particularly receptive to artificial intelligence technology within the industry. The implementation of conversational AI is based on a similar logic as IVRs, as it mainly breaks down phone conversations into a limited number of categories or entities, such as certain key words or presumed emotional states. The same logics are present in contemporary voice assistants for the home. By situating contemporary voice assistants within the broader history of semi-automation and cooperative telephonic practices based on productive sounds and voice work in the call center industry, I ultimately seek to expand existing histories of the internet and digital culture (e.g., Haigh et al. 2015) by considering the evolution of telephone-based telecommunications as an important area for the conception, testing, and mainstreaming of digitally networked media and cooperative practices.

1. Push-button Telephones and Touch-Tone Dialing: Innovation in General-Purpose Infrastructural Technologies

In the first half of the twentieth century, the handling of telephone calls in the Bell System largely remained in the hands of human telephone operators, even though a number of solutions for automatic switching, such as the Strowger switch, were at hand. Whereas technical issues and a reluctance of Bell System managers to license external patents on automatic switching formed the major reasons for clinging to manual switching (Green 1995; Lipartito 1994), opponents of automatic switching argued that establishing the connection represented a form of technical work that should be offered as part of the telephone service and hence done by operators. Harris F. Hopkins, the author of an article in the *Bell Laboratories Record*, put it this way: "Oppositionists felt that automatic switching was wrong from the customer's viewpoint. 'The public will not tolerate doing its own operating,' they said" (Hopkins 1960: 83). After the Second World War, however, rotary-dial telephones to automate the initiation of local phone calls became increasingly common. This transition to self-operating shows that the central logic of automation extended beyond the simple substitution of work by machines to the delegation or redistribution of work in general, in this case from service providers to their customers. The outsourcing of labor to both machines and customers in order to save labor cost, which forms a signature of today's digital culture, was already an economic driving force in the postwar telecommunications sector.

On November 18, 1963, Bell introduced yet another innovation in dialing automation: the push-button telephone, which featured not just a different way of

manual dialing but an entirely new way of creating dialing signals. Dialing on a rotary phone produced a train of electrical impulses, the number of which corresponded to the indicated digit on the rotary dial. Pressing a button on a push-button telephone, however, created a distinct pair of two audible sine tones generated by electronic oscillators. This so-called dual-tone multi-frequency (DTMF) dialing method was based on a four-by-four frequency scheme proposed by L. A. Meacham of Bell's Station Development Department, although initially only seven frequencies (four in the low end of the spectrum and three in the higher range of the spectrum) would generate ten unique pairs of tones (Meacham et al. 1958).[3]

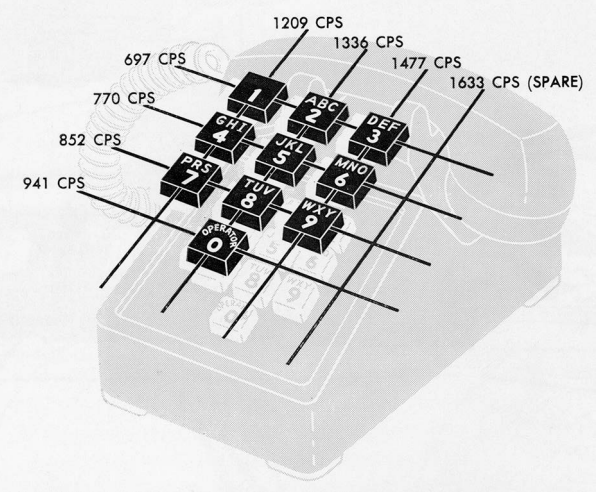


Fig. 1: Four-by-four frequency scheme for the generation of DTMF dialing signals. Image source: Noweck 1961: 314. Courtesy of AT&T Archives and History Center.

The method made use of state-of-the-art solid-state technology and was grounded in a number of field trials conducted between 1948 and 1960 (Dahlbom et al. 1949; Hopkins 1960). When dialing a number, the dual-tones provided a helpful acoustic feedback for the caller. Nevertheless, the sounds were not addressed to human ears to hear in the first place but to electronic filter banks, which were installed at the local switching stations, the so-called "call centers," for decoding. To prevent

3 The pairing of tones followed a simple rule of construction. Each vertical column has a different tone in the low frequency range assigned (FA = 697 Hz, FD = 770 Hz, FC = 852 Hz und FD = 941 Hz), while each horizontal row has a different higher frequency tone assigned (FE = 1209 Hz, FF = 1336 Hz, FG = 1447 Hz und FH = 1633 Hz). This way, each key is assigned to a different combination of a high and a low frequency tone. The necessary hardware comprised a keypad encoder and tone generator.

spoken language, noises and other sounds from interfering with the transmission of DTMF tones, the microphone was disconnected when pressing down a button. Further, because the dialing signals were audible "in-band" frequencies, the Bell technicians chose combinations of frequencies that were unlikely to occur in everyday life so as to prevent false positives and false negatives from occurring in the receivers of the switching equipment: "The frequencies that are used minimize interference from harmonics. This permits instantaneous limiting in both frequency bands, and satisfactorily guards against possible voice interference" (Hopkins 1960: 86). If you ever wondered why push-button tones sound more like the otherworldly noises of electronic music than the harmonious sounds of musical instruments, this is why.

At the 1964 World's Fair in New York, Bell presented DTMF signaling to the public under the brand name Touch-Tone. By means of Touch-Tone calling, subscribers were enabled to initiate, for the first time, long-distance calls directly without the need of a human operator as an intermediary. The introduction of the push-button telephone was therefore closely related to the more or less simultaneous introduction of electronic switching systems (ESSs) to the central switching stations. ESSs were based on digital "stored program control" (SPC), an automated and computerized method of monitoring telephone switching developed around 1954 by Bell Labs mathematician Erna Schneider Hoover (Harr et al. 1964). Electronic switching proved to be more stable and reliable than mechanical methods and eliminated almost entirely the need for human operators. Since tone-based dialing was vital for the introduction of digital switching, the use of sound was also part of a foundational step in the history of digitization. The main advantage of DMTF dialing was the fact that tones could be both generated and detected much faster than the pulse signals generated by rotary phones. The increased speed was particularly helpful for long-distance calls and calls to individual extensions, for instance within larger organizations, since this could greatly increase the number of digits to dial and hence demanded time and patience on the part of the caller. While Bell promoted Touch-Tone dialing to its customers as a more convenient way of initiating calls, the method was particularly tailored to unburden the switching centers, where the old step-by-step switches that could become serious bottlenecks in the connection process, especially during peak calling times. With Touch-Tone signaling, switching centers were able to handle many more calls within a much smaller time span.

The adoption of in-band signaling, however, was not intended to improve the dialing process and the handling of calls alone but to enable new ways of interacting with electronic, and possibly digital, systems connected to the telephone network. As Hopkins (1960) points out, Bell had confidence in offering this "possible future service" because Touch-Tone dialing would provide "the customer with a potential (slow-speed) data transmitter" (87). The first widely distributed

push-button telephone was Western Electric's Model 1500, which came with 10 buttons corresponding to the digits 0 through 9 (see fig. 1). On later models, buttons with the now ubiquitous number (#) and star (*) signs were added to enable and control the transmission of symbolic data. Transmogrified into a potential remote control or terminal device, the telephone receiver could be used to provide alpha-numerical information, such as credit card numbers, or place commands, such as vertical service codes (VSCs). VSCs are sequences of digits in combination with the signals star (*) and, less frequently, number sign (#). Dialled on a telephone keypad or rotary dial, a VSC could be used to enable or disable certain telephony service features, such as call hold, call forwarding, continuous redial or call blocking. The term "vertical" refers to commands pointing to higher-level instructions within the local telephone infrastructure rather than regular telephone numbers, which point out "horizontally" to another geographic location or switching center. AT&T began to introduce VSCs under the name "Custom Local Area Signaling Services" (CLASS or LASS) codes to subscribers in the 1960s and 1970s. With Touch-Tone, sound thus became an acoustic interface for interactions with automated electronic and digital systems.

Seen from the perspective of speech act theory (Austin 1975), the DTMF tones can be conceptualized as "sonic acts" or "sound acts," i.e., as sounds that not only *represent* something or contain information but also *act* and have consequences. As audible control signals, designed to communicate with automated electronic systems over the network, DTMF tones literally became *productive* sounds within the telephone system as they triggered switches, transmitted information and remote-controlled automatic processes. It was on the basis of productive sounds, then, that Bell engineers aimed to prepare the telephone system for the information age. Or put another way, Bell engineers realized that a technology conceived for optimizing their own infrastructure could also be used to develop and offer new information services to both their business and domestic customers. In regard to practice, the growing habit of dialing telephone numbers and using other services, such as VSCs, contributed to training subscribers to perform different forms of *data work*. As noted above, Touch-Tone dialing enabled end-to-end signaling, the transmission of control signals not only to the nearest switching center but also to switching systems anywhere in the network. Therefore, the DTMF method needs to be regarded as an infrastructural medium that played a fundamental role in the transformation of the telephone from a special-purpose technology for talking over distances to a general-purpose technology for speech, data transmission and remote control.

Fig. 2: Different potential applications for banking, retail, or domestic use interconnect customers and digital systems via Touch-Tone telephones. The original caption reads: "Many businesses are using the double-duty TOUCH-TONE® telephone and a computer to speed customer services and develop new ones as well. Banks use the Touch-Tone phone in an information retrieval system known as DIVA (for Digital Inquiry-Voice Answer). With this system, for example, a teller can query the bank 's central computer for a customer's up-to-elate balance before cashing a check (upper left). He dials the computer, taps a few buttons to identify the account number (or, if his phone is a card-dialer model as shown, inserts a DIVA account card) and the code for current balance. The computer responds with a voice answer. Data systems using the Touch-Tone telephone are being used by clerks in retail stores as well. As shown (upper right), the clerk telephones a computer to record each sale she makes. In this case, she sends the account number (for credit sales), the price, merchandise code, and her own clerk number. Billing and accounting are then handled automatically. Eventually, even a house wife (left [image not reproduced here]) may use the Touch-Tone telephone to "shop by phone," pay bills, or check her bank balance." Image source: Soderberg 1969: 203. Courtesy of AT&T Archives and History Center.

2. Speaking to Machines, Speaking in Code: The Rise of the Call Center Industry and the Semi-automation of Phone Conversations

AT&T began to offer new custom calling services based on Touch-Tone dialing in the mid 1960s. These featured new functionalities, such as call waiting, call forwarding, and three-way service or conference calls. Moreover, automatic data collection and information retrieval systems, such as the digital-inquiry/voice-answer (DIVA) system (see the textbox in fig. 2), were sought to bring new forms of distributed cooperation to the business world and domestic subscribers. Bell engineers envisioned diverse workflows of quotidian "infrastructuring" (Star/

Bowker 2002) in a number of different domains, such as banking, retail, and personal use (see fig. 2). For J. H. Soderberg, who summarized some of the potential commercial applications of Touch-Tone-based services in 1969, the switched telephone network pointed the way into the digital future of networked devices and distributed services:

> The possibilities for using the Touch-Tone telephone for control purposes are virtually unlimited. Not only can the Touch-Tone telephone bring the computer revolution into every living room or office across the nation, but it can perform many other simpler control functions. It is even conceivable that future systems will permit you to turn on your home air conditioner so that your home will be comfortable when you return from a trip, or let you "shop by phone"—merely by pushing a few buttons on your telephone. The result could be a dramatic simplification of everyday tasks. (Soderberg 1969: 203)

As Soderberg's vision shows, Bell engineers and marketers had surprisingly clear ideas about the potential of digitally networked, semi-automated services in telephone banking, distributed accounting, home shopping and smart home applications. Not least due to antitrust laws, which banned cross-subsidizing "enhanced" telecommunications services and largely prevented AT&T from venturing into computer businesses, many of these possible applications remained for more than another decade just that, a technological potential and good publicity for the Touch-Tone service. It took well until the 1980s before push button phones reached a considerable saturation.[4] But watch any Hollywood film from the time featuring 1980s yuppie culture and you will see Touch-Tone services everywhere and realize: the telephone system was the internet of services before the internet of services.

Touch-Tone-based services, however, proved tremendously successful in the customer service sector and were deeply connected to the rise of call centers. In the late 1950s and early 1960s, call centers began to form in the offices of telephone companies for their own customer and operator support. Two technical innovations fostered the spread of premise-based call centers. First, the introduction of private automated branch exchanges (PABX), later also referred to as private automated business exchanges, allowed automatic routing to an extension number in a larger organization and hence replaced the work of phone receptionists or attendants (see Bodin 2002: 20). Shortly after, automatic call distributors (ACD) extended PABX capability to collect incoming calls—for instance, to the central

4 The technology was still considered a "premium" feature until well into the 1990s, when personal computers connected to the internet via modems began to challenge the use of the telephone as the go-to interface for interacting with distributed online-services.

office of an organization—and route them to a group of customer service agents.[5] In case all agents were busy, the ACD placed the incoming call in a waiting line until an agent became available. The functionality of ACDs is based on sophisticated algorithms, such as Erlang calculations, for predicting how many agents are needed and how to best queue and assign large numbers of simultaneous calls. ACDs can therefore be seen as the foundation of call centers and represent the first kind of artificial intelligence (in the larger sense of the word), because they introduce automatic decision making to the management of calls. However, ACDs are not artificial intelligence in the narrow sense of the term but rather "conditional call routing solutions, based on if-then conditions, or rules pre-defined by the organization" (Stanley 2018). Nevertheless, ACDs assure to this day that callers are answered as quickly as possible and that the time of all agents is used evenly and effectively.

Both PABXs and ACDs reduced the need for human operators and receptionists in central telephone offices and even rendered their work entirely obsolete. Moreover, sophisticated ACDs provided reports on various aspects of the call transaction (Bodin 2002: 22-23). Automatic call distributors proved particularly valuable for organizations that faced large call volumes. However, automatic in-house routing had the obvious disadvantage, due to algorithmic procedures, of not allowing callers to contact an agent directly. Since callers were unlikely to get assigned to the same agent twice, it prevented them from forming relationships with particular agents and hence resulted in a much less personal calling experience. AT&T's introduction of toll-free 1-800 numbers in 1967 basically established automatic call distribution on a nationwide scale—the service would first redirect calls to a national or local call center, where on-premise ACDs would further route the call to available agents.[6] Toll-free numbers led to an unprecedented increase in customer service call volume and cemented the anonymous user experience as a *de facto* standard. ACDs became the foundation of large-scale, decentralized and geographically distributed call centers. Among the early ACD solutions that proved economically successful, the US-manufacturer Rockwell is one of the most credited. The company's Galaxy ACD, as the device was called, enabled Continental Airlines to start offering phone-based flight reservation in 1973.

In the 1970s, the potential of DTMF signaling was recognized by manufacturers of call center equipment. So-called interactive voice response (IVR) systems automated not only the routing of calls but also specific parts of the actual phone conversations themselves. ACDs could play welcome messages, but they

5 The job of automatic call distributors, or ACDs, is to filter, order and assign incoming calls to the best available agent.

6 The inventor of the toll-free number once stated that all he had invented was in fact a pointer in a digital directory.

featured no further functionality other than putting the caller on hold. In IVRs, prerecorded messages would inquire about the caller's needs, acoustically guide them through a menu structure and present them with choices for different services, which the caller would then be able to select by pressing the corresponding buttons on a push-button phone. Typically, these systems were semi-automated human–machine systems with IVRs at the front end and human agents who took over at predefined points or whenever an automated system would come up against limits. The division of labor between humans and machines was achieved by breaking down phone conversations into parts with greater or lesser degrees of redundancy and automating the former. Fixed sets of categories and options addressed most customer queries, delivered through prerecorded messages that caller callers could respond to using DTMF tones. We can therefore regard the relation between the customer and a respective organization, which unfolds within an IVR system, as what Susan Leigh Star and James R. Griesemer have called a "cooperation without consensus" based on a common techno-conversational "boundary object" (1989).

The self-service functionality of IVRs allowed for substituting, at least in part, not only operator work but also the actual voice and transactional work performed by customer service agents. Other than the obvious saving of labor cost, automatic call center systems had the advantage of enabling expanded service hours. The flipside, however, was that since callers were not even talking to human agents anymore—at least not until the system connected them to one—IVRs rendered the phone experience even more anonymous than the seemingly random selection process done by automatic call distributors. Over the years, vendors added voice recognition to Touch-Tone as an alternative input language. The primary goal of introducing voice control had been to extend IVR services to owners of rotary-dial telephones but the result was that with voice recognition, whoever preferred speaking to typing was now able to interact with the IVR system via spoken language. This is the point where AI techniques first enter the stage.

Most of these circuit-switched telephonic systems have since been replaced by packet-switched, IP-based technology. Their story is therefore, at least to some extent, also an archaeology or reconstruction of media-cultural visions of a semi-automated future, consisting of human operators and interactive systems. They also refer to a hybrid future of cooperative systems that were both analog and digital at the same time. George Lucas' first feature film, *THX 1138* (1971), is exemplary of the future visions in this period of telephonic information networks. Lucas paints a picture of a futuristic underground society permeated by communication and surveillance technologies, reminiscent of George Orwell's novel *1984*. He thereby extrapolates contemporaneous advancements in touch-button telephones and semi-automatic systems into a dystopia of total audiovisual mediatization and surveillance. The impression of the omnipresence of media-technological media-

tion and observation is further reinforced by frequently staging technically mediated communication situations in the form of telephone and intercom conversations, tape announcements, video transmissions, and CCTV images.

As a response to its cultural moment, *THX 1138* forms an artistic reflection on the then-incipient transformation of acoustic media into what Jonathan Sterne has termed a "speaker culture" (Sterne 2015: 113). The film's soundscape of technical communications and automated announcements, interwoven through montage, raises the question of whether the characters actually interact with human interlocutors or merely with automatically triggered answers stored on tape. Its references to telephone technology are inscribed further in its very scene design, with a Pacific Bell circuit switch room serving as a filming location, according to the IMDB trivia section:

> The seemingly endless Control Room where the android police try to corner THX and SRT, who find out LUH has been consumed for organ reclamation, was the circuit switch room of the San Francisco location of the Pacific Bell Telephone Company. Pacific Bell allowed George Lucas to shoot the film there, because the entire room and the hardware found there were about to be dismantled, as the phone company was switching to touchtone phone technology (IMDB 2019).

Lucas even named the title of the film after his San Francisco telephone number, 849-1138, where the letters THX correspond to letters found on the buttons for the digits 8, 4, and 9. Moreover, many of the electro-acoustic sound effects that populate the soundscape of the film are distilled from telephone dial tones, which editor and sound editor Walter Murch manipulated by applying compositional methods derived from musique concrete. The depiction of automatic speech systems as inhumane and anonymous is achieved largely by recreating or mimicking the user experience of early IVR systems: the messages and public announcements that are automatically triggered throughout the movie are repetitive and monotonous and leave no room for doubt that the citizens of the future society have to adjust to the system and not the other way around. Rewatching the movie almost half a century after its initial release, one cannot help but associate it with current AI-based public surveillance systems, such as China's Social Credit System.

3. "Speech is an Untapped Goldmine": The Adoption of AI in the Contact Center and Virtual Voice Assistants

Despite their still apparent limitations, recent speech recognition and synthesis systems, such as those used for voice assistants, sound more familiar and less robotic and anonymous than the mantra-like reminders and announcements that

populate the soundtrack in *THX 1138*. Early examples of automatic speech recognition (ASR) include pattern-based models for detecting a limited ensemble of spoken sounds such as digits and words, where the recognition of an uttered digit or word is determined by its correlation with a set of stored reference patterns (Davis et al. 1952: 194). Among the well-known early examples of such applications, Bell Laboratories' "Audrey" (Pieraccini 2012: 55-59) and IBM's "Shoebox" (Dersch 1962) were able to recognize spoken digits and, in the case of Shoebox, even a limited number of commands if spoken by a familiar voice.[7] "HARPY," a speech recognizer developed in the mid 1970s at Carnegie Mellon University as part of the first ARPA project on speech understanding research, was already able to recognize a vocabulary of 1,011 words (Lowerre 1976). In the late 1970s, IBM's Dragon system heralded a new era of ASR systems based on hidden Markov models, the descendants of which were used in most IVR systems from the 1990s onward (Pieraccini 2012).

As noted in the previous section, speech recognition research yielded the potential use of the human voice to control automated systems and to transmit information to them. Spoken language thus represented an alternative type of productive sound alongside DTMF tones in automated telephone systems. Moreover, the integration of voice control into major computer operating systems such as Windows or MacOS, not least in order to increase accessibility for visually impaired users, points toward the conversational systems that we now see used in current applications and platforms for smartphones and smart home devices. Today, the combination of automatic speech recognition, understanding and synthesis—now largely based on artificial intelligence approaches—is referred to as natural language processing. A crucial step toward this stage of extended voice agent interaction has been the application of machine learning and deep learning techniques, which mostly rely on learning algorithms based on deep neural networks (DNN). Compared with previous methods, following from the historical precursors in speech recognition and synthesis described above, DNNs allow for the analysis and processing of voice audio with a much higher level of accuracy and naturalness (cf. Mary 2018: 50). The improvements are primarily due to the general increase in processing power, the use of cloud computing, and the access to vast amounts of training data. This is also the reason why big tech companies have in recent years increasingly developed natural language processing and offered AI solutions for call centers and voice assistants for smartphone or home use.[8] Special apps and platforms, such as Amazon's Lex, Google's Dialog-

7 The name "Audrey" is a loose acronym of "automatic digit recognition."

8 These systems are increasingly based on a centralized internet infrastructure dominated by cloud-based services provided by a few major market leaders, Amazon (AWS), Google (Google Cloud), and Microsoft (Azure). The speech recognition models, the emotion analysis metrics, and

flow, Facebook's Wit.ai, IBM's Watson, and Microsoft's LUIS, offer considerably straight-forward solutions for creating conversational bots. Not surprisingly, one of the major professional domains of AI application is the contact center industry. Google, for instance, boasts that its cloud services provide "AI-powered virtual agents for the contact center, including phone-based conversational agents known as interactive voice response (IVR)" (Google 2019).

Call centers offer ideal conditions for the introduction of voice-centered AI technologies because they constitute, as shown in section 2, highly compartmentalized, process-oriented and automated conversational environments with a long history of human–machine integration. From the beginning, developers and vendors of IVRs have conceived systems to which both customers and agents must adapt. Now with the most recent examples of virtual assistants, we can observe this logic upheld and transformed into new semi-automated conversational settings: to ensure that the systems "understand" them, users need to adapt the way they talk and the words they use. Therefore, the processes of automating telephone work through IVRs and conversational AI can be better described in terms of what Hamid R. Ekbia and Bonnie A. Nardi (2017) have termed "heteromation," the "extraction of economic value from low-cost or free labor in computer-mediated networks"—in this case, labor performed by both call center agents and customers.

The contemporary contact center's core function does not fundamentally differ from its original, historical task of handling inquiries and improving customer satisfaction. However, the increase in online shopping and other forms of e-commerce has brought along a huge demand for virtual customer services, which coincides with the significant advancement in natural language processing capabilities and synthetic speech models over the past decade (Kopparapu 2015: 5). The use of these optimized systems promises the automation of not only call distribution and routing but also more individual customer interactions, such as complex three-factor account authentification, with the effect of further reducing the need for direct contact with human service agents—possibly until eventually conversations between humans will have shifted from the norm to the exception.[9] Since their emergence, IVR systems, which require long automated spoken menus for their extensive decision trees, have incurred criticism for being impersonal and annoying (cf. Smith 2016). A second reason for integrating intelligent personal as-

the design of synthetic voices that lie at the core of contemporary and future autonomous conversational agents are, thus, all dependent on the protocols and regulations determined by these corporates.

9 A parallel development has happened in the field of text-based chatbots, the performance of which is now convincing in standard use cases (Sheth et al. 2019), although as of now, most customer interactions still happen over the phone.

sistants is therefore to offer the customer the experience of a "personal," seemingly individual conversational behavior, with the goal of overcoming the perceived shortcomings of IVR systems, by hiding the underlying hierarchical structure from the perception of the customer. Apple's Siri, for example, has been branded from the start as a witty and fun-to-use application with personality to dissociate it from anonymous automated systems, such as IVRs.[10]

The use of automatic speech recognition to augment traditional IVR systems and replace human agents, however, is not the main purpose for introducing AI technology in the contact center. Rather, it is the tip of the AI iceberg that is generally visible or perceivable to the customer. In the contemporary contact center, we are very likely to find not only one but a growing number of different types of artificial intelligence solutions at work simultaneously. According to one of the industry's leading trade magazines, *Call Center Helper*, artificial intelligence solutions are used not only for the handling of calls but increasingly for the production of new insights about customers and call center agents by capturing data from customer interactions, applying big data analytics, predicting customer behavior or monitoring advisor performance (Call Center Helper 2018). An industry representative hence predicts that "our future with machines is going to be (and needs to be) one of partnership and enhancement, not sweeping replacement" (Call Center Helper 2019). Call centers usually have vast amounts of stored voice recordings at their disposal, which make them particularly suited for analytic AI applications, especially predictive analytics and speech analytics. As an industry white paper frames it, "Speech is an untapped goldmine." (CallMiner 2019: 5)

Predictive analytics allows call centers to generate valuable insights in real-time, such as a customer's willingness to pay off a debt, a customer service agent's effectiveness at addressing particular concerns, and a caller's overall sentiment and the actions likely to satisfy them given their history. Speech analytics, in turn,

> goes beyond recognition, interpreting not just the words a caller speaks but also the manner in which those words are spoken. [Also known as voice analytics, this technology] detects factors such as tone, sentiment, vocabulary, silent pauses, and even the caller's age, analyzing these factors to route callers to the ideal agent based on agents' success rates, specialized knowledge and strengths, as well as the customer's personality and other behavioral characteristics. (Stanley 2018, n.p.)

In particular, this concerns the backtracking of all available voice recordings for all sorts of analyses and the ambition to detect and analyze not only the semantic but also the emotional aspects of the human voice by exploiting methods of affective computing (Picard 1997; Jeon 2017). The bottom line of the current shift

10 "Siri" stands for "speech interpretation and recognition interface."

toward the integration of AI technology into the contact center is that it works only in part for the customer and primarily for the enterprise using it. What I want to argue is that the same goes for virtual voice assistants for the home, for which contact centers served as a testing ground for early adoption (Davis 2019). To this effect, the various uses of AI in the customer service industry point to a number of potentially invisible or concealed uses of AI for domestic voice assistants. Smart speakers with voice interfaces are branded as convenient interfaces to both local and cloud-based digital services. They are thus designed to simulate personality in order to be more fun to use. We should, however, not trick ourselves into thinking we are dealing with one and only one artificial intelligence alone—the workings of which are represented and condensed in the form of the artificial voice. Rather, we should realize that there are probably a dozen other AI systems listening in and analyzing the information of our voice data being transmitted to the providers' cloud servers. In the end, intelligent personal assistants work not merely *for* the users but *on* them. Domestic users and office workers embrace voice assistants for their convenience and efficiency in performing repetitive tasks such as web searches and daily routines. Businesses, tech corporations, surveillance states, and other actors, however, are competing to gain access to the users' voice itself, which is seen as a highly valuable data source—a "goldmine"—for AI-based analytics.

4. Conclusion

With the introduction of DTMF signaling in the 1960s, special-purpose telephone receivers were repurposed into general-purpose remote controls, resulting in a fundamental first step toward a long-ranging transformation of the telephone system from a mere medium of communication into a versatile medium of cooperation. Over the course of roughly two to three decades, Touch-Tone calling in conjunction with IVR systems slowly trained users in how to interact with remote automatic and semi-automatic information systems over the telephone network. Given that these technologies almost exclusively relied on all-acoustic interfaces and a small keypad, we can consider the mobilization of productive sounds to have ultimately paved the way for what could retroactively be called the first generation of everyday "online practices." Different iterations of productive sounds, as I have argued, in this way formed the basis of a slow transition from telecommunication to telecooperation: at first, in an essential and operational sense in the form of multifrequency signals; later, as voice work performed by call center agents, prerecorded messages, hold music and other design elements, which formed part of telephonic waiting loops and acoustic interfaces in automat-

ed interactive voice response systems; and finally, as conversational AI systems based on natural language processing.

As the introduction of Touch-Tone calling has shown, already the "old" media industries, especially the telecommunications sector, worked toward realizing a future based on networked information technologies and services. The automation of customer service calls revealed how infrastructural innovations laid the foundation for the emergence of new services based on both electronic and embodied "data practices" and how these transformations occurred in circuit-switched telephone networks before the growth of personal computers and the internet and well outside the computer industry. By tracing the relations between different technological agents and forms of labor within the cooperative assemblages of call centers, I have shown that the development of voice-related artificial intelligence systems should be seen as part of a larger history of human–machine interaction, the practices of which continue to shape the relations between users and contemporary voice assistants. This transformation occured not so much in the form of a disruptive revolution but in terms of historical continuities based on successive combinations and recombinations of (semi-)automatic man–machine systems and the sedulous infrastructuring, networking, and delegation of co-operative practices, ultimately leading to virtual call center agents and domestic voice assistants.

The use of voice assistants and smart speakers is reminiscent of the principles and practices of using a self-service call center system. Therefore, I tentatively like to frame them as call centers for the home. Moreover, in the coming years, voice control, especially in hands-free environments such as moving vehicles, is likely to become a ubiquitous and naturalized interface practice. In the contemporary contact center, managing and automating conversations to reduce labor cost and enhance efficiency is not the only motivation for embracing artificial intelligence solutions anymore; equally important is the analysis of user data for making predictions and producing new commercially exploitable insights. AI in virtual voice assistants is therefore used not only to create new ways of conveniently controlling our everyday tasks but also to data mine the control signals (i.e., the voice input) as exploitable customer data. Studying call center practices can therefore be a way to understand voice assistants, and their politics might thus best be explained by an uncanny pact of co-operation: One the one hand, voice assistants are devised to help us, and they do it well and will even get better as their skills improve. On the other hand, because virtual voice assistants transmit our digitized voice signals to remote cloud servers for processing, users are, metaphorically speaking, inviting into their homes and feeding nameless background AI routines with every conversation. The most common prerecorded pronouncement in call center systems is equally valid for virtual voice assistants: "Your call will be monitored."

Acknowledgements

This research has been funded by the German Research Foundation (DFG) as part of the A01 project of the Collaborative Research Center 1187 "Media of Cooperation" (Medien der Kooperation). I would like to express my gratitude to Kyle Stine for critical remarks and suggestions. I would like to express my gratitude to Kyle Stine for critical remarks and suggestions and Thomas Bjørnsten for valuable input to the paper. I would also like to thank Sheldon H. Hochheiser and Melissa Wasson of the AT&T Archives and History Center (Warren, NJ) for their generous support.

Bibliography

Aharon, Dan/Laqab, Daryush (2018): "Transforming the Contact Center with AI." Google Cloud Blog. https://cloud.google.com/blog/products/gcp/transforming-the-contact-center-with-ai/ (June 10, 2019).

Austin, John Langshaw (1975): How to Do Things with Words. Oxford: Oxford University Press.

Bodin, Madeline (2002): The Call Center Dictionary: The Complete Guide to Call Center and Customer Support Technology Solutions. Boca Raton: CRC Press.

Call Centre Helper (2018): "12 Top Uses of Artificial Intelligence in the Contact Centre." Call Centre Helper. https://www.callcentrehelper.com/12-top-uses-of-artificial-intelligence-in-the-contact-centre-123361.htm (June 10, 2019).

Call Center Helper (2019): "Artificial Intelligence in the Contact Centre: What You Should REALLY Know." Call Centre Helper. https://www.callcentrehelper.com/artificial-intelligence-contact-centre-should-know-142841.htm (June 10, 2019).

CallMiner Eureka (2019): "How AI Improves the Customer Experience. Real Use Cases of Engagement Analytics & Automation for Contact Center Success." CallMiner. https://learn.callminer.com/whitepapers/how-ai-improves-the-customer-experience (June 10, 2019).

Canalys (2019): "Canalys: Global Smart Speaker Installed Base to Top 200 Million by End of 2019." https://www.canalys.com/newsroom/canalys-global-smart-speaker-installed-base-to-top-200-million-by-end-of-2019 (June 10, 2019).

Dahlbom, C. A./Horton, Jr., A. W./Moody, D. L. (1949): "Applications of Multifrequency Pulsing in Switching." In: Trans. AIEE 68, pp. 392-96.

Davis, Jessica (2019): "Voice Assistants Coming to the Enterprise." InformationWeek. https://www.informationweek.com/strategic-cio/it-strategy/voice-assistants-coming-to-the-enterprise/d/d-id/1333642 (June 10, 2019).

Davis, K. H./Biddulph, R./Balashek, S. (1952): "Automatic Recognition of Spoken Digits." In: Journal of the Acoustical Society of America 24/6, pp. 637-642.

Dersch, W. C. (1962): "Shoebox: A Voice Responsive Machine." In: Datamation 8/6, pp. 47-50.

Ekbia, Hamid R./Nardi, Bonnie A. (2017): Heteromation, and Other Stories of Computing and Capitalism. Cambridge, Mass.: MIT Press.

Gillespie, Tarleton (2010): "The Politics of 'Platforms.'" In: New Media & Society 12/3, pp. 347-364.

Gopinath, Sumanth (2013): The Ringtone Dialectic: Economy and Cultural Form. Cambridge, Mass.: The MIT Press.

Green, Venus (1995): "Goodbye Central: Automation and the Decline of 'Personal Service' in the Bell System, 1878-1921." In: Technology and Culture 36/4, pp. 912-949.

Haigh, Thomas/Russell, Andrew L./Dutton, William H. (2015): "Histories of the Internet: Introducing a Special Issue of Information & Culture." In: Information & Culture 50/2, pp. 143-159.

Harr, J. A., E. S. Hoover/Smith, R. B. (1964): "Organization of the No. 1 Ess Stored Program." In: Bell System Technical Journal 43/5, pp. 1923-1959.

Hopkins, Harris F. (1960): "Push Button 'Dialing.'" In: Bell Laboratories Record 38/3, pp. 82-87.

IMDb (2019) "THX 1138 (1971)—Trivia." IMDb.com. https://www.imdb.com/title/tt0066434/trivia (June 10, 2019).

Jeon, Myounghoon (ed.) (2017): Emotions and Affect in Human Factors and Human-Computer Interaction. London; San Diego, Cal.: Elsevier/Academic Press.

Kopparapu, Sunil Kumar (2015): Non-Linguistic Analysis of Call Center Conversations. New York: Springer.

Lipartito, Kenneth (2003): "Picturephone and the Information Age: The Social Meaning of Failure." In: Technology and Culture 44/1, pp. 50-81.

Lowerre, Bruce T. (1976): "The HARPY Speech Recognition System." PhD Dissertation. Carnegie-Mellon University.

Mary, Leena (2018): Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition. New York: Springer.

Meacham, L. A./Power, J. R./West, F. (1958): "Tone Ringing and Pushbutton Calling: Two Integrated Exploratory Developments." In: Bell System Technical Journal 37/2, pp. 339-360.

Nexidia Interaction Analytics (2017): "AI-Powered Analytics Transform the Enterprise." www.nexidia.com/media/2522/whitepaper-using-ai-powered-analytics-jan-2017.pdf (June 10, 2019).

Noweck, H. E. (1961): "The Versatility of Touch-Tone Calling." In: Bell Laboratories Record 39/9, pp. 312-316.

Perez, Sarah (2019): "Over a Quarter of US Adults Now Own a Smart Speaker, Typically an Amazon Echo." TechCrunch. http://social.techcrunch.com/2019/03/08/over-a-quarter-of-u-s-adults-now-own-a-smart-speaker-typically-an-amazon-echo/ (June 13, 2019).

Picard, Rosalind W. (1997): Affective Computing. Cambridge, Mass.: The MIT Press.

Pieraccini, Roberto (2012): The Voice in the Machine: Building Computers That Understand Speech. Cambridge, Mass.: MIT Press.

Schüttpelz, Erhard (2017): "Infrastructural Media and Public Media." In: Media in Action 1/1, pp. 13-61.

Sheth, Amit/Yip, Hong Yung/Iyengar, Arun/Tepper, Paul (2019): "Cognitive Services and Intelligent Chatbots: Current Perspectives and Special Issue Introduction." In: IEEE Internet Computing 23/2, pp. 6-12.

Smith, Ernie (2016): "The History of the Call Center Explains How Customer Service Got So Annoying." Vice. https://www.vice.com/en_us/article/xyg4mn/the-history-of-the-call-center-explains-how-customer-service-got-so-annoying (June 10, 2019).

Soderberg, J. H. (1969): "Machines at Your Fingertips." In: Bell Laboratories Record A 47/7, pp. 199-203.

Stanley, Robert (2018): "A Comprehensive History of AI in the Call Center: From ACDs to Predictive Analytics and Beyond." CallMiner. https://callminer.com/blog/comprehensive-history-ai-call-center-acds-predictive-analytics-beyond/ (June 10, 2019).

Star, Susan Leigh/Bowker, Geoffrey C. (2002): "How to Infrastructure." In: Leah A. Lievrouw/Sonia Livingstone (eds.), Handbook of New Media: Social Shaping and Social Consequences of Icts, London: Sage, pp. 151-162.

Star, Susan Leigh/Griesemer, James R. (1989): "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science 19/3, pp. 387-420.

Sterne, Jonathan (2006): "The Mp3 as Cultural Artifact." In: New Media & Society 8/5, pp. 825-842.

Sterne, Jonathan (2012): MP3: The Meaning of a Format. Durham: Duke University Press.

Sterne, Jonathan (2015): "Space within Space: Artificial Reverb and the Detachable Echo." In: Grey Room 60, pp. 110-31.

Volmar, Axel (2017): "Formats as Media of Cooperation." In: Media in Action 1/2, pp. 9-28.

Algorithmic Trading, Artificial Intelligence and the Politics of Cognition

Armin Beverungen

In this chapter I focus on the changes in algorithmic trading in financial markets brought about by developments in machine learning and artificial intelligence (AI). Financial trading has for a long time been dominated by highly sophisticated forms of data processing and computation in the dominance of the "quants". Yet over the last two decades high-frequency trading (HFT), as a form of automated, algorithmic trading focused on speed and volume rather than smartness, has dominated the arms race in financial markets. I want to suggest that machine learning and AI are today changing the cognitive parameters of this arms race, shifting the boundaries between "dumb" algorithms in high-frequency trading (HFT) and "smart" algorithms in other forms of algorithmic trading. Whereas HFT is largely focused on data internal and dynamics endemic to financial markets, new forms of algorithmic trading enabled by AI are enlarging the ecology of financial markets through ways in which automated trading draws on a wider set of data such as social data for analytics such as sentiment analysis. I want to suggest that to understand the politics of these shifts it is insightful to focus on cognition as a battleground in financial markets, with AI and machine learning leading to a further redistribution and new temporalities of cognition. A politics of cognition must grapple with the opacities and temporalities of algorithmic trading in financial markets, which constitute limits to the democratization of finance as well as its social regulation.

Consciousness and Capitalism

Financial markets arguably are at the forefront of a battle around cognition in contemporary capitalism. If capitalism today is marked both by the way in which finance serves as a primary means to exert violence on and to extract value from life, and by the way in which capital amasses and appropriates cognitive capacities to sustain this extraction (Fumagalli/Mezzadra 2010), then financial markets

are bound to play a key role in this financial and cognitive capitalism (Beverungen 2018). Financial markets might appear then as a "collective capitalist brain" through which capital cognitively organizes the extraction of value, only hampered by "occasional and random catastrophe" associated with high-frequency trading (HFT) (Terranova 2013: 66). And HFT might be understood as the "high frontier of cybernetic innovation" in a war of capital against its enemies and the working class in which computers are "weapons wielded by advanced capital" (Dyer-Witheford 2016: 51, 35). While a closer look at artificial intelligence (AI) and algorithmic trading will yield a complex picture in which a collective capitalist brain is far from perceptible, and class war is perhaps less visible than competition between individual capitals, this is an important frame of analysis to be kept in mind as my analysis proceeds.

That finance is concerned with the extraction of value can however be taken for granted. That premise is also apparent from the perspective of the financial trader, where the question of how to extract value from financial markets becomes one of making the right trade. As Beunza and Stark argue, "What counts?" is the question which "expresses most succinctly the challenge facing securities traders in the era of quantitative finance" (2008: 253), and presumably all other financial traders as well, including high-frequency traders. The task is primarily one of information, with traders "immersed in a virtual flood of information", where "the challenge for traders is not faster, higher, stronger—as if the problem of the volume of data could be solved by gathering yet more—but selecting what counts and making sense of the selection" (Beunza/Stark 2008: 253). The "calculative practices" that traders deploy to respond to the question of "what counts?" are "distributed across persons and instruments" (Beunza/Stark 2008: 254). Beunza and Stark here presume a certain problem of information, where the task is to select relevant information that can be made to count in financial trading which yields a surplus. We can see already how AI and its key advance today—artificial neural networks—may be very helpful.[1] Below I will explore how different types of AI have been deployed in financial trading, and note how these have shifted the parameters of the challenge Beunza and Stark describe.

Before I proceed, though, I would like to historicize Beunza and Stark's premises and expand on their analysis, which will also allow me later to come back to the more abstract political analysis of financial and cognitive capitalism. First of all, it is important to note that Beunza and Stark's market characterize as being characterized by information flows and by the cognitive challenge of filtering

1 Somewhat amusingly, the futurist and "world leader in pattern recognition techniques" Ray Kurzweil has since 1999 operated a company called "FatKat" which builds "industry-leading tools for quantitatively based investing". Little is known about this company, and its website (www.fatkat.com) is still dated 2001. See Patterson 2012: 306.

information in order to yield information useful for a successful trade, is not a historical given. It took a while for the market to be understood as an information processor and for it to be designed to that end. Mirowski and Nik-Khah (2017) have extensively explored the influence of Hayek's evolving conceptions of markets on the discipline of economics and the practice of market design. They identify at least three stages in Hayek's work in which markets, knowledge and information are understood differently (Nik-Khah/Mirowski 2019: 38-44). First, knowledge is understood as something hard to amass, a task only markets can achieve. Then knowledge becomes something tacit, and therefore something only markets can bring to the fore. Finally, knowledge becomes information, something suprapersonal, residing within the market: "a new virtual kind of *information*" (Nik-Khah/Mirowski 2019: 43; emphasis in original).

Nik-Khah and Mirowski demonstrate that these conceptions of markets have influenced different schools of market design, in which economists act as engineers of markets, such as financial markets. I will note below how market designers such as Alvin Roth and others are involved in designing the financial markets in which algorithmic trading takes place and AI is deployed. The important aspect to note, coinciding with the way market design "constitutes the precepts of neoliberalism taken to their logical conclusion" (Nik-Khah/Mirowski 2019: 63), is the way in which human consciousness and cognition become increasingly irrelevant to markets and are ultimately discounted, with a market conceived as a "person-machine system", a "hybrid computational device", "with the thinking offloaded onto things" (Nik-Khah/Mirowski 2019: 53, 61). As Mirowski and Nik-Khah put it: "Agents would be folded into the person-machine system, no longer deemed capable of understanding why they made the decisions that they do. Think of their predicament as Artificial Ignorance." (2017: 238-239). It might seem ironic that this "artificial ignorance" also of financial traders is to be complemented by the artificial intelligence of machines. But as I will show below, the deployment of AI in algorithmic trading exactly follows the premises of the economists and market designers: information resides in the market, and the task of AI is to extract it—the "alpha"—in order to augment trading.

To get a handle on how markets are constituted both as human and as machinic and computational, and on how thinking is "offloaded onto things", I want to draw on Hayles' recent work around nonconscious cognition (Amoore 2019), as it offers a helpful way of making sense of how cognition is distributed in financial markets (see also Beverungen/Lange 2018). Hayles distinguishes between "thinking" and "cognition", suggesting that thinking is human, conscious cognition whereas cognition "is a much broader faculty present to some degree in all biological life-forms and many technical systems" (2017: 14). She defines cognition as "a process that interprets information within contexts that connect it with meaning" (Hayles 2017: 22), one that can also take place nonconsciously. She offers to replace

the distinction between human and nonhuman with the distinction "cognizers" and "noncognizers", where humans, biological life but also technical systems such as those deploying AI are part of the first category (Hayles 2017: 30). Importantly, this allows us to understand the make-up of financial markets as constituted by a number of cognizers (both human and machinic), to consider how cognition is distributed between these cognizers (both as conscious and nonconscious cognition), and to explore what kinds of autonomy is given to machines in algorithmic trading in "pockets within which technical systems operate autonomously" in a "punctuated agency" (Hayles 2017: 32).

Hayles (2017: 142-177) offers her own analysis of finance and HFT, and suggests that HFT may be "regarded as an evolutionary milieu in which speed, rather than consciousness, has become a weapon in the nonconscious cognitive arms race—a weapon that threatens to proceed along an autonomous trajectory in a temporal regime inaccessible to direct conscious intervention" (2017: 165). In the following sections, I want to build on Hayles and on earlier work with Lange (Beverungen/Lange 2017; 2018) to explore how this "nonconscious cognitive arms race" is shaped by AI. I will suggest that AI offers a different weapon—smartness—in a trade-off with speed in this race, one which shifts the temporal and cognitive parameters of financial markets, which can be made further sense of if discussions around financial and cognitive capitalism are kept in mind.

High-Frequency and Quantitative Trading

Prior to the automation of trading platforms of financial markets, the "cacophony of the marketplace and apparent randomness of trade" was coordinated mostly through human sociality; today, that is a matter of "managing the punctuated electronic signals that encode the orders from masses of anonymous investors", achieved by "toying with the nimble algorithms, sophisticated computer processors, hacked routers, and specialized telecommunication systems that are the material foundations of the contemporary stock exchange" (Pardo-Guerra 2019: 23). Manual trading still exists, although all orders have to be executed via automated platforms, and algorithmic trading constitutes the large majority of trading in financial markets. Kirilenko and Lo define algorithmic trading as "the use of mathematical models, computers, and telecommunications networks to automate the buying and selling of financial securities" (2013: 52). Over the last two and a half decades, its rise has been facilitated by the ways in which the financial system has become more complex, by "a set of breakthroughs in the quantitative modeling of financial markets", and by the "almost parallel set of breakthroughs in computer technology" (Kirilenko/Lo 2013: 53). Markets have been automated, trading strategies are computer-driven, and trade is executed largely by algorithms.

Financial markets, even before the introduction of automated trading platforms, offered opportunities for trading strategies based both on speed and on smartness, and implied certain forms of cognition. The introduction of the ticker tape, as discussed by Preda (2006), for example, changed the temporal regime of the stock market, offering a continuous data flow of price variations, for all means and purposes in real time: the "ragged time structure of paper slips was replaced by the smooth, uninterrupted, unique time of the ticker tape" (Preda 2006: 767). The ticker tape also came with charts and other forms of visualisation, as well as "discursive modes" which "supported the chart as a cognitive instrument, which in its turn conferred authority upon the stock analyst as the only one skilled enough to discover the truth of the market in the dotted lines" (Preda 2006: 770). The speed of the ticker tape alone did not lead to a competitive advantage; the smartness of the stock analyst was required to access the truth of the market and to act on it. This economy of speed and smartness would develop further, for example with the introduction of the Reuters Stockmaster price retrieval service in 1964 or the launch of the first automatic quotation system NASDAQ in 1972 (see Mirowski 2007: 216), and would result in a differentiation of strategies in algorithmic trading.

The ticker tape, and the development of market infrastructures such as telegraph lines spanning the globe, already foreshadows the kinds of infrastructural investments required for HFT, as a form of algorithmic trading characterized by high speed and high volume trading. HFT played a key role in the automation of financial markets since the late 1980s. For example, Mackenzie and Pardo-Guerra (2014) recount the role of Island, a new electronic trading platform launched in 1995, how it challenged existing trading platforms which had not fully automated, and how it already introduced key aspects of automated trading platforms such as ultrafast matching engines, fine-grained pricing or co-location. They also recount how symbiotic the relationship was between Island and Automated Trading Desk, one of the first HFT companies which commenced trading in 1989, and quickly became its biggest client. MacKenzie details how, through bricolage, Automated Trading Desk succeeded in becoming a HFT company, among other things playing a "causal role" in the introduction of all-to-all markets, pushing the computerization of trading, and developing the business model of HFT based on high volume and special market rates (MacKenzie 2016: 175, 180). MacKenzie summarizes: "The use of algorithms helped create markets materially better suited to algorithms" (2016: 190). The ensuing HFT "arms race" has become a "constant of the market design" of financial markets today (Budish et al. 2015: 1553).

Through infrastructural investments in things such as fiber-optic or microwave connections between trading venues, co-location centers and even computer architecture optimized for HFT (Zook/Grote 2017; MacKenzie et al. 2012), the design and temporal regime of markets has come to produce information asym-

metries that enable trading strategies based on high speed, operating in in milli-, micro- and even nanoseconds (Markoff 2018) and on "gaming the plumbing" of financial markets (Toscano 2013). In HFT, speed ultimately trumps smartness. As a consequence, trading algorithms are rather "dumb": speed requires low latency, and all information processing takes time. HFT algorithms therefore need to be kept as simple as possible in order to respond quickly to information changes and to automatically enact a trade, and therefore require constant human supervision (Beverungen/Lange 2018: 86-91). As Arnoldi (2016: 46) puts it, leaving "trading to 'naïve' algos may […] be a choice of economic necessity for high frequency traders […]. Crudely put, algos get faster but not smarter." HFT, in exploiting the plumbing of financial markets, is focused on internal market dynamics and information asymmetries, and it operates on temporal advantages of micro- or by now nanoseconds, and can therefore not afford to give time to complex computation such as that necessary for AI. The "punctuated agency" of algorithms, i.e. the space in which they "draw inferences, analyze contexts, and make decisions in milliseconds" (Hayles 2017: 142) simply doesn't leave time for AI.

That is not to say that AI could not inform HFT strategies. For example, at Automated Trading Desk, basic AI such as linear regression equations were used to predict prices: its machine would calculate an "'adjusted theoretical value' of the stock in question, a prediction of its price 30 seconds in the future", based on market data such as "'the size of the [best] bid relative to the size of the [best] offer', along with 'a short-term trend variable in the transaction prices of the stock'" (MacKenzie 2017: 182-186). That would hardly count as AI today, but it provides an early example of what kinds of models and calculations went into the design of HFT algorithms. In fieldwork conducted by Ann-Christina Lange, high-frequency traders reported that it would take years before AI would become relevant for HFT, with its use only at an experimental stage (Beverungen/Lange 2018: 89). Recent academic work developing approaches to HFT based on reinforcement learning, deep neural networks or convoluted neural networks (e.g. Kearns/Nevmyvaka 2013; Arévalo et al. 2016; Ganesh/Rakheja 2018) similarly suggests that there is a lot of experimentation but little implementation. A recent industry report argues that whereas HFT is "about speed, machine learning is about depth and breadth of insight", and while speed still matters, "it's a different kind of speed" than HFT (McCauley 2016: 4, 7).

Even though the title of Scott Patterson's book *Dark Pools: The Rise of A.I. Trading Machines and the Looming Threat to Wall Street* (2012) would suggest that HFT is largely based on AI, it is not always clear what is considered AI, and his examples either deal with trading strategies more associated with quantitative finance or with examples such as Trading Machines, which in the late 2000s operated an automated trading strategy built on expert systems but which was "a lumbering turtle compared with the rising new breed of speed Bots in the stock market"

(Patterson 2012: 38). In quantitative finance more broadly, developments such as portfolio optimization theory, the capital asset pricing model, and—perhaps most importantly—the Black-Scholes option pricing formula (Kirilenko/Lo 2013: 53-55), have offered calculative devices for deciding which financial assets to invest in, how to devise risk strategies and how to price financial assets such as options. This has allowed the "quants" to conquer Wall Street (Patterson 2010), mostly as part of hedge funds, from the 1980s onwards, and to shape financial markets in the image of their financial models (MacKenzie 2006). Quantitative trading is buy now also algorithmic, i.e. order execution is automatic and much of the trading decisions are also made by algorithms. While many hedge funds also specialize in fast trading in microseconds, in contrast to HFT the focus is on smartness rather than merely speed, and on exploiting not so much the plumbing of financial markets in high volume, high speed trading as on exploiting information asymmetries in trade that operates with holding times of hours, days or weeks rather than seconds.

Although hedge funds and their quantitative traders are extremely secretive, some instances of the deployment of AI are known and point to more recent widespread use. For example, Renaissance Technologies, one of the largest and "considered by many to be the most successful hedge fund in the world" (Patterson 2012: 107), also called "finance's blackest box" (Burton 2016), heavily recruited its staff from cryptographers from the US government and the speech recognition program at IBM (Patterson 2012: 107-117). One of their experts was Robert Mercer, who had worked on Brown clustering as part of Frederick Jelinek's speech recognition team in the 1970s.[2] Or take Haim Bodek, who worked at Hull Trading, a quantitative algorithmic trading firm, from 1997 until it was bought by Goldman Sachs in 1999 (Patterson 2012: 28-30). Bodek had previously worked in fraud detection, and used his machine learning skills at Hull (Patterson 2012: 28), later setting up Trading Machines, which operated from 2007 to 2011 as one of the first fully automated and higher frequency trading outfits (Patterson 2012: 32-60).[3] There are also more recent examples in Patterson's *Dark Pools*, for example Apama, a "complex event processing" engine founded in 1999 and taken over by Software AG in 2013 (Patterson 2012: 62), which already points to the ways in which quantitative trading is embracing a wider set of "alternative" data beyond market and

2 Robert Mercer is now notorious for his engagement in right-wing politics, such as his support for Donald Trump and for Brexit, and for his involvement in the Cambridge Analytica scandal. He resigned from Renaissance in 2017 following political pressure. See Cadwalladr 2017.

3 Haim Bodek is perhaps the most famous whistleblower of Wall Street, because he revealed a secret order type used by high-frequency traders, which was destroying Bodek's own trading strategies at Trading Machines. Bodek is the main character of the documentary *The Wall Street Code* (2013).

trading data—in particular news and social data—for analysis and feedback into trading strategies.

There are three broad current developments related to AI in algorithmic trading relevant to my discussion. First, there is a movement towards automating quantitative trading, that is using computation both for placing orders and for calculating strategies, much like HFT is already automated. Some companies seem to support this strategy by purchasing HFT outfits, such as Citadel buying Automated Trading Desk in 2016. Rebellion Research was perhaps the first fully automated hedge fund, with its "Star" algorithm based on Bayesian networks trading autonomously since 2005 and the updated "Star 2.0" launched in December 2016 (Patterson 2012: 323-335; Metz 2016). Another recent example is Aidyia, another fully automated AI hedge fund that draws "on multiple forms of AI, including one inspired by genetic evolution and another based on probabilistic logic" (Metz 2016). To what extent trading here is really automated remains questionable, however, and the industry seems to have recognized the danger of an over-reliance on and "misplaced confidence" in AI and the need for humans-in-the-loop (McCauley 2016: 14, 16). As in the case of HFT, where traders are unlikely to leave their algorithms unsupervised (Beverungen/Lange 2018), the cases here might be similar to that of Trading Machines, where Bodek also constantly supervised his algorithms operating in a volatile market: "Bodek preferred to trust his own brain. While he used AI methods such as expert systems to build his algos, he preferred to maintain control throughout the trading day. That's why he never left his seat, not even for a bathroom break." (Patterson 2012: 38; see also Satariano/Kumar 2017). Nonetheless, this automation points to a further shift towards a machine-machine ecology in financial markets.

Second, while Aidyia and Rebellion Research are comparatively small, the large majority also of the large hedge funds today claim to work with AI (see e.g. Satariano/Kumar 2017 on Man Group), and there is a significant amount of exchange between companies and research institutes currently developing AI and hedge funds. David Ferruci, developer of IBM's Watson, moved from IBM to become Senior Technologist at Bridgewater Associates in 2012 (Vardi 2016). Li Deng moved from his position as Chief Scientist of Artificial Intelligence at Microsoft to Citadel in 2017 to become Chief Artificial Intelligence Officer. Pedro Domingos, author of *The Master Algorithm* (2015) and expert in markov logic networks, joined D.E. Shaw in 2018 to lead its Machine Learning Research Group. These high-profile movements suggest that hedge funds will play a key role in the development and politics of AI in the coming decades, also through institutions such as the Oxford-Man Institute of Quantitative Finance, and it suggests that the various kinds of AI for which these researchers have expertise will be deployed extensively in algorithmic trading. That is not to say, however, that the application of AI in algorithmic trading will be simple or straight-forward. For example, Li Deng suggests

that there are at least three challenges: low signal-to-noise ratios in the information analyzed to recognize patterns; strong non-stationary with a lot of fake data that needs to be eliminated; and a diversity of data, from speech to text to images, which needs to be amalgamated and analyzed (Deng 2018; see also Frontiers A.I. 2018). Still, this constitutes a significant shift in the cognitive ecology of financial markets, with AI used to make trade both faster and smarter.

Third, there is a significant expansion of the data sources with which algorithmic trading operates and from which it seeks to extract patterns offering trading opportunities, leading to a differentiation of trading strategies (McCauley 2016: 4). In HFT data sources are limited to a clear set of market data mostly related to the order books of the trading platforms in which high-frequency traders operate, and other algorithmic trading relies on a relatively limited set of market and economic data supplied by companies such as Reuters or Bloomberg. Today, however, data sources are multiplying, and so are the companies which offer data streaming and analytics services to algorithmic trading, in particular in relation to social media. Hedge funds such as BlackRock peruse social media and monitor search engines to assist in their investment decisions (De Aenlle 2018), and there are companies such as EquBot which work with proprietary AI and IBM's Watson to parse "millions of articles and news sources to uncover catalysts and events to maximize the probability of market appreciation" ('Artificial Intelligence (AI) and the Technology behind EquBot'), including market sentiment analysis (De Aenlle 2018; McCauley 2016: 8). There are also companies such as Quandl, RavenPack, Eagle Alpha or DataMinr which offer data analytics services for algorithmic trading. DataMinr, for example, specialises in "alternative data" such as "social media, satellite imagery, weather data, and more" ('Alt Data Tips for Traders | Dataminr') and suggests that nearly 80% of traders now use such "alternative" data ('Report: Investors Embrace Alternative Data | Dataminr').[4] This expansion of the data ecology of algorithmic trading calls for AI for pattern recognition, and it would be impossible for a human cognizer to take all of this information into account.[5]

4 The big data analytics company Palantir Technologies, notorious for its involvement in the Cambridge Analytica scandal, also offers services to finance via Palantir Foundry, which however is largely used for fraud detection. On how Palantir operates, see Munn 2018: 27-56.

5 Critical art projects such as Rybn's ADM Trading Bot (see http://www.rybn.org/ANTI/ADM8/) and Derek Curry's hacktivist, tactical media project *Public Dissentiment* (see http://www.publicdissentiment.org/) seek to disrupt financial markets and to raise "awareness of how social media is now interconnected with stock trading" (Curry 2018: 108).

Shifting Cognitive Ecologies

It is perhaps no surprise that AI has been a central aspect of algorithmic trading, and that more recent developments in AI, such as varieties of deep learning, are being adapted in algorithmic trading. If markets have been designed to not put a premium on human cognition, and to assume that the truth lies in the information processor that is the market itself, then it is no surprise that human cognition is further sidestepped by the nonconscious cognition exercised by artificially intelligent machines. What is perhaps more surprising is that conscious, human cognition still plays a central role *in situ* for all algorithmic trading except its most automated variants. In HFT the "nonconscious cognitive arms race" (Hayles 2017: 165) meant that human conscious cognition was superseded by the *speed* of machinic nonconscious cognition, yet the "costs of consciousness" (Hayles 2017: 41-45)—slow response times, the bounded rationality of humans, and so on—had to be balanced against the "costs of nonconscious cognition" (Beverungen/Lange 2018: 80) which could prove financially disastrous. Investments in AI and machine learning have decidedly shifted the cognitive ecology of financial markets towards a premium put not only on *speed*, with HFT still exploiting the plumbing of financial markets, but also on *smartness*—an artificial smartness which further challenges human cognition. Now human consciousness can keep up neither with the speed in which high-frequency algorithms trade, nor with the smartness by which artificially intelligent machines interpret data and find patterns benefitting trading strategies. The costs of this different kind of nonconscious cognition—that of the various AIs at play in algorithmic trade—remain to be enumerated.

It seems a safe bet to assume that one of the costs of the cognitive ecology produced by algorithmic trading is market volatility. There are already plenty of examples of the ways in which both quantitative and HFT have produced crashes (see Kirilenko/Lo 2013: 60-67 for an incomplete list). For quantitative trading, the most serious event was the "quant quake" of August 2007, in the middle of the then emerging financial crisis. Despite seemingly little market pressure, hedge funds were involved in concerted forced liquidations and subsequent de-leveraging, which lead to huge losses for the hedge funds (Kirilenko/Lo 2013: 61-62). For HFT, the most famous example is the flash crash of 6 May 2010, in which the Dow Jones Industrial Average "experienced its biggest one-day point decline on an intraday basis in its entire history and the stock prices of some of the world's largest companies traded at incomprehensible prices", all largely due to high-frequency algorithms negatively interacting with one another (Kirilenko/Lo 2013: 62-63; Borch 2016). The flash crash was not a singular event though: Johnson et al. identified more than 18.000 "ultrafast extreme events" within a five-year period, which they see as consistent with the observation of an "emerging ecology of competitive machines featuring 'crowds' of predatory algorithms" (2013: 1). Furthermore, one

example of the volatility caused by the expanded data ecology of financial markets has been described by Karppi and Crawford (2016) as the "hack crash", in which a false news announcement on Twitter led to a jitter in financial markets caused by automated trading algorithms fed by DataMinr. These examples suggest that the "enigma of exceptional situations, rare events, and black swans"—already associated with derivatives and other aspects of financial markets—remains, and that the "terrain of a dark and confused empiricism" (Vogl 2015: 15) that characterizes financial markets is only exacerbated by AI.

All of these examples furthermore demonstrate that much of the volatility stems from the interaction of (both "dumb" and "smart") automated trading algorithms in financial markets. These automated agents significantly contribute to the ways in which financial markets are marked by interactive dynamics such as imitation (e.g. Borch 2016; Lange 2016). Yet we are "still far from having a robust understanding of how trading algorithms interact", even though how an algorithm "materially acts is shaped by interaction" so that algorithms "need to be understood relationally" (MacKenzie 2019a: 55). The "machine-machine ecology of automated trading" (Hayles 2017: 175) escapes both the understanding and control of humans as it ultimately escapes that of artificially intelligent agents. One could perhaps imagine a fruitful, symbiotic interaction between "smart" trading algorithms, and within the market design field there is certainly still the ambition and hope that multi-agent AI systems including their rules of interaction could be designed from scratch and bring forth a kind of *machina economicus* (Parkes/Wellmann 2015: 272). However, despite market design the AI trading algorithms largely operate independently, and, in that regard, financial markets also do not constitute a "collective capitalist brain" (Terranova 2013: 66); rather, the smart agents compose a sum of small capitalist brains in competition with each other.

This state of affairs is exacerbated by the multiple opacities that are pervasive in financial markets. Burrell suggests that some of the opacities of machine learning algorithms are unsurmountable and a fundamental part of how machine learning operates in terms of its architectures and scales (Burrell 2016: 4-5). Strategies such as explainable AI also currently do not deliver on reducing opacities (Sudmann 2018: 187-191). Yet these opacities of AI are only the latest addition to the other opacities of financial markets, and they are exacerbated by the secretive strategies of algorithmic traders already mentioned above. I already noted how high-frequency traders exploit the plumbing and the information asymmetries of financial markets. Since these constitute a competitive advantage they are as far as possible kept secret; only revelations such as those by Bodek mentioned above or those of Michael Lewis in *Flash Boys* (2014) have led to the microstructure of financial markets becoming more publicly known. There are also the dark pools (MacKenzie 2019b) which largely operate—as their name suggests—in the dark, with order books and many other features of their platforms largely inaccessible

to the public. Lange (2016) also recounts how the setup of HFT prop-shops produces a kind of organizational ignorance, wherein barriers between traders and coders are established which are meant to avoid imitation but can also lead to detrimental side-effects.

To politically challenge the opacities and black boxes of algorithmic trade would therefore require a serious upheaval in financial markets. Attempts at regulation have only addressed these opacities in a limited way, for example by demanding that HFT algorithms be identifiable (e.g. Coombs 2016). Other attempts at changing the design of markets in order to decrease opacities also exist. For example, the Investors Exchange (IEX) is a trading platform celebrated by Lewis (2014) as fighting HFT: a coil of a 61 km long cable around the data center adds around 7 milliseconds to the "round trip" of the algorithms and effectively excludes HFT from being operable on the platform. IEX also has a much more transparent fee structure and offers "fairer" trading conditions. Another suggestion comes from Budish et al. (2015), who suggest to replace continuous limit order books—currently the way trading platforms organize order matching—with batch auctions, which could take place every second and would thereby also largely deny high-frequency traders their temporal advantages (see also Hayles 2017: 165-169). Roth, a key proponent of market design and a teacher of Budish, supports these suggestions (2015: 81-100). Mirowski and Nik-Khah (2008), in a different context, warn against taking on this constructivist perspective of market design, with its neoliberal tint. While there are other nuanced considerations of the politics of algorithmic trading (see e.g. Lange et al. 2016), none of these suggestions address the opacities of AI in algorithmic trading.

It would also be unclear to what extend these changes would lead to a democratization of algorithmic trading and AI. As MacKenzie and Pardo-Guerra reflect in relation to Island, whose order book was open, "allowing anyone real-time sight of its order book", in contrast to all current trading venues: "information might have wanted to be free, but capitalism had other priorities" (2014: 171). Particularly the developments around the expanding data ecologies of financial markets discussed above suggest that rather than democratization, these developments in algorithmic trading and AI lead to a further financialization of daily life (Martin 2002). The social life recorded on social media and elsewhere can now feed into "financial Social Machines, which integrate the innovative high-speed network, social media information, and trading decisions of individuals to provide more accurate price predictions leading to improved financial market integration" (Ma/McGroarty 2017: 245). Here the "great promise" of deep learning, which 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" (Sudmann 2018: 193), is enrolled in what Hayles calls

"vampiric capitalism" (2017: 159) and what I discussed above in terms of financial and cognitive capitalism.

A focus on "the infra-medial conditions of modern AI technology and their political dimension" (Sudmann 2018: 185), as they present themselves in relation to financial markets, and the "shifting our analytical focus toward infrastructures" of financial markets (Pardo-Guerra 2019: 31), as attempted in this contribution, reveals how thoroughly algorithmic trading and the more recent deployment of AI as part of it are enthralled to financial and cognitive capitalism. To get to grips with the politics of AI in algorithmic trading requires an analysis of how AI is enrolled in the service of the extraction of value, most recently from social life as it is recorded on social media and elsewhere. The outline above demonstrates that the politics of AI are increasingly closely entangled with finance and the cognitive ecologies in which it operates. As part of an expanded understanding of the politics of operations (Mezzadra and Neilson 2019), AI deployed as part of finance reveals how it partakes, through financialization, in an extraction of value which it would take more than some tweaks of market design to break out of. Most immediately, the politics of AI in financial markets appears as a politics of cognition, one in which currently the "nonconscious cognitive arms race" (Hayles 2017: 165) is decidedly shifting towards a terrain in which AI is complicit with neoliberal finance capital. This calls for a politics of cognition which thinks through the ways in which AI maybe be extracted from this complicity and be put to other ends not necessarily so congruent with financial and cognitive capitalism.

Acknowledgments

Thanks to Ann-Christina Lange for so thoroughly introducing me to the phenomenon of high-frequency trading and for the collaborative writing projects that ensued. Thanks to the Center for Advanced Internet Studies (CAIS) in Bochum, where I was a Fellow during the summer of 2019 as I completed this text.

References

"Alt Data Tips for Traders | Dataminr". n.d. (https://www.dataminr.com/resources/tips-for-traders-to-take-advantage-of-alt-data). Accessed 11 July 2019.

Amoore, Louise (2019): "Introduction: Thinking with Algorithms: Cognition and Computation in the Work of N. Katherine Hayles." In: Theory, Culture & Society 36/2, pp. 3-16.

Arévalo, Andrés/Niño, Jaime/Hernández, German/Sandoval, Javier (2016): "High-Frequency Trading Strategy Based on Deep Neural Networks." In: De-

Shuang Huang/Kyungsook Han/Abir Hussain (eds.), Intelligent Computing Methodologies, Cham: Springer International Publishing, pp. 424-36.

Arnoldi, Jakob (2016): "Computer Algorithms, Market Manipulation and the Institutionalization of High Frequency Trading." In: Theory, Culture & Society 33/1, pp. 29-52.

"Artificial Intelligence(AI) and the Technology behind EquBot". n.d. (https://equbot.com/technology/). Accessed 11 July 2019.

Beunza, Daniel/Stark, David (2008): "Tools of the Trade: The Socio-Technology of Arbitrage in a Wall Street Trading Room." In: Trevor J. Pinch/Richard Swedberg (eds.), Living in a Material World: Economic Sociology Meets Science and Technology Studies, Cambridge, MA: MIT Press, pp. 253-90.

Beverungen, Armin (2018): "'Kognitiver Kapitalismus'? Nichtbewusste Kognition Und Massenintellektualität." In: Zeitschrift Für Medienwissenschaft 18, pp. 37-49.

Beverungen, Armin (2017): "Zeitlichkeit und Kognition im Hochfrequenzhandel." In: Archiv für Mediengeschichte 17, pp. 9-20.

Beverungen, Armin/Lange, Ann-Christina (2018): "Cognition in High-Frequency Trading: The Costs of Consciousness and the Limits of Automation." In: Theory, Culture & Society 35/6, pp. 75-95.

Borch, Christian (2016): "High-Frequency Trading, Algorithmic Finance and the Flash Crash: Reflections on Eventalization." In: Economy and Society 45/3-4, pp. 350-78.

Budish, Eric/Cramton, Peter/Shim, John (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." In: The Quarterly Journal of Economics 130/4, pp. 1547-1621.

Burrell, Jenna (2016): "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." In: Big Data & Society 3/1, pp. 1-12.

Burton, Catherine (2016): "Inside a Moneymaking Machine Like No Other." In: Bloomberg Markets, November 21 2016 (https://www.bloomberg.com/news/articles/2016-11-21/how-renaissance-s-medallion-fund-became-finance-s-blackest-box). Accessed 11 July 2019.

Coombs, Nathan (2016): "What Is an Algorithm? Financial Regulation in the Era of High-Frequency Trading." In: Economy and Society 45/2, pp. 278-302.

Curry, Derek (2018): "Public Dissentiment: Hacktivism in the Age of High-Frequency Traders." In: Visual Resources 34/1–2, pp. 93-115.

De Aenlle, Conrad (2018): "A.I. Has Arrived in Investing. Humans Are Still Dominating." In: The New York Times, 8 June, sec. Business (https://www.nytimes.com/2018/01/12/business/ai-investing-humans-dominating.html). Accessed 11 July 2019.

Deng, Li (2018): "AI in Finnace." Presented at the EmTech Digital, San Francisco, March 27 2018 (https://events.technologyreview.com/video/watch/li-deng-ai-finance/). Accessed 11 July 2019.

Dyer-Witheford, Nick (2016): "Cybernetics and the Making of a Global Proletariat." In: The Political Economy of Communication 4/1, pp. 35-65.

Frontiers, A.I. (2018): "A Journey with AI: From Speech to Finance." August 24 2018 (https://medium.com/aifrontiers/a-journey-with-ai-from-speech-to-finance-2a60cb3422a). Accessed 11 July 2019.

Fumagalli, Andrea/Mezzadra, Sandro (eds.) (2010): Crisis in the Global Economy: Financial Markets, Social Struggles, and New Political Scenarios, Los Angeles/Cambridge, MA: Semiotext(e)/Distributed by the MIT Press.

Ganesh, Prakhar/Rakheja, Puneet (2018): "Deep Neural Networks in High Frequency Trading." In: ArXiv:1809.01506 [Cs, q-Fin, Stat], September (http://arxiv.org/abs/1809.01506).

Hayles, N. Katherine (2017): Unthought: The Power of the Cognitive Nonconscious, Chicago/London: The University of Chicago Press.

Johnson, Neil/Zhao, Guannan/Hunsader, Eric/Qi, Hong/Johnson, Nicholas/Meng, Jing/Tivnan, Brian (2013): "Abrupt Rise of New Machine Ecology Beyond Human Response Time." In: Scientific Reports 3/September (https://doi.org/10.1038/srep02627).

Karppi, Tero/Crawford, Kate (2016): "Social Media, Financial Algorithms and the Hack Crash" In: Theory, Culture & Society 33/1, pp. 73-92.

Kearns, Michael/Nevmyvaka, Yuriy (2013): "Machine Learning for Market Microstructure and High Frequency Trading." In: David Easley/Marcos Lopez de Prado/ Maureen O'Hara (eds.), High Frequency Trading—New Realities for Traders, Markets and Regulators, London: Risk Books.

Kirilenko, Andrei A./Lo, Andrew W. (2013): "Moore's Law versus Murphy's Law: Algorithmic Trading and Its Discontents." In: Journal of Economic Perspectives 27/2, pp. 51-72.

Lange, Ann-Christina (2016): "Organizational Ignorance: An Ethnographic Study of High-Frequency Trading." In: Economy and Society 45/2, pp. 230-50.

Lange, Ann-Christina/Lenglet, Marc/Seyfert, Robert (2016): "Cultures of High-Frequency Trading: Mapping the Landscape of Algorithmic Developments in Contemporary Financial Markets." In: Economy and Society 45/2, pp. 149-65.

Lewis, Michael (2014): Flash Boys: A Wall Street Revolt, New York: W.W. Norton & Company.

MacKenzie, Donald (2006): An Engine, Not a Camera: How Financial Models Shape Markets, Cambridge, MA: MIT Press.

MacKenzie, Donald (2017): "A Material Political Economy: Automated Trading Desk and Price Prediction in High-Frequency Trading." In: Social Studies of Science 47/2, pp. 172-94.

MacKenzie, Donald (2019a): "How Algorithms Interact: Goffman's 'Interaction Order' in Automated Trading." In: Theory, Culture & Society 36/2, pp. 39-59.

MacKenzie, Donald (2019b): "Market Devices and Structural Dependency: The Origins and Development of 'Dark Pools'." In: Finance and Society 5/1, pp. 1-19.

MacKenzie, Donald/Beunza, Daniel/Millo, Yuval/Pardo-Guerra, Juan Pablo (2012): "Drilling Through the Allegheny Mountains: Liquidity, Materiality and High-Frequency Trading." In: Journal of Cultural Economy 5/3, pp. 279-96.

MacKenzie, Donald/Pardo-Guerra, Juan Pablo (2014): "Insurgent Capitalism: Island, Bricolage and the Re-Making of Finance." In: Economy and Society 43/2, pp. 153-82.

Markoff, John (2018): "Time Split to the Nanosecond Is Precisely What Wall Street Wants." The New York Times, June 30, 2018, sec. Technology (https://www.nytimes.com/2018/06/29/technology/computer-networks-speed-nasdaq.html). Accessed 11 July 2019.

Martin, Randy (2002): Financialization of Daily Life, Philadelphia: Temple University Press.

McCauley, Dennis (2016): Ghosts in the machine: Artifical intelligence, risks and regulation in financial markets, London: Thought Leadership Consulting (www.euromoneythoughtleadership.com/ ghostsinthemachine).

Metz, Cade (2016): "The Rise of the Artificially Intelligent Hedge Fund." *Wired*, January 25, 2016 (https://www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund/). Accessed 11 July 2019.

Mezzadra, Sandro/Neilson, Brett (2019): The Politics of Operations: Excavating Contemporary Capitalism, Durham: Duke University Press.

Mirowski, Philip (2007): "Markets Come to Bits: Evolution, Computation and Markomata in Economic Science." In: Journal of Economic Behavior & Organization 63/2, pp. 209-42.

Mirowski, Philip/Nik-Khah, Edward (2008): "Command Performance: Exploring What STS Thinks It Takes to Build a Market." In: Trebor Pinch/Richard Swedberg (eds.), Living in a Material World: Economic Sociology Meets Science and Technology Studies, Cambridge, MA: The MIT Press, pp. 89-128.

Mirowski, Philip/Nik-Khah, Edward (2017): The Knowledge We Have Lost in Information: The History of Information in Modern Economics, New York City: Oxford University Press.

Munn, Luke (2018): Ferocious Logics: Unmaking the Algorithm, Lüneburg: meson press.

Nik-Khah, Edward/Mirowski, Philip (2019): "The Ghosts of Hayek in Orthodox Microeconomics: Marekts as Information Processors." In: Armin Beverungen/Jens Schröter/Philip Mirowski/Edward Nik-Khah, Markets, Minneapolis/Lüneburg: University of Minnesota Press/meson press.

Parkes, David C./Wellman, Michael P. (2015): "Economic reasoning and artificial intelligence." In: Science 349/6245, pp. 267-272.

Patterson, Scott (2010): The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It, New York, NY: Crown Business.

Patterson, Scott (2013): Dark Pools. The Rise of A.I. Trading Machines and the Looming Threat to Wall Street, London: Random House Business.

"Report: Investors Embrace Alternative Data | Dataminr", n.d. (https://www.dataminr.com/resources/report-investors-embrace-alternative-data). Accessed 11 July 2019.

Roth, Alvin E. (2015): Who Gets What—and Why: The New Economics of Matchmaking and Market Design, Boston, MA: Houghton Mifflin Harcourt.

Satariano, Adam/Kumar, Nishant (2017): "This Massive Hedge Fund Is Betting on AI." In: Bloomberg Markets, September 27, 2017 (https://www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai). Accessed 11 July 2019.

Sudmann, Andreas (2018): "On the Media-Political Dimension of Artificial Intelligence." In: Digital Culture & Society 4/1, pp. 181-200.

Terranova, Tiziana (2013): "Ordinary Psychopathologies of Cognitive Capitalism." In: Arne De Boever/Warren Neidich (eds.), The Psychopathologies of Cognitive Capitalism: Part One, Berlin: Archive Books, pp. 45-68.

Toscano, Alberto (2013): "Gaming the Plumbing: High-Frequency Trading and the Spaces of Capital." In: Mute Magazine 3/4, pp. 74-85.

Vardi, Nathan (2016): "Wall Street Keeps Raiding Silicon Valley For Tech Talent", Forbes March 11, 2016 (https://www.forbes.com/sites/nathanvardi/2016/03/11/wall-street-keeps-raiding-silicon-valley-for-tech-talent/). Accessed 11 July 2019.

Vogl, Joseph (2015): The Specter of Capital, Stanford, CA: Stanford University Press.

Zook, Matthew/Grote, Michael H. (2017): "The Microgeographies of Global Finance: High-Frequency Trading and the Construction of Information Inequality." In: Environment and Planning A 49/1, pp. 121-40.

The Quest for Workable Data
Building Machine Learning Algorithms from Public Sector Archives

Lisa Reutter/Hendrik Storstein Spilker

This chapter analyzes one of the early efforts within the Norwegian Government to improve public services with data from public sector archives. It explores an initiative to develop AI-based services within the Labor and Welfare Administration (NAV). The Norwegian public sector is in a pioneering mood. A new wave of digitalization is drawing attention to platforms, clouds and algorithms. Artificial intelligence holds the potential and promise to revolutionize the public sector. Supervised machine learning, especially, has become the method of choice to achieve the ultimate and somehow diffuse goal of becoming data-driven.[1] There is a lot of excitement about how machine learning algorithms might be used to provide better and more personalized services, changing the way we do bureaucracy and empower citizens. Recording, storing and processing information on citizens has long been a key element of the modern state; however, the calculative systems and techniques to do so have become ever faster, more comprehensive and more autonomous (Beer 2017).

In comparison to private tech-enterprises, public sector organizations possess one obvious advantage—at least "on paper". They possess massive datasets about citizens, of a personal character, often recorded through a long historical span, and continually updated. As Redden notes, "this makes them incredibly valuable from a data analytics perspective" (2018:1). Our informants are very well aware of this potential advantage—some refer to big government data as "our gold". The gold is described as rich, comprehensive, exciting and unique by its miners. Machine learning presents itself as an opportunity to mine the gold lying within the archives, providing the administrators with new and surprising insights into their own work and the citizens they govern.

However, as with real-world mining, extracting gold from its ores is not necessarily a straightforward affair. Someone must dig it out, distinguish it from other

1 The employment of techniques associated with artificial neural networks (ANN) is not allowed in public service, since it is non-transparent and decisions cannot be explained.

items, wash and clean it, to make it suitable for the production of public goods. In NAV, the answer to this has been to establish a new data science environment. This chapter is a story of the unexpected challenges that the AI-division has had to face—and the mundane work that underlies the practices of doing machine learning. Thus, our research question is twofold: *What are the challenges connected to developing AI-based services from public sector archives? How do these early challenges reflect the uncertainties that lie behind the hype of AI in public service?*

These are important perspectives, because the responsibility to realize the supposed empowering and democratizing potential of AI in government-citizen relations ultimately hinges on the ones preparing the data and tinkering with the algorithms. Within the public sector, there has so far been a remarkable amount of optimism and hype related to the development of AI-based services (Vivento AS/Kaupan AS 2015; Teknologirådet 2017). At the same time, there is a growing awareness of the concerns that dominate much of the social science discourse on AI. Ever more aspects of our everyday life are affected by datafication, where human activity and behavior is converted into an analyzable form of digital data and put to multiple uses (Mayer-Schönberg/Cukier 2013). The utilization of big data raises serious questions of privacy, data security and ethics. These questions are, of course, even more critical when AI is employed in the public sector compared to the private sector (cf. Sudmann 2018). There is a significant potential for surveillance as well as a risk of automating unjust practices (cf. Pasquale 2015; Cheney-Lippold 2017; Crawford/boyd 2012).

Of course, these concerns also represent an impetus for research to investigate and develop a deeper understanding of the processes whereby (traditional) public sector archives are transformed into (modern) machine learning algorithms. In order to enable and safeguard democratic influence and control, it is important not only to study the effects of ready-made algorithms but also to investigate algorithms as they are constructed (to paraphrase Latour 1987). Theoretically, we are informed by the work done within the new field of "critical algorithm studies" (Beer 2017; Kitchin 2017; Gillespie 2014). Algorithm studies represent a move beyond the study of digital content and interactions to look at infrastructures that condition the visibility of digital content and the patterns of interaction. The central task for the critical algorithm studies has been to uncover the structures and dynamics and consequences of algorithm-based infrastructures, as these infrastructures often come across as technical and neutral, opaque and impenetrable (Burrell 2016).

However, as algorithm-based infrastructures form the basis for more and more decisions and recommendations in social, political and economic fields, it becomes urgent to address their role and functioning. Pasquale (2015) has famously invoked the metaphor of "the black box" to designate how vital societal decisions are formed beyond visibility and control. Pasquale sketches a scenario

with an *inside* consisting of technology firms, data scientists and their secret and opaque algorithms, in power and control, and a disenfranchised outside, where the rest of us reside, citizens, costumers, the whole old society.

Critical algorithm studies have contributed with valuable insights into the actors and organizations behind or underneath data structuring practices and how they contribute to social ordering. However, according to Flyverbom and Murray, they have so far had "little to say about the actual, inside processes whereby data get organized and structured" (2018: 5-6). Also, boyd and Elish highlight the importance of the mundane work of collecting, cleaning and curating data, because "it is through this mundane work [that] cultural values are embedded into systems" (2018: 69). Despite repeated calls for more ethnographic studies, few have so far been conducted (Kitchin 2017). Thus, an important motivation for our decision to carry out a "laboratory study" of the NAV data science environment was based on the recognition of the absence of such studies and the desire to investigate the minutiae of the processes of algorithm construction. The ultimate goal was to examine the actual practices involved in doing machine learning and the uncertainties and methodological challenges that lie behind the hype of AI in public service (boyd/Elish 2018).

Case Study: The Labor and Welfare Administration

NAV, one of the biggest Norwegian public agencies, is in the forefront of an ongoing nationwide digital transformation. NAV is a public welfare agency that delivers more than 60 different benefits and services, such as unemployment benefits and pensions. The public agency manages approximately one third of the overall Norwegian state budget and operates under the ministry of labor. NAV has about 19.000 employees, of whom approximately 14.000 are employed by the central government, with an additional 5000 at the local level.

The NAV data science environment is part of a newly established division in the IT department. This division intends to concern itself with all environments developing and managing data products in the Labor and Welfare Administration. Hence, its assignment is to arrange for the datafication of citizens. The data science environment was founded in 2017 and consisted, at that point, of observation on the part of a few data scientists and a team leader. The members of this team are key elements of the imagined data-driven public agency.

The urge to become data-driven has its origins both within and outside the organization. Within the organization, individuals have started experimenting with big data for a while. Outside the organization, societal and economic trends, such as downswings in the oil sector, higher immigration rates and the automation of industries present new challenges to the administration and the welfare

state in general. The solution proposed? A data-driven welfare state. Political directives have thus requested an investigation of machine learning and big data:

> It is natural to assume that big data, alongside technologies such as automation and artificial intelligence, will be able to change how the government operates service production in the future (Kommunal- og moderniseringsdepartementet 2016: 109).

In this first phase of the data-driven digital transformation, machine learning algorithms are developed mainly as decision support tools. This can for example be illustrated through a project which wants to bring together municipal and governmental data to improve user follow-up. One of the ambitions of the project is to identify vulnerability in new unemployment cases. The projected end-product is a classification tool, categorizing newly unemployed citizens into two groups, those who are likely in need of intensive follow-up from NAV, and those who are likely to become employed within a short period of time with little intervention required. This assessment has been previously done by the human user support.

> The first assessments of the user's needs should to the furthest extent be automated and based on knowledge of which factors that affects the user's possibilities of entering the workforce. (NAV-ekspertgruppen 2015: 13)

The fieldwork was conducted in January 2018 and included a three-week observation of the data science environment, 11 in-depth interviews with key employees within and outside of the team and a document analysis of internal documents, discussing and presenting the work on big data utilization through machine learning.

Mining the public archive gold mine: The quest for workable data

The modern state and data are inseparably woven together, insofar as the availability of statistical information to the public is a condition and necessity for any democracy (Desrosières 1998: 324). The amount, granularity, immediacy, and variety of digital data about subjects to be governed are unique to contemporary governments (Ruppert/Isin/Bigo 2017). NAV is the second biggest producer of data in the Norwegian public sector. Data have always played an important role in the administration, as it produces official statistics and reports for political decision-making for example on sick leave and unemployment.

The Labor and Welfare Administration practices a culture of archiving, collecting, and storing vast amounts of information on citizens and their own work.

Surprisingly, government agencies tend to forget about the data they possess, unless a crisis or inquiry leads them to deal with the data they forgot or misfiled, or the dots they failed to connect (Prince 2017: 236). Data have so far been used in the production of statistics and then transferred to a public archive or database. The archive changes its role within the organization with the emergence of machine learning—from passive receiver and collector of data, to active provider of data. Rather than gathering dust, the data are projected to drive the day-to-day work of the administration. The administration assumes a yet undiscovered value within public archives which may be key to the administration's survival. The archive hence becomes a source of value and power. The information stored within, becomes an active target of exploration.

Gold mining, however, is a messy business. Companies, such as, for example, Google/Alphabet, Facebook, and Amazon seem to effortlessly feed data back into practice and mine the gold as they create it. By contrast, the creation of machine learning algorithms within the public sector can and has to rely on already existing data and infrastructures. In addition, it has to align with long-existing practices and sets of values. The vast public archives carry the promise of being an invaluable and limitless data source for the creation of machine learning algorithms. However, in practice there exists a broad range of challenges connected to their utilization.

The Labor and Welfare Administration has to build a data-utilization infrastructure on top of the already existing digital infrastructure, which both limits and renders possible the work on machine learning. Which data and how data are used will influence predictions made by algorithms. To produce machine learning algorithms, one needs large amounts of data, against which algorithms can be refined and tested. One of our informants summarizes the overall importance of data work by describing it as a foundation for the data-driven future of the public administration on which the failure or success of initiatives depends:

> So, knowing what data you have and the quality of data, what you are allowed to use it for, I think you have to count on spending a lot of time on that. I think that will be the foundation. And what you are building on top of that will not be better than the foundation.[2]

Much of the work done in the data science environment is described as far from confined to the practice of data analysis and computer science. Before any algo-

2 Due to a disclosure agreement with the administration, none of the informants is identified by any meta-information or pseudonym. All unmarked quotes are thus obtained from any of the 11 interviews. Although this compromises the transparency of the analysis, it was necessary due to the size of the team during observation.

rithm can be constructed, the data scientists themselves need to assemble data, which can be fed to algorithms. The team needs to negotiate the access to training and test data, understand legal frameworks supporting the ethical utilization of data and assess the quality of data. The AI staff needs to make the data machine learnable. This leads to a certain degree of frustration and uncertainty among data scientists, which is however regarded as necessary to ensure the proper use and production of machine learning. So, let's take a closer look at the processes of assembling the data, the organization and structuring of data in practice.

Access

The overall change of the public archive's role requires that the data scientists actively engage with the dusted archive, hence accessing its inner workings. The public agency has standardized and good routines for accumulated data used in public statistics. The data can be accessed and found in a data warehouse. These data are cleaned and adjusted for traditional analysis.

> But what we are concerned with now is the 95 percent of data that are not in the data warehouse, but which are in the raw databases.

The data required are a different from what are used in traditional statistics and described as raw. The latter are a kind of natural, unprocessed and unlimited resource. So how to access this resource and what kind of data does the organization actually have? The supposedly raw data are far from easy to access. Previous reorganizations have led to a distributed data storage system in the administration. Data therefore have a huge variety of owners and are placed all over the organization. Our datafied selves are far from centralized, united entities. The amount, content and whereabouts of the bits and pieces of information on citizens are often uncertain.

> And the practical, technical access to the data seems delayed to say the least. We could have had the time to do so much more if it had not taken so much time for the data scientists to figure out for themselves which data we have and where they are and which unit in the organization you need to consult in order to gain access.

An organizational and administrative divide between municipal offices and the central government does in addition complicate data recirculation. Data stored in different organizational units have not yet been allowed to be assembled or been set up to be put together. Access to data is for example granted on specifically formatted computers, but not necessarily on computers with the right tools

to analyze those data. In addition, putting municipal data and government data together has not yet been possible.

Again, there is a clear sense of old and new. This is not only about a historical perspective on data, but also about the role data are expected to play. New data are projected to be agile and dynamic, flawlessly migrating through the whole of the organization. Accessing the gold mine is about bringing together data from different sources and formatting these data or—metaphorically speaking—building tunnels and shafts to access and transport the gold, so that it can be processed. It is about connecting the dots, building an infrastructure on an already existing infrastructure to direct a data flow towards machine learning algorithms. As previous attempts of assembling data have often failed and few people seem to feel responsible for the overall management of data access and what data in which format are available, the data scientists use a significant amount of time seeking allies in the distributed public archives. These archives, however, show distinct signs of never being intended to be mined, with gatekeepers who are not yet aware of their role as gatekeepers.

Quality

After gaining access to data, the data are often visualized and examined to determine their quality. Quality is here measured in both the amount and completeness of data and the accuracy of information stored in the data. There is a significant amount of uncertainty connected with data quality, as the owners of data know little about their data sets. Machine learning algorithms do not only depend on huge amounts of data, they also depend on data with a certain degree of quality to produce any kind of classification or prediction.

> But it is important we understand how the data are affected and what those data might tell us and how they also will affect the models we are building. Because our models are despite everything not more than what we feed into them and train them to do.

It is in this stage of the gold mining process, that the overall gold metaphor cracks. Data, unlike gold, do not naturally appear in the wild (Cheney-Lippold 2017). Several informants highlight the importance of understanding that most of the data stored in the administration have been produced by human beings collaborating with machines. There is no such thing as raw data. The concept of raw data is, as Bowker (2005) points out, an oxymoron.

Before being stored in a database or archive there are many selection and manipulation opportunities. Even if data sets appear more or less complete, an additional complexity arises connected to the interpretation of the data entries: what

are exactly measured, and how were the measurements made? Data are situated knowledge, socially constructed, historically contingent and context dependent. A sufficient understanding of how data have been registered and stored is regarded as key to the overall goal of becoming data-driven. Data found in the public archive are a result of the work practices in the administration. Without context, the data will appear meaningless to their users. When, for example, visualizing easy register data on the employment/unemployment status of citizens, the team soon discovered blank spaces. What then are these blank spaces? Is it an employer, who forgot to register an employee, or is it an unemployed person, who did not register his or her unemployment? Maybe there has been a misspelling along the way, or maybe there was an error in one of the registration infrastructures? It is simply not easy to tell what happened, and therefore challenging to deal with. A user support employee has therefore been consulted to contextualize the data registered, discussing work practices with the data science environment. The desired quantification of error had however not been achieved at the point of observation.

Machine learning is often accused of legitimizing its social power in that it appears to be mathematical, logical, impartial, consistent, and hence objective (Gillespie 2014). Surprisingly, objectivity is not an element of the team's articulation work. Here participants stress that their prototype itself, the public agency user support, is not objective. Their methods do therefore not need to produce hard facts. Accuracy is more important to the team than objectivity. There are no perfect data or raw data available. Still, the informants think they will be able to extract some applicable meaning from the data sets that extend the knowledge derived from traditional statistics.

Data protection

A third complexity for the data scientists in preparing the data is related to security issues. Who can use data? What data can be used? What data cannot be analyzed together? How to safely transport the gold from the mine to the algorithm? This is an interdisciplinary and wide-ranging challenge. Several informants regard the work on data privacy and information security as the most important, and at the same time most demanding part of their work. As there is no specific framework on how data can and should be utilized and what data can be used, the participants need to negotiate new frameworks for the ethical and legal utilization of data in the public agency. The utilization of big data is new to the organization, as well as the Norwegian public sector. Several official reports do point out the lack of legal guidelines within big data utilization through machine learning (Teknologirådet 2018). Although often mentioned in political speeches, the data-driven welfare state is a future imaginary without practical present guidelines.

The non-existing legal framework leads to uncertainty among the data scientists. Just because data are accessible, it is not automatically ethical to process these data. Machine learning is touching not only the field of privacy, but also justice. Although the administration has long been responsible for handling huge amounts of highly sensitive data, the recirculation of data in its own practice has not yet been explored. The 95% of data previously ignored are not sufficiently regulated. Depending on common sense and gut feelings when working on highly sensitive data is regarded as demanding and unwanted. The consequences of errors are imagined to be significant.

> We cannot let that happen. Everything would stop. We have an incredible amount of information about the whole population of Norway for the most part. And a lot of information about the most vulnerable and difficult situations in people's lives.

Data protection is about assessing the ethical and safe use of data. It is about implementing good HSE in your gold mining project. Several informants compare the work performed in the administration with work on machine learning algorithms done in the private sector. Although the private sector has come a long way in the field of machine learning, participants do not necessarily want to adopt practices and models produced by private sector agents. To produce and facilitate trust among their users in a proper way is important to them. Citizens do expect them to manage data safely. The non-existence of legal guidelines here is tantamount to a free space for experimentation. Several informants highlight that it is important to act not only legally, but also morally and ethically. To quantify and apply moral and ethical behavior in the work on data is however far from straightforward. So far, rather than making mistakes that may affect the trust of citizens, the administration refrains from the use of data.

Discussion and conclusion

We will start this discussion and conclusion part by returning to the metaphor of algorithmic infrastructures as "black boxes". The metaphor invokes an imaginary of a corporate inside in power and control and disempowered and unknowing outside. Of course, as more and more decisions are informed by machine learning models such a lack of transparency and influence constitutes a serious democratic threat. Thus, a central task for critical algorithm studies has been to unpack and examine the constitutive elements of such "black boxes".

Here, transparency cannot be achieved simply with a publishing code, which has been suggested by some in the public sector. We believe that an important contribution from ethnographic studies of the minutiae of algorithm construc-

tion is a more nuanced notion of the degree of control that prevails on the inside. Seaver's (2017) fieldwork depicts the complexity and messiness of programming and the uncertainty among data scientists about the connection between the input to and the outcome of algorithmic processing. Our study dismantles another part of the control imaginary, by demonstrating the uncertain basis for the algorithms. Decisions on the data to feed into algorithms are rarely unambiguous and forthright, but involve dealing with missing values, textual contingencies, context dependencies and interpretative gaps. The process of making data machine learnable is often rendered invisible.

Some of these challenges are generalizable to all types of data preparation, also within private enterprises and applications of deep-learning and neural networks. There is after all no AI without data. Others are more specific to the exploitation of public sector archives. The massive datasets that reside within public bodies have been described—also by our informants—as a "gold mine" for the development of machine learning algorithms that can be used to provide citizens with better and more personalized services. A lot of hope and excitement has been placed on the data gold mine by politicians and decision makers. However, our case study shows that the challenges related to utilizing such archives are, if not insurmountable, at least far larger and more demanding than expected. There is a sense of magic tied to machine learning that minimizes attention to the methods and resources required to produce results (boyd/Elish 2018).

Our first research question was about the challenges related to developing AI-based public services from public sector archives. In this chapter, we chose to present three types of challenges that confronted the data scientists in the early stages of their work. First, there are major obstacles related to getting access to data, both organizationally and technically. These obstacles result from the fact that government data have a huge variety of owners and are placed all over the organization, since previous reorganizations have led to a distributed data storage system. Furthermore, the gatekeepers of specific data sets within the administration are often not easy to find or are unaware of their role as gatekeepers. Also, due to information security risks, data are difficult to flawlessly migrate through the organization. Another challenge relates to the quality of the data in the data sets and the interpretation of their meaning. The data scientists soon discovered that many of the data sets were filled with missing values and approximations and that the numbers were difficult to interpret without knowledge of the aim and context of their registration. What exactly has been measured? How were the measurements made? Finally, the data science environment has to deal with a lot of complex legal and security issues, which makes the progress of its work cumbersome. Who can use the data? What data can be used? Which data sets can be linked together? As there is no existing formal legal framework on how to work

with data in conjunction with machine learning, the data scientists have to develop guidelines along the way—with extra safety margins added.

Interestingly we can find many similarities between the negotiated challenges of the data science environment and critical questions raised by social scientists (Crawford/boyd 2012). The data scientists working on machine learning algorithms are well-aware of the complexity and flaws of the field they are operating in. In addition, we can find similarities of methodological challenges between the social sciences and the doing of machine learning. Like boyd and Elish (2018), we therefore want to point machine learners to an exchange of expertise between data scientist and social scientists. Involving a broader set of expertise is one way forward to increase societal influence on the shaping of digital infrastructures (Ananny/Crawford 2018).

The amount and complexity of the preparation work has to some degree come as a surprise to the administration– the data scientist having had to spend countless days wandering up and down corridors and in and out of offices, searching dusty archives, looking into and interpreting old data sets, and familiarizing himself with unclear legal frameworks and confusing organizational security guidelines. Thus, his days got filled up with tasks that supposedly lay outside his area of expertise, while he hardly got started with the tasks for which he was employed—to create and tinker with machine learning algorithms. The fieldwork was conducted in a phase of exploration and uncertainty. The newly established data science environment had not yet reached what is called the smash point. The data science environment was still working on paving the way toward machine-learning algorithms, making data machine learnable. The future data-driven imagery was diffuse and had no present guidelines. The challenges encountered thus represent a break with the data-driven myth of seamless and impressive functionality and raised serious questions of what is possible and what is actually realistic (boyd/Elish 2018). Rather than describing their work as working toward becoming data-driven, the data scientists perceived it as initiating a more conscious relationship with data.

Ultimately, it appears that the data science environment was set on a quest to reconfigure the organization's overall data practices. This was however not limited to the sheer automation of data practice. The team was intended to change the relationship between data stored in the administration and the administration itself. Data are here imagined to be assigned more power and trust to achieve an overall goal of personalization, enhancement of efficiency, and empowerment. However, those who attributed the most power to the public archive were not the people directly working on machine learning algorithms. For the data scientists, there was a constant struggle between the grand myth of the data-driven welfare state and the real-world experiences with machine learning. This is also reinforced by our own struggle to align the gold mine metaphor given to us by in-

formants with findings in our empirical evidence. The supposed gold mine might not even contain any gold. The very foundation of the data-driven imagery seemed uncertain.

There is no standard solution on how one can and should approach the data-driven imagery yet. This also means that there is still room for reconstructions and configurations of data practices related to the development of AI-based public services. As Cheney-Lippold (2017: 13) argues: "Who speaks for data, […] wields the extraordinary power to frame how we come to explain a phenomenon." The call for democratization of machine learning itself is diffuse and fluid, and so is the overall goal of becoming data-driven within the public sector (cf. Sudmann 2018). Realizing the empowering and democratizing potential of AI in government-citizen relations ultimately hinges on the ones preparing the data and constructing the algorithms. It depends on how data scientists and organizations meet the uncertainties and methodological challenges encountered. To avoid being carried away by the myths and hypes surrounding AI, we need to research mundane negotiations and decisions and turning our attention towards methods and resources required to produce machine learning. Only with insight into the real-world experiences with this kind of work, will we be able to start asking the right questions and be in charge of our data-driven future.

References

Ananny, Mike/Crawford, Kate (2018): "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." In: New Media & Society 20/1, pp. 973-989.

Beer, David (2017): "The social power of algorithms." In: Information, Communication & Society 20/1, pp.1-13.

Bowker, Geoffrey C. (2005): Memory Practices in the Sciences, Cambridge, Massachusetts: MIT Press.

boyd, danah/Elish, Madeleine Clare (2018): "Situating methods in the magic of Big Data and AI." In: Communication Monographs 85/1, pp. 57-80.

Burrell, Jenna (2016): "How the machine 'thinks': Understanding opacity in machine learning algorithms." In: Big Data & Society 3/1, pp. 1-12.

Cheney-Lippold, John (2017): We Are Data: Algorithms and the Making of Our Digital Selves, New York: NYU Press.

Crawford, Kate/boyd, danah (2012) "Critical questions for big data—Provocations for a cultural, technological, and scholarly phenomenon." In: Information, Communication & Society 15/5, pp. 662-679.

Desrosières, Alain (1998): The politics of large numbers: a history of statistical reasoning, La politique des grands nombres. Cambridge, Massachusetts: Harvard University Press.

Flyverbom, Mikkel/Murray, John (2018) "Datastructuring—Organizing and curating digital traces into action." In: Big Data and Society 5/2, pp. 1-12.

Gillespie, Tarleton (2014): "The Relevance of Algorithms". In: Tarleton Gillespie/Pablo J. Boczkowski/Kirsten A. Foot (eds.), Media Technologies: Essays on Communication, Materiality, and Society, Cambridge, Massachusetts: The MIT Press.

Kitchin, Rob (2017) "Thinking critically about and researching algorithms." In: Information, Communication & Society 20/1, pp. 14-29.

Kommunal- og moderniseringsdepartementet (2016): Digital agenda for Norge: IKT for en enklere hverdag og økt produktivitet (Meld. St. nr. 27 (2015-2016)), Oslo: Kommunal- og moderniseringsdepartementet.

Latour, Bruno (1987) Science in action: how to follow scientists and engineers through society, Milton Keynes: Open University Press.

Mayer-Schönberg, Viktor/Cukier, Kenneth (2013) Big Data: A Revolution That Will Change How We Live, Work and Think, London: John Murray.

NAV-ekspertgruppen (2015): Et NAV med muligheter. Sluttrapport fra ekspertgruppen, Oslo: Arbeids- og sosialdepartementet.

Pasquale, Frank (2015) The black box society: The secret algorithms that control money and information, Cambridge, Massachusetts: Harvard University Press.

Prince, Christopher (2017): "Big Data and Privacy: Why Public Organizations Adopt Big Data." In: Canadian Journal of Information & Library Sciences 41/4, pp. 233-244.

Redden, Joanna (2018) "Democratic governance in an age of datafication: Lessons from mapping government discourses and practices." In: Big Data and Society, pp. 1-13.

Ruppert, Evelyn/Isin, Engin/Bigo, Didier (2017) "Data politics." In: Big Data & Society 4/2., pp. 1-7.

Sudmann, Andreas (2018) "On the Media-political Dimension of Artificial Intelligence: Deep Learning as a Black Box and OpenAI". In: Digital Culture & Society, Volume 4, Issue 1, Pages 181-200.

Teknologirådet (2017): Denne gangen er det personlig: Det digitale skiftet i offentlig sektor, Oslo: Teknologirådet.

Teknologirådet (2018): Kunstig intelligens—muligheter, utfordringer og en plan for Norge, Oslo: Teknologirådet.

Vivento AS/Agenda Kaupan AS. (2015): Kartlegging og vurdering av stordata i offentlig sektor. Oslo: Kommunal- og moderniseringsdepartementet.

Plural, Situated Subjects in the Critique of Artificial Intelligence

Tobias Matzner

1. Introduction

Many current critical standpoints on information technologies from the field of artificial intelligence (AI) focus on a difference between human subjects and technology. Such standpoints come in two variants. The first variant is the idea of technical neutrality. Most fortunately, the old argument that technology is neutral, that its social impact "just depends on what you do with it", is losing influence.

However, this argument is often debunked by saying: algorithms are not neutral because they are made by humans. Similarly, on a more abstract level it is often claimed that data sets that are used to train machine learning algorithms mirror human society and thus import its injustices and prejudices (Campolo et al. 2017; O'Neil 2017). That implies that algorithms could be neutral, if humans would not constantly spoil them with their biases. This is a very determinist, platonic story, where human ideas and actions are decisive, which are then put into code and executed by machines (Chun 2008).

Thus, it is important to turn to the second variant of critique. It comprises the positions that show that human subjectivity is not something external to information technology—which is then represented by that technology in a biased or unbiased fashion.[1] Rather, they argue, digital technology does something to human subjectivity itself.

However, most of these approaches form a general verdict on data-based or algorithmic subjectivity, which is usually described as a kind of loss of features that are endorsed. In the following, I will engage with such theories and show using a few cases why such general verdicts harbor the danger to miss the important factor that specific applications of AI connect in quite different manners to pre-existing socio-technical situations and the respective forms of subjectivity. I will use the work of postcolonial theorist Linda Martín Alcoff in order to provide a

1 Such approaches that hinge on an epistemic critique of representation are discussed in detail in (Matzner 2016).

concept of subjectivity that can grasp the impact of recent technological changes but at the same time highlights differences between particularly situated subjects as its resource of normativity—rather than a general feature or lack of algorithmic forms of subjectivity.

2. Applications of AI and two Forms of Critique

Technologies from the field of AI increasingly structure digital communication and interaction, but also what is perceived as "offline" spaces. Especially predictive technologies from machine learning are central to the current services of digital platforms. They are used to personalize search results, to filter posts on social media, to suggest which content we should watch and with whom we should interact. Such predictive technologies also have permeated various institutional and commercial processes. Famously, decisions on credit, insurance and hiring are influenced by scores provided through machine learning algorithms. Security agencies and polices all over the world use AI-enhanced surveillance technologies, in border controls, the processing of visa and asylum applications, the automated evaluation of CCTV footage or—the posterchild of algorithmic bias—recidivism prediction (Angwin and Larson 2016).[2] Predictive uses of machine learning also drive targeted advertising and the creation of other "prediction products" as Shoshanna Zuboff calls them (Zuboff 2019). However, the exact relation of algorithmic technologies, labor, and value creation in the digital economy has yet to be clarified (Heilmann 2015; Srnicek 2016).

A lot of critical work has been done regarding the information that can be derived from such algorithmic predictions, their epistemic status and their tendency to veil biases in the aura of machinic objectivity (Aradau and Blanke 2015; Kitchin 2014, 2017; Pasquale 2015). Elsewhere I have argued that these important inquiries must be amended with critical scrutiny regarding what these algorithmic practices do to subjects (Matzner 2016). For example, the use of daily interaction on social media for surveillance purposes imports meanings and practices of suspicion and mistrust into these interactions.

Following this intuition, it is important to ask which new forms of subjectivity, or which shifts in forms of subjectivity, the increasing impact of AI-based technologies engenders. Many critical accounts, including those from activist positions, implicitly presuppose the model of subjectivity predominant in liberal political

2 The research on each of the applications of AI I have mentioned here is growing almost daily. Cathy O' Neil's (2017) book is a good starting point for references on the applications I have mentioned here—even if her criticism falls within the line of defending autonomous subjects against technology that I criticize.

thought: a rational, self-reflexive and autonomous subject. Algorithmic processes that apply machine learning technologies are seen as an imposition on each of these aspects. For example, discourses on the so called "filter bubble" focus on the prevalence of emotional rather than rational discourse through algorithmic filtering, the lack of transparency of the algorithms so that the self-reflective thinking necessary for autonomous judgements is impaired and thus an autonomous use of technology is no longer possible (Pariser 2011; Zuiderveen et al. 2016). However, the clear opposition between liberal subjects and technological impositions is too simple. The entire story of cybernetics, which led up to current connectionist AI (Sudmann 2018), has been structured by a deeply ambivalent relation to liberal ideas. On the one hand, cybernetics was driven by the idea to develop new and powerful tools for free and more effective human actions. On the other hand, the ensuing idea of the human, the animal and the machine as essentially matters of control and communication is a deep threat to ideas of autonomy and self-reflexivity (Hayles 1999: 87). Also the recent applications of AI can in many regards be considered as a liberal project (Matzner 2019). Furthermore, the concrete challenges that current applications of AI pose cannot be easily solved on an individual level. For example, issues of privacy and data protection, if solved within the liberal paradigm, presuppose a partition of data into personal data, which each respective individual can control (Matzner 2014). However, the attractivity of current AI-driven data analysis is to use data on an aggregate level, which finds patterns and associations that cannot be reduced to single users' contributions. Even personalized systems like recommender systems or timeline filtering algorithms usually do not store a digital model of the user, as the use of "data doubles" and other concepts might suggest (Lyon 2014). Rather, the decision is taken for each individual item, regarding which an approximation of the user's interest is derived from the current stream of data and state of the user's connections.[3] Thus, such problems need to be addressed on the aggregate level of data usage rather than only individualized parts. Finally, liberal theory has come under scrutiny from feminist and other critical theories for engaging what Hayles calls the "practices that have given liberalism a bad name" (Hayles 1999: 87).

For these reasons, critical theories of applications of AI that take recourse to other sources of normativity are preferable. A prominent and elaborated example is Antoinette Rouvroy's concept of algorithmic governance. She derives her nor-

3 As usual, it is hard to know exactly how prominent applications like Twitter's timeline or Facebook's newsfeed are filtered. Thus, I derive my observation from the published research. Already early research done at Yahoo (De Francisci, Morales et al. 2012) that has spearheaded a lot of research on personalized content, did not use persistent models of the user. The approach uses support vector machines for classification. In the meantime, personalization, like most other machine learning tasks, has switched to neural networks, and thus to even more data driven and dynamic approaches. See for example a recent paper by Microsoft Research (Zheng et al. 2018).

mative stance from an idea of humanity that is precisely based on the absence of full autonomy and rationality. Rouvroy follows theories of Judith Butler and Louis Althusser (Rouvroy 2013: 158). Both describe subjects as never in control of themselves, because they are essentially dependent on others. However, these others are not simply determining. The influence of others on us happens in social interaction which neither we nor the others fully control. It is particularly that excess and openness of human action that enables critique and meaningful interaction. Albeit, this very excess is threatened by algorithms:

> [W]hat has to be preserved as a resource antecedent to both the 'subject' and sociality, as excess of the world over the algorithmic reality, is 'the common'; this 'in between', this space of common appearance (*comparution*) within which we are mutually addressed to each other. (Ibid.: 159-60)

Thus, Rouvroy sees human interaction yielding a potential for novelty and spontaneity that computing never can grasp. In her account, algorithmic governance, much in line with the description above, is not focused on individual subjects. Rather, algorithmic governance is "[e]ffected through the reconfiguration of informational and physical architectures and/or environments within which certain things become impossible or unthinkable, and throwing alerts or stimuli producing reflex responses rather than interpretation and reflection." (Ibid.: 155) This description clearly echoes cybernetic worries of the loss of the subject. Algorithms, in Rouvroy's words, reduce the virtual to the actual, the possible to the statistically probable, the living to the computational (Rouvroy 2017). Thus, the main line of critique Rouvroy harnesses has against algorithmic governance is again a certain loss of subjectivity, in this case a form of relational subjectivity that can contribute to the creation of politics and resistance.

3. Critique on a general level and the importance of situated subjects

Such analyses provide important insights into the consequences of the application of AI. In particular, Rouvroy's account does justice to the specifics of many recent forms of AI-based verdicts and activities, which work on the supra-individual level and which provide incentives for action rather than information. It is important to note that there are some applications of AI that can be seen very much in line with more Foucauldian forms of disciplinary power (Matzner 2017). In particular, these can be found at the borders of the Western, capitalist societies that Rouvroy and most other critics of AI take into focus. Yet, within these societies, such analyses are pertinent. However, in their attempt to find a general verdict on a specific loss of subjectivity through applications of AI, they miss important qualifications.

This is not only a matter of descriptive accuracy but also means that AI is not per se such an anti-political technology as which it appears in these analyses. Its anti-political effects do not fall on subjects as such but on particular subjects—and on each in a different manner.

The problems of such general verdicts can e.g. be seen in Wendy Hoi Kyong Chun's analysis of filter bubbles. She shows that the theory of the filter bubble is based on the concept of homophily: The idea that human beings tend to orient themselves towards others who are or think similarly. Critics of the filter-bubble argue that algorithmic content creation tends to enforce that human tendency in a dangerous manner, which can lead to all kinds of extreme and racist communities. However, the problem of the algorithmic selection is not seen in the content itself, but in the concept of similarity that applies to all content in the same manner. That way, homophily

> serves as an alibi for the inequality it maps, while also obviating politics: homophily (often allegedly of those discriminated against)not racism, sexism, and inequalitybecomes the source of inequality, making injustice 'natural' and 'ecological.' (Chun 2018: 76)

Algorithmic filtering, which is an exemplary case of what Rouvroy calls the "reconfiguration of informational […] architectures" (Rouvroy 2013: 155), is criticized regarding a universal trait of human subject formation. Chun shows that it is necessary to take the social situation of subjects, which enable racism, sexism, inequality into account. Another case in point would be the infamous analysis by ProPublica, which has shown that a recidivism prediction software was biased against blacks (Angwin and Larson 2016). This case has been discussed almost too much, so I just want to highlight that the software did not use any racial features as input. Thus, even if the efficacy of algorithms does not work in terms of race, it still addresses and produces race.

In order to overcome the line of critique mentioned in the beginning, which implies a neutral technology spoilt by biased data, it is necessary to show how any kind of media and AI in particular engage with socially and culturally situated subjects—including race.

4. Situated subjects

In her book on what she calls "habitual new media," Chun describes data analytics and their turn away from individuals quite similar to Rouvroy. Her analysis centers on the concept of habit: rather than focusing on an individual subject, data analytics try to grasp habits, established ways of acting, and consequently tries

to form and influence these habits. In order to achieve this, they focus on the correlation between habits rather than individual acts or even individual patterns. "Through this, individual actions become indications of collective patterns rather than exceptions." (Chun 2016: 57) These patterns are the object of optimization, quite similar to Rouvroy's description of the reconfiguration of architectures and environments in order to achieve certain behaviors.

Here, I cannot follow the detailed conceptual work in which Chun engages with the notion of habit. However, I want to follow her suggestion to connect this take on habit from media theory with thoughts on habit that relate to alterity:

> habit is publicity: it is the experience, the scar, of others that linger in the self. Habits are remnants of the past—past goals/selves, past experiences—that live on in our reactions to the environment today, as we anticipate tomorrow. Through habit we inhabit and are inhabited by alterity. (Ibid.: 95)

Chun encourages us to ask how such habits are changed through recent developments in digital media and how they can change again in order to change society (ibid.: 8). This implies that not habit per se is the problem, but differences among habits. However, Chun herself does not take these differences serious enough. Her main preoccupation are liberal injunctions to protect the subject from alterity and technological impositions. By fusing both, she urges to find ways to "inhabit" our habitualized relations to others, which includes to "warily embrace" the many new flows of data, connections, configurations of subjectivity. Here she has a much more positive outlook on technology than Rouvroy. Yet, she underestimates how any form of exchange and ensuing subjectivity is formed by power—not just the private, liberal space. Some socio-technical positions are quite hurtful to inhabit. Thus, in the following I want to suggest a middle ground, which however shares the outlook that changes in the ways we perceive and the ways we (can) act in a given situation are not only the aim of algorithmic means of governing. They are a fundamental way how subjectivity works. This is analyzed in detail by Linda Martín Alcoff in her book visible identities.

Alcoff starts from the Foucauldian insight that power is not just an imposition from the outside. Rather, being a thinking and acting subject also means to be situated in power. However, contrary to Foucauldian analysis which focuses on the disciplinary subjection under norms, Alcoff shows via a theory of alterity and habitualization that our perceptions and actions are formed by the practices we perform and by the situations we have found ourselves in. Our past experiences leave traces that Alcoff describes in line with central insights from what is commonly discussed as theory of social practices (Reckwitz 2002): "[T]he interpretive horizon that constitutes our identity is undoubtedly constituted […] by a wealth of tacit knowledge located in the body." (Alcoff 2006: 106) Such tacit knowledge and

habitualizations have their location in practices. They are not necessarily imposed on us, rather they are the growing residue of the way we act—or are forced to act. The latter of course remains important, but is not the only way how habitualization comes about. It is an integral part of the way we make meaning of our situation and how we structure our actions. A lot of these ways of perceiving and acting come from others—via education, the various contexts we live, work, play, learn, etc. All of these contexts or situations are structured by collective practices. Practices in which we do something but at the same time attain a subjectivity. Others tell us—more or less implicitly—who we are, what we become or should become by doing certain things, what is apt or usual for "someone like you" etc. As Alcoff states:

> Part of what the collective praxis creates are aspects of the self. Our preferences, our dispositions toward certain kinds of feelings in certain kinds of situations, what typically causes fear, anxiety, calmness, anger, and so on, are affected by our cultural and historical location. Sometimes people take such internal feelings as proof of a natural origin, as when a homosexual kiss elicits feelings of disgust. The feelings may well be quite real, but this is not proof that homosexuality is unnatural; physical reactions can be altered by knowledge and acquaintance. This example suggests the most powerful role that the other plays in self-formation: the character of the other determines in no small part the self. (Ibid.: 115)

Regarding theories of the subject, it is often important to highlight this influence against ideas of innate characteristics or the demand to become as self-reflexive as possible. Then it suffices that "the situation" of the subject is important—but not so much what that situation actually is. Alcoff highlights that the practices we become habituated in are structured by all kinds of social difference. She mainly analyses race and gender, but points at social strata, education and financial resources as others. Thus, apparently quotidian practices are different for subjects inhabiting different social positions. E.g. she lists all kinds of things that are particular for women, with regard to the work of Simone de Beauvoir and Iris Marion Young:

> There is not only throwing and sitting, but standing, walking, running, patterns of conversation involving interruptions and dominating the topics, perceptual orientations that can encompass sideline issues so as to notice household dirt, distressed children, bored interlocutors, and so forth, as well as the very interior experience of one's own emotional subjectivity. (Ibid.: 106)

She has similar lists for race and cross cultural and intersectional indices (ibid.: 106 et. seqq.). Alcoff describes that we perceive situations, spaces and persons dif-

ferently, depending on our preceding experiences, the cultures and meanings in which we have moved. We enter a subway differently as man or a woman, as person with white skin or person of color. Here, cultures and meanings should not be understood as externalizable structures. They only persist in collective practices and particularly in what Alcoff calls "perceptual practice" (ibid.: 115).

It is of course possible to reflect and to engage with one's own habitualization and the practices in which this happens—but not by rendering them fully transparent to oneself. We can act very consciously of the fact that our perception and the possible forms of action are deeply intertwined with contingent practices. Nevertheless, these practices are the very context in which meaning and perception emerge. Furthermore, experiencing something means to be somewhere and thus does not only enable knowledge, but also the possibility to be changed in one's subjectivity: "Knowing is a kind of immanent engagement, in which one's own self is engaged by the world […] rather than standing apart and above."(Ibid.: 111) Thus, when we attempt to engage with our own situation, practices form both the context and the site of this engagement. In consequence, habitualization can only yield to another form of habitualization:

> The phenomenal world constantly folds back on itself, adding to what has come before and what remains still in the background of the present moment; the past is that which has been surpassed, yet remains within. There are no complete breaks or total separations, only folds within a continuous cloth, pregnant with latent meaning. (Ibid.: 110)

This also entails that a lasting change of subjectivities cannot be based on individual attempts. Rather, the practices, the ensuing social relations need to be changed in order to bring about different forms of habitualization and subjectivities:

> Experiences matter, but their meaning for us is both ambiguous and dynamic. We are embodied, yet not reduced to physical determinations imagined as existing outside of our place in culture and history. This account helps to capture the dialectics of social identities, in which we are both interpellated into existing categories as well as making them our own. (Ibid.: 111)

This analysis of situatedness, also the situatedness of social change has consequences for the kind of politics that Rouvroy advocates. Alcoff denies the necessity for an account of (human) beings as always in excess, or a "pure capacity of negation or of flight" (ibid.: 112). Even if such ideas of politics are deeply inspired by critiques of the subject, Alcoff contends that they still contain remnants of the "dualism" that inspires liberal accounts, which try to somehow separate the individual from others or society. However, the habitual situatedness within practices

is not just part of oppressive and determining identities—although these are in the foreground of Alcoff's discussion. They are part of any subjectivity, including those with which we identify, in which we find pleasure, friendship, solidarity, luck. In consequence, to attain these we do not need to exceed situatedness, we just need to change the situation. In Alcoff's words: "Moral agency, subjectivity, and reasoning capacities are made possible within social networks of certain types. There is no amorphous substance or pure capacity lying pristine below the layer at which social constructions of identity take hold." (Ibid.)

5. Situated Subjects, AI and Politics

Alcoff herself does not discuss media and technologies. However, her thinking is deeply inspired by Merleau-Ponty's phenomenology, which contains the mediatized structure of experience at its core—represented by the infamous example of a blind person's stick (Merleau-Ponty 1962: 152, see also Alcoff 2006: 188). Thus, Alcoff's thought can be easily amended with the necessary reflections on media technology.

In his discussion of interfaces, Christoph Ernst shows that interaction with digital technology via interfaces implies a situated subject, including the body (Ernst 2017: 100). Interfaces only work because they can address implicit knowledge which is rooted in practices and thus is structured by social rules (ibid.: 102). Interface research and design even tries to consciously address that using what Ernst calls in reference to cognitive science a "conceptual model" (ibid.). While this bears the potential of manipulative attempts, it is not manipulation per se but a necessity for an interface to work, i.e. to do justice to the fact that interfaces do not just interact with generic human beings but concretely situated subjects.

Ernst discusses interfaces, not the more abstract adjustment of architectures or environments that Rouvroy emphasizes, which work through "stimuli and signals that produce reflex responses". However, if this efficacy is precisely the defining factor of technologies in algorithmic governance, they need to connect to the habitualized subjectivities not unlike interfaces (see also Distelmeyer 2017). Thus, even if these technologies do not aim at a set of norms and ensuing subjectivity, they are still entangled with situated subjects.

This also is confirmed by Chun's observation that predictive analytics is tied to habitual practices, which I have cited above. Using Alcoff's theory, we now return to the point that habitualization itself is not the problem. That a lot happens on a pre-conscious and habitual level, does not mean that the applications of AI work deterministically on us. Rather they interact with structures of perception and action that can certainly be influenced by algorithms, but that are also characterized by a pre-formed depth that results from prior experiences. This can change

the previsioned result of algorithmic governance in many ways, ensuing in frictionless, almost unnoticed influence, as well potentials to inhabit and embrace and potentially evolve one's situation as Chun suggests—but also many kinds of tensions, hurt and resistances. This is the main point here. Subjects are concretely situated subjects and algorithmic governance, particularly because it acts one supra-subjective level connects quite differently to the various forms of subjectivity.

This already starts on the level of perception: For example, EU citizens that are not recognized by automated border control terminals that use AI based face recognition will immediately see this as malfunction of technology. Migrants might perceive this as a threatening decision. Also, the even less tangible adaptations of environments connect to situated subjects. This is precisely the reason why applications of AI are not neutral. Not just because they are based on biased training data; but because they connect differently to different forms of subjectivity. The algorithmic filtering of news is problematic because it connects better to certain subjects and communities structured by hate and othering than to other forms. Recidivism prediction enlarges and continues a security system that is based on race discrimination. John Chenney-Lippold has shown that the algorithmic selection of merchandise based on a machine learning system that tries to predict the users gender connects better to heterosexual, commodified forms of gender than others (Cheney-Lippold 2011).

AI has yielded many technologies that have enhanced the efficacy of technologies in the sense that they directly impact the way we perceive and act in the world. This impact, however, does not amount to a loss of subjectivity in general. Rather it reconfigures different forms of subjectivity in different manners. The normative source of critique then does not lie in a difference between a new form of subjectivity under algorithmic governance and one that is somehow beyond that. Rather, the source of critique lies in the differences that already exist between subject positions, and the many ways in which they are shifted through technology. Chun's suggestion to "warily embrace" this situation can be one way of trying to achieve a change for the better in such a situated manner. However, other ways lie in the refusal to accept a situated subject position, which might include demands for privacy protection as well as ceasing to use particular technologies altogether. These demands will need a socio-technical index. That is, they are not the demand to return to an independent subject position like the liberal strands that both Chun and Rouvroy criticize would have it. Still, privacy, cloaking of data, refusing to be implicated in automated analysis might be a necessary resource to find better and viable situations for persons whose subject position becomes entangled with applications of AI in hurtful, abusive, disempowering ways.

To repeat, the challenge of critique is not to escape situatedness but to change the situation. Alcoff's reflection shows that this will always be a situatedness with others; and as my amendment of her theoretical outlook illustrates, it will always

be a situatedness with technology. Thus, in the end this amounts to a political and democratic challenge. Our situation is always already related to others. Applications of AI make that very clear: they focus on relational data and as data driven technologies only make sense at the aggregate level.

At the same time, as Alcoff shows, is it impossible to fully reflect that situatedness and relationality. It is not a system or environment but an encroachment of many different "past goals/past selves—past experiences" as Chun writes. This creates many differences in perception and possibilities for actions for each subject. Thus, issues of epistemology and of power are fused. In this sense, the political challenges are first to get to get to know the situation of others, the way that technologies connect to their subject position. In a second step these positions need to be reconciled to achieve a new and better configuration of technology. This needs to be a democratic solution, not in the sense of finding a compromise between pre-existing interests, but in the sense that subjects always already form a related, socio-technically situated plurality.

References

Ahmed, Sara (2014): *The cultural politics of emotion*. Edinburgh University Press.

Alcoff, Linda Martín (2006): *Visible identities: race, gender, and the self*. New York: Oxford University Press.

Angwin, Julia; Larson, Jeff (2016): "Machine Bias". *ProPublica*. Retrieved on 13.06.2018 (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Aradau, Claudia; Blanke, Tobias (2015): "The (Big) Data-security assemblage: Knowledge and critique". In: *Big Data & Society* 2 (2), pp. 1-12.

Campolo, Alex; Sanfilippo, Madelyn; Whittaker, Meredith; et al. (2017): *AINow 2017 Report*. New York: AI Now Institute.

Cheney-Lippold, John (2011): "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control". In: *Theory, Culture & Society*. 28 (6), pp. 164-181.

Chun, Wendy Hui Kyong (2008): "On 'sourcery,' or code as fetish". In: *Configurations*. 16 (3), pp. 299-324.

Chun, Wendy Hui Kyong (2018): "Queerying Homophily". In: Apprich, Clemens; Cramer, Florian; Chun, Wendy Hui Kyong; et al. (eds.) *Pattern Discrimination*. Lüneburg: meson pp. 59-98.

Chun, Wendy Hui Kyong (2016): *Updating to remain the same: habitual new media*. Cambridge, MA: The MIT Press.

De Francisci Morales, Gianmarco; Gionis, Aristides; Lucchese, Claudio (2012): "From chatter to headlines: harnessing the real-time web for personalized

news recommendation". In: *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 153-162.

Distelmeyer, Jan (2017): "AN/LEITEN – Implikationen und Zwecke der Computerisie- rung". In: *Navigationen* 12 (7), pp. 99-116.

Ernst, Christoph (2017): "Implizites Wissen, Kognition und die Praxistheorie des Interfaces". In: *Navigationen* 12 (7), pp. 99-116.

Hayles, N. Katherine (1999): *How We Became Posthuman*. Chicago: University of Chicago Press.

Heilmann, Till A. (2015): "Datenarbeit im 'Capture'-Kapitalismus. Zur Ausweitung der Verwertungszone im Zeitalter informatischer Überwachung". In: *Zeitschrift für Medienwissenschaft*. (2), pp. 35-48.

Kitchin, Rob (2014): "Big Data, new epistemologies and paradigm shifts". In: *Big Data & Society* 1 (1), pp. 1-12.

Kitchin, Rob (2017): "Thinking critically about and researching algorithms". In: *Information, Communication & Society*. 20 (1), pp. 14-29.

Lyon, David (2014): "Surveillance, Snowden, and Big Data: Capacities, consequences, critique". In: *Big Data & Society*. 1 (2), pp. 1-13.

Matzner, Tobias (2016): "Beyond data as representation: the performativity of Big Data in surveillance". In: *Surveillance & Society*. 14 (2), pp. 197-210.

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

Matzner, Tobias (2019): "The Human Is Dead—Long Live the Algorithm! Human-Algorithmic Ensembles and Liberal Subjectivity". In: *Theory, Culture & Society*. Online First.

Matzner, Tobias (2014): "Why privacy is not enough privacy in the context of 'ubiquitous computing' and 'big data'". In: *Journal of Information, Communication and Ethics in Society* 12 (2), pp. 93-106.

Merleau-Ponty, Maurice (1962[1945]): *Phenomenology of perception*. London: Routledge.

O'Neil, Cathy (2017): *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York: B/D/W/Y Broadway Books.

Pariser, Eli (2011): *The filter bubble: what the Internet is hiding from you*. London: Viking.

Pasquale, Frank (2015): *The Black Box Society: the secret algorithms that control money and information*. Cambridge: Harvard University Press.

Reckwitz, Andreas (2002): "Toward a theory of social practices: a development in culturalist theorizing". In: *European journal of social theory* 5 (2), pp. 243-263.

Rouvroy, Antoinette (2017): "Gouverner hors les normes: la gouvernementalité algorithmique". In: *Lacan Quotidien* 733.

Rouvroy, Antoinette (2013): "The end(s) of critique: data-behaviourism vs. due-process.". In: Mireille Hildebrandt, Katja de Vries (ed.) *Privacy, Due Process and the Computational Turn—The Philosophy of Law Meets the Philosophy of Technology*. London: Routledge pp. 143-167.

Srnicek, Nick (2016): *Platform capitalism*. Cambridge: Polity.

Sudmann, Andreas (2018): "Szenarien des Postdigitalen". In: Engemann, Christoph; Sudmann, Andreas (eds.) *Machine Learning-Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*. Bielefeld: transcipt, pp. 55-74.

Zheng, Guanjie; Zhang, Fuzheng; Zheng, Zihan; et al. (2018): "DRN: A Deep Reinforcement Learning Framework for News Recommendation". In: *Proceedings of the 2018 World Wide Web Conference* (WWW '18), pp. 167-176.

Zuboff, Shoshana (2019): *The age of surveillance capitalism: the fight for the future at the new frontier of power.* London: Profile Books.

Zuiderveen Borgesius, Frederik; Trilling, Damian; Moeller, Judith; et al. (2016): "Should we worry about filter bubbles?". In: *Internet Policy Review* 5 (1).

Deep Learning's Governmentality
The Other Black Box

Jonathan Roberge/Kevin Morin/Marius Senneville

Introduction

Frank Pasquale's 2016 book *The Black Box Society* is now considered a landmark study in law-related disciplines, in the social sciences and beyond. The topic in itself, digitalization and the invasiveness of the internet, is of the utmost importance. The book reveals the inversion of operational secrecy by digital platforms and the extensive access to users' private data. Facebook, Google and the like collect and aggregate bits and bytes of information to create massive profiling schemes, the *modus operandi* of which little is known. Problematically, as private actors, they acquire massive data and "knowledge" about the society and our individual behaviors that we don't. To that complex and most certainly unpleasant reality, Pasquale's analytical rigor and finesse contribute valuable insights on potential regulations and on the possibility of developing a smarter citizenry. However, we want to argue that there is another, broader and maybe more cultural reason why *The Black Box Society* draw so much attention—thus further explaining the success of the book. The image of a concealed, networked entity is evocative of some common fears. It captures a sense of "loss of control" vis-à-vis the latest automation processes. Such an algorithmic black box, in other words, taps into a diffused anxiety regarding what is to be called a "known-unknown", i.e. something we recognize to be a mostly hidden form of knowledge production. The image of a black box is a disenchanted one; here lies its strength *as well as* its weakness.

While mostly in line with Pasquale's effort to decipher the opaqueness of our data-driven world, it also appears significant to question the limits of the black box as a heuristic if not holistic image. Scholars such as Geiger (2017), Sudmann (2018), Burrell (2016) or Bucher (2016) have explored this territory. The latter, for instance, has argued that "the widespread notion of algorithms as black boxes may prevent research more than encouraging it", noting that the notion is "too readily used" (84). She then calls for critical scrutiny of algorithms and algorithmic systems using a three-step method: i) "do not fear the black box"; ii) "do not expect the solution to be inside" and iii) "consider the boxing of the box". Whereas

the first step could be understood as encompassing the entire process, the others could be conceived as forming a complementary pair, together examining the inside and outside of the box. Moreover, such an approach is particularly suited to analyzing the recent shift towards deep learning algorithms, but also to machine learning techniques of different sorts, and everything nowadays labelled artificial intelligence. How and to what extent these algorithms are practically and symbolically different from the previous so-called 'generations'. What would they entail in terms of opacity, ambiguity and vagueness? Where, what, and whom should we look at to develop critical insight and robust interpretation? These questions, once docked to Bucher's steps, could serve as guide to this chapter.

Examined closely, the idea of "not expecting the solution to be inside" matches perfectly with the ingrained logic and historical development of deep learning algorithms. "Open sesame!" is a task that cannot be programmed, or "learned" for that matter. Despite its biological inspiration and the romantic-teleological accounts of the field's historical development (Rosenblatt, 1958; Hinton & al. 2006), the fact remains that what stands for "learning" is in fact adaptation and self-tweaking. The mathematical structure modifies itself while interacting and coping with the data stream coming from the outside world (Litvinski 2018; L'heureux & al. 2017). Backpropagation, recursive loops and other subtleties thus not only reflect but also enact a reality in flux. Another way of looking at such uncertainty is with the discrepancy between the more classical symbolic approach to AI and today's connectionist or neo-connectionist shift (Cardon & al. 2018). Whereas the first relied on deduction, explicit modeling, abstract rules and programmable languages to create a logical and formal mode of reasoning, the second is based on induction, whereby connected hypotheses and approximations produce "optimized" perceptions and predictions about what is going on in the data, inasmuch as data translates into improved rates of predictability (Mackenzie 2017, Sudmann 2018). Layers of non-linear calculus thus inform something of a "deep" but shallow architecture which does not necessarily form an inexplicable AI, but which pushes the limits of its explicability further away. If not fully black, the box of current AI is very grey, to say the least. This can also be seen in the problems scholars are now facing concerning the reproducibility of small-scale theories, where current practices of publication generally prevent them from sharing both source code and training database, or to address the hazards related to the arbitrary setting of hyperparameters, or even the unavoidable randomness inherent in the process of generating training values (Hutson 2018). Managing and massaging that much data is never an easy task, especially not in an experimental environment, and even less in the real world where the saying "garbage in, garbage out" remains thoroughly valid. The multiple problems nowadays with bias fall under the same

category, and could serve as one last example here, namely that the box cannot by its very definition be the solution to something bigger than itself.[1]

The internal problems sketched above might not even compare to what is at stake with Bucher's insight probing to "consider the boxing of the box". In fact, there is a long history of social sciences in general and in STS in particular to devote a great deal of attention to everything surrounding a given piece of technology (Bijker & al. 2012). In the specific case of deep learning algorithms, it is all the more important to remember that "they are embedded in larger, far more complex assemblages" that never cease to influence their shape and content (Gillespie 2014: 3). The question, then, is how to make sense of such molding pressures and the kind of opacity they produce. It is about the complexity of a given context or a given "ecology", yet we want to argue that the best way to consider such boxing is through a *networked* approach. As stated elsewhere, "[…] there is not one box, but multiple boxes. The opacity of algorithms is more precisely expressed in different forms of opacity, all of which, in specific ways, are contingent on the *in-betweenness* of a plethora of actors, both human and non-human" (Roberge & Seyfert 2018: 2; Latour 1987). First, it is difficult not to acknowledge an intense division of labor within this domain of innovation—a situation that often translates into developers working on a dataset without fully knowing for whom, to which end and why. Here agency is divided amongst many little hands. Second, a networked approach would consider the actual implementation of deep learning tools and techniques as more in flock than in a row, adjusting to one another more than collaborating. This has been well documented in the literature on algorithmic finance, for instance, where competing stakeholders deploy "algotrading" tools to bolster attack or defense maneuvers (Seyfert 2018; Knorr-Cetina & Preda 2011; Castelle & al. 2016). Whether social sciences will be more attentive in the future to the combined, butterfly-like effects of all these actors' efforts into "boxing the box" remains to be seen. However, considering the understudied state of this phenomenon we urgently need to give more attention to the management, ordering and decision processes that shape what deep learning algorithms come to be about in the real world. More straightforwardly, the political economy of AI is one of our biggest and most opaque boxes today. In this chapter, we intend to contribute to the ongoing debate by analyzing what is at stake in this new form of socio-technical governmentality, i.e. what are the tensions, struggles, efforts at coopting knowledge, power, etc. Taking the Montreal AI hub as a case study, and following a 2016-2018

1 Of late, IBM has announced that it would allow access to a library comprising over two million images for facial recognition training with the hope that enhanced accuracy would help curb bias. The position of NGOs such as the *American Civil Liberty Union* in that case and in other similar ones is that better facial recognition is still bad news for minorities facing discrimination across a variety of social settings. See Browne 2019.

ethnographical investigation[2], we will focus on how stakeholders deploy multiple strategies and resources, including the building of legitimacy through symbolically-laden media operations. Our is thus empirical, while also network approach being theoretically informed; in that sense we hope to answer calls by preeminent scholars to develop critical thinking through studies which are *in situ* (Kitchin 2014; Mackenzie 2018).

I. Governmentality—What about it, and what Does it Change to the Study of Deep Learning?

Debates surrounding the increasing complexification of today's political economy and how it should translate into new understandings of power gain a great deal of intelligibility once one adheres to Michel Foucault's concept of governmentality (2004a, 2004b). This practice-oriented analysis of the 'problem of government' has allowed scholars to re-orient their focus on "the ceaseless transactions which, variably, modify, displace, upset, or insidiously shift the funding, the investment modalities, the centers of decision, the forms and types of control, the relations between local and central authorities, etc." (Foucault, quoted in Lascoumes 2004: 3; our translation.) Indeed, the French philosopher has had a pivotal role in identifying—at least—these three logics: i) how power and knowledge are inseparable, ii) how these introduce mobile and networked dynamics, and iii) how all of this allows to think about authority as enacted *by* and *as a* set of technologies. That said, the difficulty with Foucault is that he never properly wrote about the digital. Of late, it is Mackenzie who has endeavored to apply the concept to the study of what he calls 'machine learners', i.e. naive Bayes classifiers, decision trees, neural networks, and a range of others that fall under the broad category of AI (2013; 2017; 2018). All of this, according to him, corresponds to a "data practice that reconfigures local centers of power and knowledge by redrawing human-machine relations" (2017: 9). How research, development and implementation is organized; by whom, for what purposes, and through which means and discourses, is different from London to the Silicon Valley, or China and Canada for that matter. Likewise, how power relations and the distribution of authority are shaped specifically by the structure of its organizations and institutions varies from subfield to subfield—finance, military, transport, etc. Mackenzie is thus very helpful by providing such ecosystemic, if not ecological views. At the same time, in his book, he runs the risk of over-emphasizing an internal examination of the technology,

2 We totalized 12 interviews with machine learning specialists, 4 additional ones with scientific journalists, and over 400 articles from local francophone and anglophone newspapers and monthly publications.

and thus is only partially able to 'consider the boxing of the box'—to refer to Bucher once again.

What would it take to be able to provide insights that would be both local and 'architectural'—that is, able to demonstrate how particular constructions and transformations occur from the outside in? One such way is by looking at the distance between governmentality and what is deemed today as 'governance', and how, in fact, the latter is the topic of the former. In Montreal and probably elsewhere, the discourse related to governance enjoys great momentum, as the idea itself serves as a sort of empty-signifier that can tactically be given meaning. Governance, like progress or innovation, readily means "*good* governance" and many stakeholders involved in the construction of the Montreal hub conflate the two in order to bolster the institutional-public support for market-oriented developments in deep learning, the details of which will be presented shortly. For now, suffice to say that the very idea of a deep learning governance in Quebec's metropolis seeks to implicate pretty much everyone as "partners" in a game of collective self-management and purposive social change. In play is what scholars such as Walters have identified as "[an] emphasis on self-governing networks" drawing heavily on "the imagery of cybernetics and complexity theory" (2004: 29-30; see also Simard 1979 for a similar theoretical approach applied to Québec). Power and authority here are conceived as enablers: they allow for the circulation of resources, not for their constraint or restriction. As will be made clear below, everything related to ethics—the industry-backed Partnership on AI or the Montreal Declaration for a Responsible Development of AI—is tainted by an idea of self-regulation and its distinctive way of translating into a loose, sickly effort to *not* legislate. Power and politics have not disappeared for that matter; while governance might present itself in the best light, as lightweight *government at a distance*, the point is that it represents itself as an efficient, if understudied form of governmentality.

What is it about the Montreal deep learning hub that makes it worthy of scientific analysis? Part of the answer relates to the fact that Quebec is a rather small society, well developed but still marked by the concentration of its elites—social, political, economic, cultural, etc. As for the historical context in which the province has addressed the most recent "AI awakening", it is important to recall the role of Canada's CIFAR in subsidizing deep learning research, even when the technique was highly unfashionable (Hernandez 2014; Cardon & al. 2018; Engemann and Sudmann 2018). Star scientists such as Hinton (University of Toronto), former students Bengio (Université de Montréal) and to a lesser extent[3], LeCun (NYU and now Facebook) are both the inheritors and the best promoters of what is now a C$125 million pan-Canadian AI strategy. When, for instance, the talk of

3 LeCun has worked less in Canada and more in France and USA in recent years, although he still enjoys important media coverage.

an ongoing "AI revolution" emerged in Quebec's francophone mediascape, Bengio himself came to be introduced as one of the leaders of Montreal's AI ecosystem, itself presented as one of the most world-renowned hubs of cutting-edge AI innovation (See Bourgault 2017). This peculiar dynamic can be further demonstrated by the fact that its name appears in 126 of 161 articles focusing on AI developments published by Montreal newspaper *La Presse* between May of 2016 and July of 2017 (Bourgault 2017). The point here is that, when considering the Quebec's AI field, two correlations appear clearly: firstly, between the emergent rhetoric of a revolution and the rise of a charismatic leader; and secondly, between the accentuated hype surrounding deep learning and AI and the capacity of local actors to rapidly set in motion the relevant institutions. "Hype is low on informative content," scholar Guice rightly observes, "but directly states the relevance of the information to a social context" (1999: 85). In order to bolster the Montreal hub, former Quebec's Economy and Innovation Minister Anglade noted that her government would not "sprinkle" public investment (Rettino-Parazelli 2017). That led first to the creation of an advisory committee and, subsequently, of an AI Cluster initially equipped with a budget of C$100 million. The two most interesting facts about the cluster is that it devoted 80% of its funds to Bengio-directed, Université de Montréal-led MILA (Montreal Institute for Learning Algorithms) all while being officiated by Breton—Université de Montréal's dean—and well-known businessman Boivin, who several months later also became the head of MILA's board (see IA.Québec 2018). This suggests that in this particular context, and in this rather short period of time, what good governance meant was delivering efficiency; whereas a broader, more reflexive and critical perspective would instead have interrogated what it means in terms of circulating elites, and why the effort to maximize efficiency still needs to justify and legitimate itself through at least the appearance of duly-conducted administrative processes.

Another way of considering 'the boxing of the box' in the Montreal case is to have a look at the conjunction between efforts geared towards the launch of the aforementioned Montreal Declaration for a Responsible Development of AI and the creation, in late 2018, of the International Observatory on the Societal Impacts of AI. Fully endorsed by the government and its main scientific institutions, both make claims to an epistemological posture of "knowledge co-construction" with the public, the different stake-holders, etc., that in practice serves as a malleable, if not shallow, signifier. The Declaration, for instance, proposes a list of ten principles that are all more general and abstract than the other, with some overly naive or in contradiction with the current economic reality of deep learning—Principle 6.2 for instance states that "AI development must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge" (IA responsible 2017). For its part, the Observatory is still nascent, but in its very constitution already signals a poor understanding of social sci-

ences' role in studying social impacts, with for instance more members coming from computer sciences than sociology and communication studies altogether. Importantly, what the Observatory and the Declaration have in common is the cybernetic view of governance introduced above. On the one hand, the management of knowledge production obliterates any notion of checks and balances or arms-length regulatory principles, notions central to the very idea of modernity. On the other hand, it appears that all current virtue signaling efforts, including the Declaration, the Partnership for AI and the like, emerge as what Wagner calls "an escape from regulation" (2018).[4] All in all, the Quebec government's involvement in the development of its Montreal hub is one not of creating barriers and obstacles, but rather one to usher and foster the circulation of whatever is deemed 'positive', namely any twists and turns that exhibit a form of action from the government or the stakeholders, knowing that the legitimacy of who gives reflects on who receives and vice versa.

II. The 'Triple Helix' Remix and the Role of Open Science

At this point, it would be tempting to declare that, in spite of the initiatives of numerous actors, including significant gestures by the government of Quebec, it still is "business as usual". This, however, would be misleading in at least two separate ways. First, while it is accurate to say that the Cluster, the Declaration and the Observatory all participate in building a certain public perception of everything AI, it is not possible to adequate it to an ideology that would hide any sort of naked truth.[5] In other words, to be critical is to question how the box is made, not to put it on fire. Second, the expression 'business as usual' undercuts how much the advent of deep learning and related AI techniques is changing the power-knowledge topography of the province, notably the pivotal role universities are called upon to play. A governmentality approach must therefore be attentive to the structuration as well as the tensions involved here—which is also to say the historical and geographical subtleties that make higher education in Quebec something both North American but also profoundly influenced by the French universalistic approach,

4 Many have indeed noticed how increasingly frequent calls for "ethical AI" from industry figures often correlate with ongoing campaigns against "overly coercive" government regulation; see both Wagner, 2018 and Greene & al., 2019. A recent variation on this theme seems to be industry-backed regulations (Simonite, 2019); already, accusations of "regulatory capture" have been expressed (Biddle, 2019).

5 Such a tradition finds an emblematic figure in the early Habermas while he was for instance saying that "[…] [ideologies] replace traditional legitimations of power by appearing in the mantle of modern science and by deriving their justification from the critique of ideology. Ideologies are coeval with the critique of ideology" (Habermas 1971, 99).

and how, starting in the 60s and 70s, it made a substantial push towards a democratization of access.[6] Interestingly, Université de Montréal and McGill University, both at the forefront of MILA, are historically considered more 'elite' schools while still enjoying a great deal of public support. The MILA itself is important not only because it attracted most of the government-backed Cluster's money, but also as it comes to embody the displacement and, really, the refinement of what the literature calls the triple helix—a schematic model of innovation where corporate actors come to mesh with university and government ones (Etzkowitz & Leidesdorff 2000)—to now a "quadruple helix" where start-ups, too, are considered key strategic partners. Confusing small and large, it is not rare in Montreal to see international corporations such as Microsoft being equated with the local Maluuba, Facebook being considered as the emerging FAIR-MTL or Google as an embryonic DeepMind—with media celebrating even the smallest of investments.[7] All of this participates in what we argue is an ecological mentality that blurs the symbolic—a hub is positive by its very nature—and the practical, by the virtue of the latest trend in what Hoffman and others have called "academic capitalism" (2017; see also Slaughter and Rhodes 2010). In turn, the reality corresponds less to the early French influence on Quebec's higher education system than to a mode of "Silicon Valley-isation" or "Stanford-isation", terms borrowed from Salter (2018).

Common to all AI developments is the fact that they are guided by and inseparable from a specific ethos or model of "open science" (Leonelli 2013; Mirowski 2018). Researchers see the sharing of information as, *prima facie*, progress in and of itself; discovery and innovation are meant to be picked up by and benefit the entire "community" in what is thus an ecological as well as cybernetic mentality which, again, has roots in a certain Californian "rebelliousness".[8] Today, these norms prove to be very efficient, especially with regards to the following three dimensions. First, the obligation to choose to pursue either an academic career or a career in private R&D becomes less of an issue when one can publish freely, which is now allowed, if not encouraged in most basic research-inclined industrial labs—in fact, it is not rare to see papers co-authored by scientist at Facebook or DeepMind along with university-affiliated researchers. Knowing the shortage of qualified personnel in AI, such open science practices are thus instrumentally adopted

6 The ten institutions networked across the territory under the umbrella of Université du Québec is emblematic in that regard.

7 See for instance the summary of investments made to the local ecosystem in 2017 in Mathys 2017.

8 Here, we want to refer to what Saxenian (1994) and others have described as the characteristically innovative way Silicon Valley academic and industry actors had to produce new organizational forms at an impressive rate. On many accounts, this distinctive way of establishing collaborative ties between actors pertaining to different professional categories but to a common cultural background has spearheaded the privileged understanding of how to lead technological innovation these days—see also Storper & al. 2015.

as part of the repertoire necessary to navigate the "war to attract talent" (Hernandez & King 2016; Metz 2016a; 2017). Second, unrestricted circulation of people and ideas should allow for companies to track the best of university research. Internships, grants, and philanthropic donations large or small, contribute to secure access to computer science labs and to reach researchers where they are. For a city such as Montreal, this has proven very helpful, even if, from this decentralization and openness, it is impossible to conclude that its hub is a "plaque-tournante"—after all, others like Paris, Singapore, Pittsburg, etc., have benefited too. Thirdly, this openness is not only geographical, but temporal, as the adaptation between the different helixes, companies and university labs in particular, is intended to happen more or less in real-time. The pace of research here is as important as the commercial turnover rate that transforms an algorithmic architecture into an API, an innovation in a recommendation system, etc. While openness translates into windows of opportunity and good timing into fierce competition between companies, it is especially important to understand that the logic sustaining the entire model really is one of "strategic openness" (Ananny & Crawford 2018). The knowledge being produced in universities turns out to be "open for business" in a new and understudied sense, especially with regards to its wider implications in contexts such as Quebec.

The fact that deep learning and associated AI technologies signal a substantial displacement of wealth, prestige and power finds numerous and all the more empirical examples to which we will come in a few moments. For now, however, it appears that a necessary transition implies to question the broader significance of the "exploitable epistemology" (Levy & Johns 2016) set in motion through the quadruple-helix and open science nexus. As part of this research, a series of interviews with individuals involved in AI in Montreal were conducted, with most expressing largely consensual views, except for two or three more critical figures. The first one came from a computer scientist working in healthcare. Her critique points to structural elements in the transformation of research financing in Quebec and Canada:

> It's a concentration of millions of dollars, it's as if you're betting on a single number at the casino roulette. There's a variety of different types of research done, not all from the deep learning or big data strain [in AI] but that are also innovative—but you're not betting on them, you're only betting on deep learning. You're pushing everyone in the same direction and you forget that innovation is not necessarily of all going in the same direction. You also need to leave some to be sure that research in its totality is somewhat diversified. That, I see as a threat. It's going to siphon everything in the same direction […]. In fact, everyone is rushing into it.

This comment could serve as a proxy or hint as to how and why even scientific institutions partake in the kind of self-fulfilling prophecy that makes deep learning a reality. Again, institutions, hype and the pressure towards "Stanfordized" research go hand in hand. Another key example would be the attribution of 29 CIFAR Research Chairs late in 2018, some to prominent Facebook associates such as Pineault (McGill) or Vincent (Université de Montréal) (CIFAR 2018). For less trendy research streams, of course, this draws a path in which difficult access to funding would blend with its equivalent in terms of strenuous access to students—at least two other computer scientists in public universities from our samples talked about how they barely have any grad students nowadays. Looking at the longer term, chances are that the situation will become only more cyclical and detrimental.

Another related issue emerging with respect to the meaning of the "exploitable", even weaponized, epistemology implicated here, concerns the handling of databases: who owns them, how are they released, and for what purposes. For Big Tech companies as Google, who just open-sourced GPipe, or Microsoft, who acquired and now runs Github, gigantic libraries of data are acting in both performative and legitimating ways. Their flaws and limitations are scarcely if ever exposed—the fact for instance that such companies still pursue patents aggressively (Simonite 2018)—especially when compared to the ecological and cybernetic benefits attributed to these platforms and widely praised in the media. It is then at a more mezzo or local level that things get more challenging. The problem is that open data for training is not exactly the same as "real-deal" data or value-added data. For instance, our interviewee who works with deep learning applications in healthcare insisted that a dangerous dynamic is developing, where start-up businesses search for any sizeable bases to access in exchange for deep learning services, or at least, make a contract allowing them to share data with a third party. In places such as Montreal, to make a profit means finding clients—insurance, banks, clinics, biotech, etc.—not yet accustomed to deep learning techniques, in a legal environment still unsure about the best way to defend privacy or to regulate any potential wrongdoing.[9] Yet, it is probably at the micro level of the different university labs that the difference between the data "haves" and "have nots" is the most striking. Star researchers such as Bengio in Montreal—or, for that matter, Hinton in Toronto—attract funding and students because of their close connec-

9 An important parallel should be established with the way failed unicorn Theranos capitalized on the biotech industry's important regulatory leeway to position itself as one of the biggest (if short-lived) success stories of this emergent field. Its ability to rely on the reputability of early-backers such as Gen. James Mattis, Oracle founder Larry Ellison, media mogul Rupert Murdoch or present-day Secretary of Education Betsy Devos to sustain increasing investment rounds should be understood precisely as the result of the field's relative newness and its lack of proper regulations (O'Brien 2018).

tion with Google and the like. But what about the other, lesser-known researchers in the field? The conundrum is that they almost never gain access to the data actually prone to broad commercial applications. To give one final example in this section, our team met with another scientist in the summer of 2017 and talked about the general sense of community, and what it meant that open science was a way for private and public actors in the field to communicate. His answer was laconic: "It's just fake. It's just fake. They share the algorithm but not the data, you can do nothing with this. [As for the meaning of "open"], it's just a word because I cannot use it".

III. Deep Learning is Redefining the Private-Public Partnership

To say that today's developments in the AI field's political economy blur the preexisting distinctions between what is deemed private and public—or, for that matter, that it amounts to a "Stanfordisation" of higher education in places such as Quebec—is not to succumb to any nostalgia for a utopian past. A descriptive and agnostic approach is indeed needed to account for, as Hoffman stated, "the complicated, subtle, and sometimes contradictory ways that commercial logics have diffused across academic culture" (2017: 727). The point is that, in Montreal and most probably elsewhere, ambiguity is in itself a form of governmentality. Weakened institutional autonomy is translated into more collaboration; buzzwords in the semantic region of "hub", "clustering", "ecosystem" and the likes are repeated and celebrated in what is then hard to decipher from public relations endeavours (see Turkina 2018, for instance). A turning point of this development was the January 2019 relocation of the MILA to Mile-Ex, a post-industrial inner-city in Montreal. The relocation of the lab occurred as it got elevated to the status of "Quebec Artificial Intelligence Institute" and came to be positioned at the forefront of the Mile-Ex's Cité de l'IA, with multiple small and big companies establishing their new facilities either in the same building or in its immediate surroundings. O Mile-Ex, the converted textile-manufacture the lab moved in, already accommodated the offices of up-and-coming startup Element AI, Royal Bank of Canada's AI branch Borealis, French military contractor Thales' AI research division and the para-public Institute for Data Valorization (IVADO), with Microsoft's Maluuba also a close neighbour (Bachand 2018; Dubuc 2018). Importantly, the idea to create the Cité represents the fourth pillar of the government's strategy in everything AI—along with the Industrial Cluster, the Declaration and the Observatory—yet, because of its weight in terms of jobs, investments, square feet of office space and the like, it is possible to argue that it is the most important. The people in charge there understand rather well the leverage associated with their interstitial position. Indeed, in interviews with media about the relocation, they were keen to ask

for additional public funding: "Attracting researchers to Montreal by telling them we only have two years of funding left, that won't work. We need to be part of a broader, longer-term vision. We're in the order of tens and hundreds of millions" (Pisano quoted in Rettino-Parazelli 2019).

In terms of practical, yet non-official public-private blending, there might be no equivalent in Quebec to Element AI, the fast-growing company co-founded by MILA's director Bengio. As rightly expressed by one media commentator, "the business model is not easy to understand" (quoted in Mercure 2016); not only did it attracted historic amounts of venture capital without a proper product on the market, but it continually operates under an ethics-oriented discourse of public good and social benefits while also positioning itself as an active player in the rather traditional and profit-savvy fields of logistics, insurance and banking (The Economist 2017). Bengio himself appears willing to play on both levels as he dedicates genuine efforts to promote an ethical and socially-minded development of his field while lionizing the commercial success of his company, one apparently set to become one of the first Canadian AI Unicorn (George-Cosh 2018; Vara 2018). In addition, he sometimes confuses his own numerous public and private affiliations in talks, Power Points and elsewhere, in what is now emblematic of a bigger issue, namely how the value and wealth created in the public domain tends to move away from it. The very nature of Element AI—and part of the reason for its initial valuation—is to capitalize on its access to star academics to develop 'business solutions' for its private-sector clients. As acclaimed in the Journal of Small Business & Entrepreneur, the company has "a faculty fellow network composed of over 20 world-renowned AI scientists from the top academic labs across Canada. These professors not only do research-related work for the company but […] provide valuable advice […]. This unique arrangement gives Element AI access to cutting-edge research" (Turkina 2018: 2). Again, what there is in this quotation relates to everything cybernetic about the new model being implemented discussed above. Helixes rotate, openness signals access, pace equals circulation and innovation, etc., in a movement that is certainly difficult, yet not impossible, to track.

While Element AI is cybernetic by essence, it is as well ecological in a very practical way. Proximity to the MILA shapes the urban space around it, and could be measured in meters. Of course, such proximity is not something to be found only in the Cité de l'IA; numerous incubators in North America, Europe and elsewhere use the model with the justification that it contributes to the cross-pollination of ideas and resources. The problem, however, is slightly different when it comes to the blurring of public and private assets. Emblematic in that regard is a Facebook post by Element AI saluting the arrival of MILA "to the neighborhood", which showcased a picture of Bengio while emphatically adding, "see you in the stairwell" (Element AI 2018). Such metaphor usually refers to a more or less licit space; one with more or less fuzzy codes and boundaries. Who goes up, what goes

down, when, and how? In the case particularly of students-becoming-interns-becoming-students, the lack of explicit limitations never cease to be problematic both on an individual and on a cohort basis. Once aggregated, these public and private part-time or twofold affiliations reinforce a model that is poorly checked and balanced, especially in the face of its long-lasting socio-political and economic impacts and how these could be discussed and amended in the public sphere.

At current, it is such public-private intermingling that comes to colonize the many layers of the AI Montreal hub—despite certain pleas from Bengio against a reality he actually contributes to.[10] The recent wave of investments made by foreign corporations in the Montreal hub has been going hand-in-hand with the increasing adoption by the newly 'partnered' scientists of this new organizational arrangement, namely, the dual affiliation model. This university-to-industry collaborative form, imported from the fields of law, management and medicine—and probably at its strongest in the biotechnology industry; see Mirowski 2012—allows scholars to keep their university professorship appointment while adding to it a commitment, on at least a part-time basis, to their new corporate employer (Serebrin 2017a; 2017b). To the list parsed throughout this chapter, we should still be adding the many names of MILA-affiliated scientists such as Larochelle at Université de Montréal and Google Brain, Precup at McGill and DeepMind or Pal at Element AI and Université de Montréal. Dual affiliation is justified by actors of the field as a novel solution where scientists are able to continue teaching and conduct basic research while also participating in industrial R&D, whereas previously, such participation would entail a complete retreat from their university teaching and basic research activities (Plamondon Emond 2017). The growing dissemination of the model thus operates at the junction of two distinct but concomitant dynamics. On the one hand, corporate actors are increasingly aware of the necessity, for their business model, to achieve an all-essential balancing-act between the preservation of the "ecosystem sustainability"—i.e., to ensure the continued formation of future generations of AI researchers and the further advancement of basic research endeavours (LeCun 2018)—and, as described in section II, the conflicting urge of immediate appropriation of specialized human resources (Metz 2016a; 2017). On the other hand, scientists are responding to constraints which are mostly presented as incentives: besides the alluring possibility of alternative sources of private funding, researchers also have to deal with the fact that access to state-of-the-art corporate computational infrastructures and some of the widest proprietary databases are indeed technological means increasingly needed for the pursuit of cutting-edge deep learning research. In turn, this new public-pri-

10 See Shead 2018. On his criticizing the increasing concentration by major tech corporations of both technological means and specialized human resources while scarcely rejecting opportunities to collaborate with them—see Mathys 2017 and Vara 2018.

vate assemblage and its peculiar way of providing many solutions at once is presented, justified and legitimized as a form of necessity, and not as a cascade of contingent choices and principles. In that sense, however, it is nonetheless highly political and a form of governmentality.

Conclusion

This chapter started by acknowledging how the black box is a powerful, yet disenchanted figure to reflect on technologies *in the making* such as deep learning, and AI more broadly. While there is an urgency that should spur social sciences inquiry, it is nonetheless important to do things right, with a certain dose of agnostic and critical reflexivity. In this light, we attempted to follow Bucher's triple advice to 'not fear the black box'; 'not expect the solution to be inside' and 'consider the boxing of the box'. So how did it go? How did the theoretical concepts apply to the practical reality and how, in turn, can better understanding of a case such as the Montreal hub inform broader and more critical reflection? Parts of the answer came in section I, where the argument was made that what is mostly at stake is the present and future political economy of AI, i.e. how the automation of knowledge production transforms power relations and how the different actors involved in deep learning are engaged in what Crandall names a particular form of "cooperative struggles" (2010). Substantial resources including money, state support, media coverage, etc. are flowing and aggregating, the details of which are precisely what must be understood about this dense and tense regime of governmentality. The new normal brought about by deep learning and AI-related technologies will be messy and ambivalent, if this is not already the case. We insisted throughout the chapter that power is more than ever a transaction and that what "control" means in these circumstances relates to a new sense of cybernetics and ecology that shall account for all types of mutualism and parasitism. In section II, we described this by digging into the Montreal example, especially how it exhibits a peculiar form of rotary motion between the helixes that are the governmental, university, established and upcoming corporate actors. Whereas actors repeatedly proclaim there is a "community" and that the Cluster, the Observatory, the Declaration and the *Cité de l'IA* make for an integrated whole, we propose a somewhat less optimistic, more realistic analysis. Open science is a case in point, as "open" translates into aggregation and as it signals an important shift in the educational model in vogue. There is such thing as a privatisation of higher education in place like Quebec, in which deep learning and AI related technologies are instrumental. This was the principal conclusion of section III. Whether you call it the "double affiliation" or the "see you in the staircase" model, what is clear is that the benefits are not equally redistributed at current and have very poor chance—at least the-

oretically—of being so in the future. In the end, it might then be such unfolding, in its many twists and turns, that constitutes the proper object of another, still-in-the-making ecology, one which could be called Critical AI Studies.

References

Ananny, Mike, and Crawford, Kate (2018 [2016]): "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20/3, pp. 973-89. https://doi.org/10.1177/1461444816676645.

Bachand, Olivier (2018): "Une Cité de l'intelligence Artificielle Dans Le Mile-Ex à Montréal." Radio-Canada, retrieved from https://ici.radio-canada.ca/nouvelle/1081709/technologie-montreal-saint-urbain-intelligence-artificielle (accessed March 11, 2019).

Biddle, Sam (2019):"Should We Trust Artificial Intelligence Regulation by Congress If Facebook Supports It?" The Intercept, retrieved from https://theintercept.com/2019/03/07/artificial-intelligence-facebook-ibm-ro-khanna/ (accessed March 9, 2019).

Bijker, Wiebe E., Hugues, Thomas Parke, and Pinch, Trevor (eds.) (2012 [1987]): *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Mass: MIT Press

Bourgault, Mathieu (2017): "Synthesis on the Quebec's AI Mediascape". *Unpublished Manuscript*, pp. 1-30.

Browne, Ryan (2019): "IBM Hopes 1 Million Faces Will Help Fight Bias in Facial Recognition." CNBC, retrieved from https://www.cnbc.com/2019/01/29/ibm-releases-diverse-dataset-to-fight-facial-recognition-bias.html (accessed March 9, 2019).

Bucher, Taina (2016): "Neither Black Nor Box: Ways of Knowing Algorithms." In *Innovative Methods in Media and Communication Research*, edited by Sebastian Kubitschko and Anne Kaun, pp. 81-98. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40700-5_5

Burrell, Jenna (2016): "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society*, 3/1, pp. 1-12. https://doi.org/10.1177/2053951715622512.

Cardon, Dominique, Cointet, Jean-Philippe, and Mazières, Antoine (2018): "La revanche des neurones: L'invention des machines inductives et la controverse de l'intelligence artificielle." *Réseaux* 211/5, pp. 173-220. https://doi.org/10.3917/res.211.0173

Castelle, Michael, Millo, Yuval, Beunza, Daniel, and Lubin, David C. (2016): "Where Do Electronic Markets Come from? Regulation and the Transforma-

tion of Financial Exchanges." *Economy and Society* 45/2, pp. 166-200. https://doi.org/10.1080/03085147.2016.1213985.

CIFAR (2018): "29 Researchers Named to First Cohort of Canada CIFAR Artificial Intelligence Chairs." CIFAR, retrieved from https://www.cifar.ca/cifarnews/2018/12/03/29-researchers-named-to-first-cohort-of-canada-cifar-artificial-intelligence-chairs (accessed March 9, 2019).

Crandall, Jordan (2010): "The Geospatialization of Calculative Operations: Tracking, Sensing and Megacities." *Theory, Culture & Society*, 27/6, pp. 68-90. https://doi.org/10.1177/0263276410382027.

Dubuc, André (2018): "Microsoft Déménage Ses 200 Employés Dans Le Mile-Ex." La Presse, retrieved from http://plus.lapresse.ca/screens/60fda222-94fb-49ee-ac0a-05a933c166a0__7C___0.html?utm_medium=Twitter&utm_campaign=Internal+Share&utm_content=Screen (accessed March 11, 2019).

Element AI (2018): "See you in the stairwell". *Facebook Post*, January 29, retrieved from https://www.facebook.com (accessed March 9, 2019).

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

Etzkowitz, Henry, and Leydesdorff, Loet (2000): "The Dynamics of Innovation: From National Systems and 'Mode 2' to a Triple Helix of University–Industry–Government Relations." *Research Policy* 29/2, pp. 109-23. https://doi.org/10.1016/S0048-7333(99)00055-4.

Foucault, Michel (2004a): "Naissance de La Biopolitique." In *Cours Au Collège de France 1978-1979*. Paris: Gallimard-Seuil.

Geiger, R Stuart (2017): "Beyond Opening up the Black Box: Investigating the Role of Algorithmic Systems in Wikipedian Organizational Culture." *Big Data & Society* 4/2, pp. 1-14. https://doi.org/10.1177/2053951717730735.

George-Cosh, David (2018): "Element AI Aims for Unicorn Status with Record Canadian Financing: Sources." BNN Bloomberg, retrieved from https://www.bnnbloomberg.ca/element-ai-aims-for-unicorn-status-with-record-canadian-financing-sources-1.1101206 (accessed March 9, 2019).

Gillespie, Tarleton (2014): "Algorithm [Draft] [#digitalkeywords]." *Culture Digitally* (blog), retrieved from http://culturedigitally.org/2014/06/algorithm-draft-digitalkeyword/ (accessed March 9, 2019).

—(2017): "Regulation of and by Platforms." In *SAGE Handbook of Social Media*, edited by Jean Burgess, Thomas Poell and Alice Marwick, pp. 254-78. London: Sage.

Greene, Daniel, Hoffman, Anna Lauren and Stark, Luke (2019): "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." Proceedings of the 52nd Hawaii International Conference on System Sciences.

Guice, Jon (1999): "Designing the Future : The Culture of New Trends in Science and Technology." *Research Policy*, 28, pp. 81-98.

Habermas, Jürgen (1971): *Toward a Rational Society: Student Protest, Science and Politics*. London: Heinemann.

Hernandez, Daniela (2014): "Meet the Man Google Hired to Make AI a Reality." Wired, retrieved from https://www.wired.com/2014/01/geoffrey-hinton-deep-learning/ (accessed March 9, 2019).

Hernandez, Daniela, and King, Rachael (2016): "Universities' AI Talent Poached by Tech Giants." *Wall Street Journal*, retrieved from http://www.wsj.com/articles/universities-ai-talent-poached-by-tech-giants-1479999601 (accessed March 9, 2019).

Hinton, Geoffrey, Osindero, Simon and The, Yee-Whye (2006): "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18/7, pp. 1527-1554.

Hoffman, Steve G. (2017): "Managing Ambiguities at the Edge of Knowledge: Research Strategy and Artificial Intelligence Labs in an Era of Academic Capitalism." *Science, Technology, & Human Values* 42/4, pp. 703-40. https://doi.org/10.1177/0162243916687038.

Hutson, Matthew (2018): "Artificial Intelligence Faces Reproducibility Crisis." *Science* 359/6377, pp. 725-726. https://doi.org/10.1126/science.359.6377.725.

IA.Québec (2018): "Effervescence de l'intelligence Artificielle." Comité d'orientation de La Grappe En Intelligence Artificielle, retrieved from http://ia.quebec/ (accessed March 9, 2019).

IA responsable (2017) : "Déclaration de Montréal Pour Un Développement Responsable de l'IA." Déclaration de Montréal IA Responsable, retrieved from https://www.declarationmontreal-iaresponsable.com/ (accessed March 9, 2019).

Kitchin, Rob (2014): "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1/1, pp. 1-12. https://doi.org/10.1177/2053951714528481

Knorr-Cetina, Karin and Preda, Alex (eds.) (2011 [2006]): *The Sociology of Financial Markets*. Oxford: Oxford University Press.

Lascoumes, Pierre (2004): "La Gouvernementalité : De La Critique de l'État Aux Technologies Du Pouvoir." *Le Portique*, no. 13-14, pp. 1-16.

Latour, Bruno (1987) : "Science in Action: How to Follow Scientists and Engineers through Society" Cambridge, Mass: Harvard University Press.

LeCun, Yann (2018): "Facebook's Chief AI Scientist Says That Silicon Valley Needs to Work More Closely with Academia to Build the Future of Artificial Intelligence." Business Insider, retrieved from https://www.businessinsider.com/facebook-yann-lecun-dual-affiliation-model-ai-experts-2018-8 (accessed March 9, 2019).

Leonelli, Sabina (2013): "Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production and the Political Economy of Contemporary Biology." *Bulletin of Science and Technology Studies* 33/1-2, pp. 6-11. https://doi.org/10.1177/0270467613496768.

Levy, Karen EC, and Johns, David Merritt (2016): "When Open Data Is a Trojan Horse: The Weaponization of Transparency in Science and Governance." *Big Data & Society* 3/1, pp. 1-6. https://doi.org/10.1177/2053951715621568

L'heureux, Alexandra, Grolinger, Katarina, Elyamany, Hany F. and Capretz, Miriam A. M. (2017): "Machine Learning with Big Data: Challenges and Approaches." *IEEE Access* 5, pp. 7776-7797.

Litvinski, Oleg (2018): "On Social Mechanisms of Algorithmic Opacity." *Unpublished Manuscript*, pp. 1-32.

Mackenzie, Adrian (2013): "Programming Subjects in the Regime of Anticipation: Software Studies and Subjectivity." *Subjectivity* 6/4, pp. 391-405.

— (2017): *Machine Learners: Archaeology of a Data Practice*. Cambridge, MA: The MIT Press.

— (2018): "From API to AI: Platforms and Their Opacities." *Information, Communication & Society*, June, pp. 1-18. https://doi.org/10.1080/1369118X.2018.1476569.

Mathys, Catherine (2017): "Le Rebelle de l'intelligence Artificielle." L'actualité, retrieved on http://lactualite.com/techno/2017/11/10/le-rebelle-de-lintelligence-artificielle/ (accessed March 9, 2019).

Mercure, Philippe (2016): "Montréal, Future Plaque Tournante de l'intelligence Artificielle?" La Presse, retrieved from https://www.lapresse.ca/affaires/economie/201610/26/01-5034411-montreal-future-plaque-tournante-de-lintelligence-artificielle.php (accessed March 9, 2019).

Metz, Cade (2016a): "Giant Corporations Are Hoarding the World's AI Talent." Wired, retrieved from https://www.wired.com/2016/11/giant-corporations-hoarding-worlds-ai-talent/ (accessed March 9, 2019).

— (2016b): "GOOGLE OPENS MONTREAL AI LAB TO SNAG SCARCE GLOBAL TALENT." Wired, retrieved from https://www.wired.com/2016/11/google-opens-montreal-ai-lab-snag-scarce-global-talent/ (accessed March 9, 2019).

— (2017): "FOR GOOGLE, THE AI TALENT RACE LEADS STRAIGHT TO CANADA." Wired, retrieved from https://www.wired.com/2017/03/google-ai-talent-race-leads-straight-canada/ (accessed March 9, 2019).

Mirowski, Philip (2012): "The Modern Commercialization of Science Is a Passel of Ponzi Schemes." *Social Epistemology* 26/3–4, pp. 285-310.

— (2018): "The Future(s) of Open Science." *Social Studies of Science* 48/2, pp. 171-203. https://doi.org/10.1177/0306312718772086.

O'Brien, Sara Ashley (2018): "Elizabeth Holmes Surrounded Theranos with Powerful People." CNN Business, retrieved from https://money.cnn.com/2018/03/15/technology/elizabeth-holmes-theranos/index.html (accessed March 9, 2019).

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

Plamondon Emond, Étienne (2017): "Intelligence Artificielle: Un Pied à McGill et l'autre Chez Un Géant Des Technos." Le Devoir, retrieved from https://www.

ledevoir.com/societe/education/511873/intelligence-artificielle-un-pied-a-mcgill-et-l-autre-chez-un-geant-des-technos (accessed March 9, 2019).

Rettino-Parazelli, Karl (2017): "L'intelligence Artificielle, Moteur Économique." Le Devoir, retrieved from https://www.ledevoir.com/economie/500340/dominique-anglade-au-devoir (accessed March 9, 2019).

—(2019): "Nouveaux Bureaux, Nouveaux Besoins Pour Mila." Le Devoir, retrieved from https://www.ledevoir.com/economie/546567/nouveaux-bureaux-nouveaux-besoins?fbclid=IwAR3XDUBbIUrIqnGIt5oXUfZTdp8GqYtY6kbVQsGl8pp_slKzpPQFgv_s2IQ (accessed March 9, 2019).

Roberge, Jonathan and Seyfert, Robert (2018 [2016]). "What are algorithmic cultures?" in *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies*, Robert Seyfert and Jonathan Roberge (eds.), pp. 1-25, New York: Routledge.

Rosenblatt, Frank (1958): "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65/6, pp. 386-408.

Salter, Christopher (2018). Personal communication, *"AI Talks: The New Silicon Valley of the North, Really?"*, December 4, Montreal, Concordia University.

Saxenian, AnnaLee (1994): *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, Mass: Harvard University Press.

Serebrin, Jacob (2017a): "Facebook to Open AI Lab in Montreal Headed by McGill Professor." Montreal Gazette, retrieved from https://montrealgazette.com/news/local-news/facebook-to-open-ai-lab-in-montreal-headed-by-mcgill-professor (accessed March 9, 2019).

— (2017b): "Google-Linked AI Company to Open Research Lab in Montreal." Montreal Gazette, retrieved from https://montrealgazette.com/business/local-business/google-affiliated-ai-company-deepmind-to-open-research-lab-in-montreal (accessed March 9, 2019).

Seyfert, Robert (2018): "Automation and Affect: A Study of Algorithmic Trading." In *Affect in Relation—Families, Places, Technologies. Essays on Affectivity and Subject Formation in the 21th Century*, Birgitt Röttger-Rössler, Jan Slaby (eds.), pp. 197-218. New York: Routledge.

Sudmann, Andreas (2018): "On the Media-political Dimension of Artificial Intelligence: Deep Learning as a Black Box and OpenAI.", in: Rethinking AI. Neural Networks, Biopolitics and the New Artificial Intelligence, *Digital Culture & Society* 4/1, pp. 181-200. (available on academia.edu)

Shead, Sam (2018): "AI Pioneer Yoshua Bengio Says Universities Deserve More Credit." Forbes, retrieved in https://www.forbes.com/sites/samshead/2018/11/13/ai-pioneer-yoshua-bengio-says-universities-deserve-more-credit-for-their-ai-research/#5cb7d54273ab (accessed March 9, 2019).

Simard, Jean-Jacques (1979): "Québec et frères, inc. La cybernétisation du pouvoir." *Recherches sociographiques*, 20/2, pp. 239-261. https://doi.org/10.7202/055840ar.

Simonite, Tom (2018): "Despite Pledging Openness, Companies Rush To Patent AI Tech." Wired, retrieved from https://www.wired.com/story/despite-pledging-openness-companies-rush-to-patent-ai-tech/ (accessed March 9, 2019).

— (2019): "AMAZON JOINS MICROSOFT'S CALL FOR RULES ON FACIAL RECOGNITION." Wired, retrieved from https://www.wired.com/story/amazon-joins-microsofts-call-rules-facial-recognition/ (accessed March 9, 2019).

Slaughter, Sheila, and Rhoades, Gary (2010): *Academic Capitalism and the New Economy: Markets, State, and Higher Education*. Baltimore: Johns Hopkins University Press.

Storper, Michael, Kemeny, Thomas, Makarem, Naji Philip, and Osman, Taner (2015): *The Rise and Fall of Urban Economies: Lessons from San Francisco and Los Angeles*. Innovation and Technology in the World Economy. Stanford, California: Stanford Business Books.

The Economist (2017): "A Hybrid Startup Offers AI Services to Business." The Economist, retrieved from https://www.economist.com/business/2017/06/22/a-hybrid-startup-offers-ai-services-to-business (accessed March 9, 2019).

Turkina, E. (2018): "The Importance of Networking to Entrepreneurship: Montreal's Artificial Intelligence Cluster and Its Born-Global Firm Element AI." *Journal of Small Business & Entrepreneurship* 30/1, pp. 1-8. https://doi.org/10.1080/08276331.2017.1402154.

Vara, Vauhini (2018): "Can This Startup Break Big Tech's Hold on A.I.?" Fortune, retrieved from http://fortune.com/longform/element-ai-startup-tech/ (accessed March 9, 2019).

Wagner, Ben (2018): "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?" In *Being Profiling. Cogitas Ergo Sum.*, Amsterdam University Press, pp. 1-7. M. Hildebrandt.

Walters, William (2004): "Some Critical Notes on 'Governance.'" *Studies in Political Economy* 73/1, pp. 27-46. https://doi.org/10.1080/19187033.2004.11675150.

Reduction and Participation

Stefan Rieger

> Three years ago, researchers at the secretive Google X lab in Mountain View, California, extracted some 10 million still images from YouTube videos and fed them into Google Brain—a network of 1,000 computers programmed to soak up the world much as a human toddler does. After three days looking for recurring patterns, Google Brain decided, all on its own, that there were certain repeating categories it could identify: human faces, human bodies and … cats. (Jones 2014: 146)

1. Deep Learning

By this point, talk of the omnipotence of algorithms is everywhere. This discourse proceeds without interruption and is seemingly impossible to stop—not least because algorithms operate quietly and inconspicuously in the background (cf. Bunz 2012; Seyfert/Roberge 2017). Many of the discussions about their influence concern the status of their opacity and, by concentrating on the refusal of firms to make them transparent, bring arguments into play that seem like relics from another era. Whereas then the focus of critics rested on the activities of a discredited culture industry, today it is the economization of hitherto unimaginable volumes of data that is considered a violation. The economic valence of data has become the object of a media critique that lost one of its favorite subjects from the previous century: the critical and autonomous media user (or that which was once regarded as such). The algorithms of large corporations such as Google, Amazon, or Facebook rightly seem to have subsumed the latter subject's potential for action, autonomy, resistance, and subversion (cf. Sudmann 2017). This process has been so successful that it has even led to counter-movements that do not casually lament the end of the private sphere as collateral damage of digitalization but have rather adopted agendas that enthusiastically promote its undoing (cf. Rieger 2018). For the internet exhibitionists of the so-called *Post-Privacy Spackeria*, data protection is nothing more than a historically datable remnant, a vestige from the last millennium: "The private sphere is so 1980s." (Reißmann 2019, n. pag.)

The areas of application for the use of algorithms, which, for their part, have been the object of a brief evolution and whose optimization has been oriented not least

toward meeting the specifications of nature, are ubiquitous and so varied that they cannot be surveyed in full:[1] whether recognizing faces in everyday life for reasons of delayed surveillance or future-oriented forensics, identifying sequences of behavior or engaging in biopolitics, clarifying the authorship of images and texts (cf. Rodriguez et al. 2019; Rehman et al. 2019), classifying works of art according to the style of a given epoch or comparing signatures supposedly written by the same hand, intervening in the business of science and confronting apparently non-computable objects of knowledge with big data and algorithmization (cf. Rieger 2019), affecting the self-perception and self-assessment of certain disciplines over the course of the "computational turn" and "humanities computing," associating the latter disciplines with different forms of reflection and thereby contributing fundamental changes within the humanities itself (cf. Hall 2013), or otherwise intervening in the order of things—such activity typically draws upon processes of artificial intelligence, artificial neural networks, and deep learning. Their manner of dealing with large volumes of data has become a knowledge-promoting game and has even opened up new possibilities for Foucauldian discourse analysis, which is seldom applicable to technological developments (cf. Engemann/Sudmann 2018). The possibilities of artificial intelligence play right into the hands of Foucault's basic intuition that "empirical knowledge, at a given time and in a given culture, *did* possess a well-defined regularity" and that "the history of non-formal knowledge had itself a system." (Foucault 2002 [1966]: x) Over the course of his book *The Order of Things*, Foucault sought to reveal an epistemologically stringent (but, in technical terms, hardly realizable) *positive unconscious of knowledge* and thus to give expression to the supposition that there is a "well-defined regularity"—a formal code behind non-formal knowledge as well. It would therefore be possible to process the science of this knowledge in a different way: it could become the object of an algorithmic discourse analysis and remain removed from individual understanding and comprehension. In the modes of access employed by cultural analytics, such a positive unconscious of knowledge is brought up to technical speed and made visible in the form of regularities and repetitions. Data mining and text mining make patterns and thus forms of knowledge visible that are not necessarily exhausted in intentional questions. Here, everything that human intelligence, in its scientific narcissism, regards as its genuine field of activity—ordering and classifying things, identifying similarities, and creating genealogies—is relegated to algorithms. In this case, the business of science is therefore not at the mercy of chance in its efforts to produce knowledge; rather, identities and differences are processed automatically—with algorithmic and not anthropogenic support.

Yet this concerns not only the sciences, with their broad subject areas and the claim to complexity associated with them. The activity of algorithms even extends to

1 The keywords in question would be evolutionary algorithms, evolutionary or genetic programming.

the lower senses, which, for long stretches, received hardly any attention in cultural history but have since come into the spotlight thanks to the efforts of various naturalization movements (cf. Kortum 2008). Like almost everything else, the detection of smells can also be delegated to algorithms—with the effect that, where olfactory data can be processed automatically in large quantities and at high speeds (*in real time*, to use one of the favorite terms of several protagonists), a familiar danger looms. In the case of smells, this danger has been called "odorveillance." In addition to seeing everything, Jeremy Bentham's panopticon can now smell everything as well (cf. Stark et al. 2018a). The consensus over this seems to be that such a regime of odors should be regarded as an outgrowth of other biometric activities and should accordingly be opposed. Of course, the following is just a rhetorical question: Is this sort of odorveillance really what we want? (Stark et al. 2018b: 18) And there also seems to be a consensus over the fact that automated activities of this sort should be the object of fundamental reflection concerning the nature of "veillance" in all of its varieties (the latter now include "sousveillance" and "metaveillance") (cf. Kammerer/Waitz 2015). Indeed, this idea has even been spelled out in a programmatic way—in works with titles such as "Declaration of Veillance (Surveillance is a Half-Truth)" (Mann 2015).

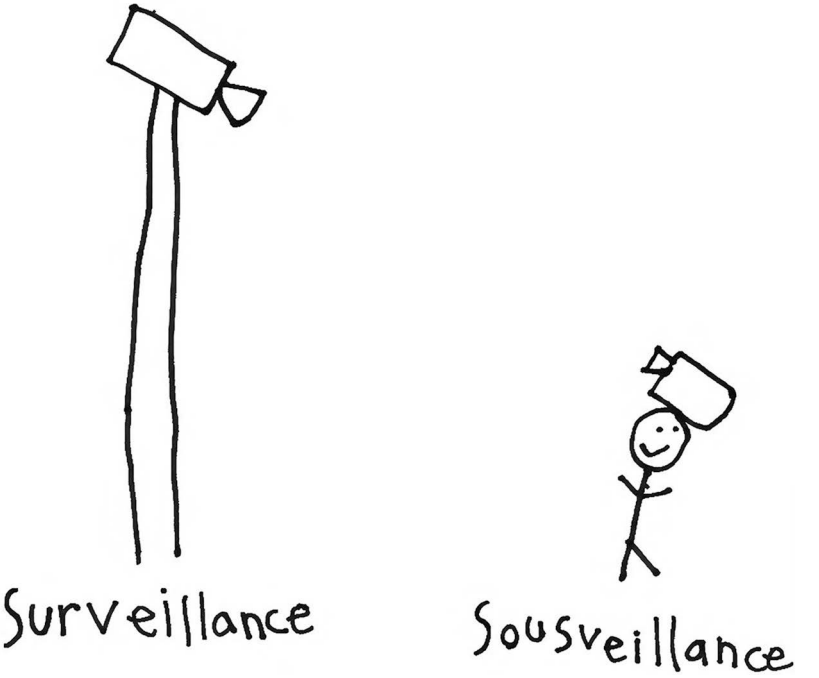


Fig. 1: Surveillance versus Sousveillance (https://en.wikipedia.org/wiki/File:SurSousVeillanceByStephanieMannAge6.png, accessed June 4, 2019)

Protagonists such as Steve Mann (2016) or José van Dijck (2014) should be mentioned here, the former for introducing concepts of veillance beyond surveillance, the latter for his concept of datafication, which describes the normalization of data politics and its ambit as a new sort of currency. With datafication and its basic suspicion concerning the opaque modus operandi of such data processing, the media-critical impetus of earlier days seems to have survived and not to have capitulated to the demands for a total relinquishment of the private sphere. In his book *Post-Privacy: Prima leben ohne Privatsphäre*, for instance, the internet activist Christian Heller comes to appreciate the latter, even though there are arguments in favor of its complete abandonment. He cites an example of algorithms being able to determine the sexual orientation of individuals from their social behavior—without any regard, of course, for the safety of the people in question:

> His sexual orientation is private, and so it should remain. However, he created his account without considering the inventors at the Massachusetts Institute of Technology (MIT). There a process was developed for identifying, with a high probability, the homosexuality of men on the basis of their Facebook profile, even if they posted no photos or listed no preferences of any sort. All that is needed is to analyze their social environment on Facebook, which is used above all to stay in touch with friends, relatives, and acquaintances. Often enough, profiles include a list of friends that is visible to anyone in the whole world (it is possible to make this information private, but few bother to do so). The researchers at MIT discovered that it is possible to make approximate predictions about whether a male student is gay on the basis of the portion of men among his Facebook friends who have outed themselves as gay on their own profiles (Heller 2011: 12).

2. Strategies of Participation

Artificial intelligence is at work everywhere, regardless of whether we know it, whether we can know it, or whether we even want to know it. The concerns of surveillance studies or critical code studies aside, moreover, everyday user behavior is often defined by a fundamental and reckless indifference to the activities of algorithms and issues of security. As is clear not only from people's access codes and passwords (the use of easily decipherable sequences of numbers, birthdays, the nicknames relatives and pets) but also from their willingness to disclose their consumer preferences and other habits, this behavior exemplifies flippant and careless negligence. Yet there is another aspect that defines what is going on and increasingly determines how we engage with artificial intelligence, one that is perhaps less visible and at first glance far removed from concrete political action. Whereas algorithms are monopolizing autonomy everywhere, whereas they operate in a self-determined or partly self-determined way, whereas they are exe-

cuting the grand scheme of automated knowledge with greater and greater efficiency and on hardware that is ever increasing in capacity, and whereas—as one repeatedly reads—they conduct their business without notice and in the mode of operative latency, a peculiar counter-movement is taking place on the level of use and participation, social engagement, and the campaign for acceptance. This process is peculiar because it seemingly overturns the order of the grand narrative that surrounds technology in general and digitalization in particular. The grand narrative about the technology around us and the associated politics of the internet is typically bound to a principle of quantitative growth. This can be narrated in the form of large numbers and is written as the history of progress of an utterly relentless triumph of increasing complexity.

It is thus all the more striking to see tendencies in dealing with technical (or perhaps it would be better to say socio-technical) infrastructures that move in a different direction and are based on the opposite of growth—that is, on what will be discussed here under the title "Reduction and Participation." This interruption of the customary success story and the intentional reduction of technically possible complexity are noteworthy—and in various ways they revolve around aspects of internet politics, democratization, and the question of who should have access at all (and in what way). What is especially remarkable is a fundamental expansion of that which is considered fit for participating on the internet and thus for being addressed. Over the course of this expansion, as will be shown, different and additional agents have been put in position to participate—agents who are situated outside of the dominant concerns of human-computer interaction (HCI) and who endorse the argument for reduction or at least provide some indication of the gestures associated with it. Those who have somewhat systematically become part of the plan include users who, with their specific profiles, veritably embody the issue of reduction. These particular users are phenotypically diverse and thus, not least, children and people with challenges have attracted increasing attention as extreme cases of those with special user profiles: "Alterations of HCI methods is common when interaction design is planned for 'extreme' human users." (Hirskyj-Douglas et al. 2016, n. pag.)

Yet this is not just a matter of differentiating human beings according to their stages of development (children) or according to their particular challenges (deaf, blind, autistic, elderly people; people with cognitive or other challenges). Beyond human beings, the aspired reduction of complexity also brings new agents into play. Noteworthy in this regard are such things as "animal-computer interaction" (ACI). Clara Mancini, one of the leaders of this movement, is quick to point out that there is more than just casuistry behind such approaches and that there is more to them than mere anecdotes about Skyping dogs and chatting cats (cf. Ritvo/Allison 2014; Pongrácz et al. 2016; Golbeck/Neustaedter 2012). Rather, Mancini's program stands for a system that is fundamentally related to the field of altered social

forms ("interspecies communities") and is dedicated to promoting "multi-species awareness" (cf. Mankoff et al. 2005). She combines her endeavor with the promise of an overarching systematic approach and with the self-confidence of a newly emerging discipline, as is impressively clear from her manifesto and its positive reception (cf. Mancini 2011; Hirskyj-Douglas et al. 2018).

Not least, this obligation is a matter of social responsibility. As with overstepping the boundaries between species, this is due to more expansive ideas of participation (cf. Kelty 2016; Stahl 2014). This attentiveness is accompanied by a deeper consideration for the particular features of semiotic systems and by reconsiderations of one's own ethical positions (cf. Mancini 2011, 2017). By encouraging the intermingling of species and a political awakening, approaches such as ACI are part of a larger intellectual movement known as transhumanism or posthumanism. The latter is defined by figures that programmatically renounce differentiation. This renunciation is exemplified in Donna Haraway's book *Staying with the Trouble: Making Kin in the Chthulucene*—especially in her use of the word *critter*, which stands at the center of her thinking. As she notes, this term serves as a placeholder for a peculiarly broad range of beings (and machines): "In this book, 'critters' refers promiscuously to microbes, plants, animals, humans and non-humans, and sometimes even to machines." (Haraway 2016: 169n1) A similar argument has been put forth by the philosopher Rosi Braidotti (2013), who considers all species to be equally vulnerable to the threats of anthropogenic climate change and thus urges interspecies collaboration, which, as she vigorously pleads, should be part of the political agenda.[2] It is high time, according to Braidotti, for humans to create new social bonds—not only with other species but also with the techno-others that we tend to keep at a distance and reduce to their operational functionality. Only in such a way does she think it will be possible to ensure our common survival as a community facing the same threat.[3] What Braidotti proposes is a fundamental dedifferentiation of the social, which is comparable to the dedifferentiation of the ontological in Haraway's definition (or non-definition) of critters.

Such figures of dedifferentiation, which are central to the theoretical position of post- and transhumanism and thus seek to avoid the habitual accusation of anthropocentrism, are necessarily associated with intuitive gestures—a finding that unites the numerous movements in favor of openness and expansion against the dominance of human-computer interaction. After long phases of political abstinence, this expansion was joined on the agenda by categories such as respon-

2 This collaboration should not, moreover, be dictated by a logic of precariousness (cf. Bennke et al. 2018).

3 Such an attitude toward the techno-other is being fostered by a number of anthropophilic gestures being made on the part of machines (cf. Seaman 2011).

sibility, ethics, and participation. The tone of all this is demanding, immodest, and programmatic; as Braidotti herself concedes, it is impatient and hardly free of pathos. The way in which the concerns of individual participation offensives interrelate with those of certain theoretical formulations can be seen, for instance, in the work of Fredrik Aspling. The Swedish sociologist is a committed critic of anthropocentrism and considers himself a close ally of post- and transhumanism:

> The increased involvement of nonhuman species in interactive contexts supported by digital technology, which could be framed as multispecies-computer interaction, leads to new possibilities and forms of interactions, and consequently, a need to reconsider what this is and can be in terms of interaction. (Aspling 2015: 1)

Multispecies interaction thus becomes the operational basis for a new concept of interaction. On this basis, Aspling places a concept of inclusion on the agenda and encourages people to consider the particular needs and features of different species:

> The addition of nonhuman species challenges conventional interaction approaches and theoretical frameworks in HCI. There is a need to think beyond the human and confront the challenges associated with the inclusion of other species with dissimilar cognitions, experiences, senses, abilities, timescales, wants and needs. For further advancement we need appropriate approaches and theoretical foundations to better understand the emerging dynamics of these new forms of interactions. The attention given to nonhuman species in HCI (e.g., animal as legitimate users to design for and with) is in analogy with posthumanism and its critique of anthropocentrism. (Ibid; Aspling et al. 2018)

The issue of going beyond human-computer interaction and integrating new agents and processes is part of what is being negotiated by way of concepts such as post- and transhumanism and by way of new epochal designations such as the Anthropocene or the Chthulucene (cf. Haraway 2016). Alongside gestures of ontological opening, which feature prominently in Haraway's work, there are thus also gestures of opening up social interaction. The development of ACI (animal-computer interaction), PCI (plant-computer interaction), CCI (child-computer interaction) or RCI (robot-computer interaction) stand for this. The logic of subdividing forms of interaction into appropriate departments is just as striking as the aspect of promoting all sorts of interspecies collaboration. Interactive relationships prevail between the knowledge about various individual user groups. These relationships make it possible for such groups to learn and profit from one another: "The aim is to strengthen connected thinking whilst highlighting the exchangeable connecting methods from both ACI and HCI and their subfields in-

cluding Child Computer Interaction (CCI) and Human Robot Interaction (HRI)." (Hirskyj-Douglas et al. 2016: n. pag.; cf. Hourcade/Bullock-Rest 2011; Hourcade et al. 2018) These interactive relationships and this act of learning from one another ("discussing what these fields learn from each other with their similarities and differences mapped") lead to common design criteria. And the latter criteria keep the special or extreme user in mind—as children, as people with cognitive or sensory limitations, as autistic people, and so on (cf. Gennari et al. 2017; Eisapour et al. 2018; Lindsay et al. 2012; Satterfield et al. 2016).

Several of the maxims expressed by proponents "participatory design" are syntactical peculiarities. Now it is common to encounter expressions with dual prepositions; in order to include special users in advance, for instance, programmers are now encouraged to work *for* and *with* them. This double use of prepositions is important to the movement and therefore often seen. Noteworthy, too, is the unusual use of the preposition *with*. In this context, it is often attached to the word *becoming*, which was one of post-structuralism's objects of fascination. This mode of "becoming-with" (with animals, plants, stones), which concerns both the molecular as well as the technical and artificial, is believed to be a key element in the struggle for global survival ("the necessity to become-with animals and techno-objects as a matter of survival" (Davis 2016: 210).

Cats and children—but also people with challenges, disabilities, or highly individual needs—have become the respected target groups of special interfaces made particularly for them (cf. Maaß/Buchmüller 2018; Westerlaken/Gualeni 2016). Their participation takes place via the reduction of complexity—and this, as I have already remarked, in a field that is otherwise defined by gestures of increasing complexity. Not least, it is defined by a further gesture that involves the systematic integration of playfulness; in fact, the impression left is that playfulness is the order of the day and that play itself has a central role in eliminating the barriers between species (cf. Nijholt 2015).[4] Such measures almost make it seem as though the professionalization of algorithms is being accompanied by an infantilization movement—as a sort of counter-movement. This is tied to gestures of reduction or can at least be understood under that formula. The programmatic nature of the formula owes itself to the discovery that wherever there is talk of technology, another narrative is being expressed as well. For such an argument in favor of reduction, which is conceived in functional terms and not meant disrespectfully, one should look toward venues that break up and diversify the primacy of HCI. The recent concentration on children and the efforts—referred to by Aspling—to blend child-computer interactions with those of ACI are therefore more than mere symptoms: They modulate a praxis of their own. Atypical allianc-

4 This applies not only to the design of interfaces but also to the design of data and the practices associated with it (cf. Anderson et al. 2017).

es are now becoming visible and possible, as is evident from the following title of an article about ACI: "Of Kittens and Kiddies: Reflections on Participatory Design with Small Animals and Small Humans." (Chisik/Mancini 2017) This organized focus on children *and* cats as representatives of a desired form of intuition exemplifies some of the concerns of participatory design. The goal is to produce a user-friendly interface design that does not have to be laboriously explained but is rather intuitive, self-explanatory, and based on tacit knowledge. Participatory design is negotiated both with as well as between humans, animals, and machines. With its focus on small animals and people, it makes reduction tangible. What is more, it makes reduction the keystone of participation.

3. Asymmetries

The naturalization of designs meant for interaction, collaboration, or communication requires the use of surfaces and has operative dimensions (cf. Norman 2010). Thus it is not exhausted by gestures of dedifferentiation but rather goes hand in hand with strategic considerations. One of these is the discovery of the multisensory—or, as Caon et al. (2018) have called it, "multisensory storming." Storming the senses has been able to take place, first of all, through the increasing discovery of the tactile and the haptic—a discovery over whose course the manners of speaking about computers and algorithms have themselves been changing. Gestures of naturalization, which have been described as well as criticized within the discussion about interfaces, concern not only the problems of dealing with hardware but also manners of programming (cf. Bruns 1993; Hornecker 2008). Not only does the computer require massive strategies for accommodating the senses; the activity of programming is also under pressure to recreate itself in a new image. It has to abandon its cognitive solipsism and, beyond merely working with symbols, become a tactile undertaking. Thus, yet again, the body will become the natural guarantee of a form of comprehensive participation that can or should be able to take place without effort, intuitively, and in the transparent mode of self-evidence.

The issue is not only computer use and literacy but also a life world that allows technology to exist in any given ambient form. By now there are abundant examples of this and, on a systematic level, they tend to have certain features in common, most notably the development of new channels, the integration of different senses, and the emergence of new forms of communication. The latter free up scenes of asymmetrical communication—scenes that invalidate the common conceptions of communication theory. The abundance of examples extends from applications for remotely caring for pets to interacting with plants, which are often grown in artificial environments (cf. Lee et al. 2006; Kuribayashi et al. 2007).

They alter forms of sociality. One of the most theoretically ambitious protagonists in this field is the Japanese researcher Hill Hiroki Kobayashi.[5] His goal is to transcend a paradigm of communication and interaction that is measured solely on the basis of human beings (in full possession of their mental faculties) and a particular form of linguistic communication. Kobayashi's notion of "human-computer-biosphere interaction" (HCBI) has a virtually unlimited field of operation. It not only changes the sphere of actors but also, and necessarily, the ways in which communication takes place: "HCBI extends the subject of Human Computer Interaction (HCI) from countable people, objects, pets, and plants into an auditory biosphere that is uncountable, complex, and non-linguistic." (Kobayashi 2010: n. pag.) This abandonment of the anthropocentric standpoint is as much a program as it is a collaboration with agents that elude the principle of countability (cf. Kobayashi 2014). In this way, possible forms of expression beyond articulated speech are assigned a central role. Regarding the use of wearables that are meant to bring people closer to nature ("Wearable Forest-Feeling of Belonging to Nature" is the title of his article), Kobayashi writes: "Thus, wearable computer systems have become an inter-medium to express the telepresence of various species in the biosphere in such a way that their non-linguistic expression is perceived and understood by each participant, which violates all the rules of linguistic science." (Kobayashi 2008: 1133)

The locus for such applications is thus close to life and by no means limited to art installations. An indication of how lifelike they can be is provided by a device called LumiTouch. At first glance, LumiTouch looks like a regular pair of picture frames. One inconspicuous frame is connected to an equivalent through the internet, and it is able to trigger signals that correspond to someone's mere touch. Depending on the type of touch (its intensity, frequency, duration), various light patterns and color constellations are released that can be associated with an individualized code. According to its designers, the latter is suitable for implementing a special form of expression and thus encourages the development of a private emotional language (cf. Kaye/Goulding 2004). LumiTouch changes the simplified (because idealized) models of communication theory, and the act of touching the picture frame has useful advantages for people with impairments. What its designers envision are forms of asymmetrical exchange for which one of the communication partners does not need to be in full command of his or her cognitive or physical abilities: "People who are unable to actively communicate for long periods of time (e.g. sick or elderly) might be able to use the passive transmission of LumiTouch." (Chang et al. 2001: 314) The potential of overtaxing motor skills or cognitive faculties in certain situations, such as when someone is bed-ridden, can be counteracted with communicative systems that are less demanding: "Similar-

5 See his homepage at http://hhkobayashi.com (accessed June 2, 2019) and Nijholt 2015.

ly users who lack the required dexterity or concentration for pushing numerous buttons might appreciate this system due to its small number of simple grasping inputs." (Ibid.)

Fig. 2: LumiTouch (Chang et al. 2001: 314)

Another system that is based on reduction is a product called Tsunagari-kan' Communication, which is devoted to the goal of ensuring communication between distant family members (cf. Miyajima et al. 2005). Here, too, what is favored is a non-linguistic form of intimate communication (""Tsunagari' communication aims to foster a feeling of connection between people living apart by exchanging and sharing the cue information via network everyday." (Itoh et al. 2002: 810) Expanding upon LumiTouch's model, it also allows communication to take place in the mode of the unconscious and passive. Using a so called "Family Planter" as a communicative tool, it is meant to enable firm social bonds to form through exchanges of "cue information" (ibid.: 811). By means of infrared and ultrasound sensors, Tsunagari's interconnected terminals react to a person's presence and movement. This information is transmitted and converted on the receiving end into a non-linguistic signal:

> Optical fibers at the top of the terminal will gleam to indicate the remote human presence and will rotate to indicate the remote human motion. This is intended to exchange presence and movement information implicitly (without explicit intervention from users) and constantly. (Ibid.)

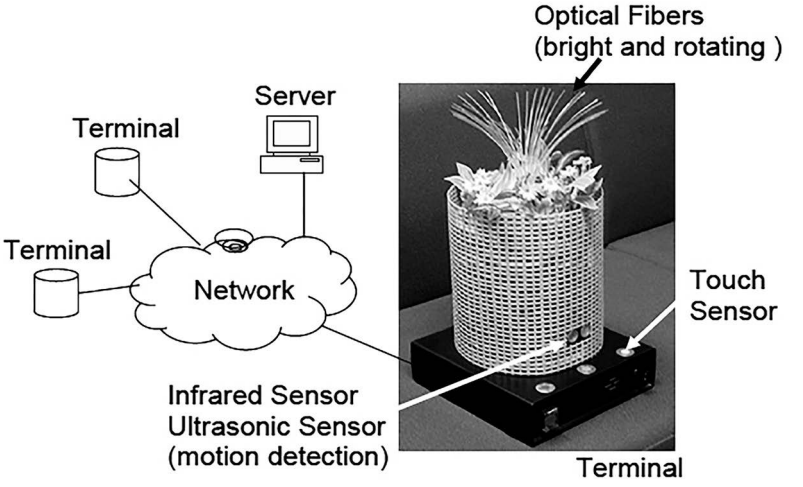


Fig. 3: Family Planter (Itoh et al. 2002: 810)

The design of this planter-based sensory device contains several important aspects that also happen to be central to interspecies communication and interaction. In the mode of implicit and thus unconscious participation, the system allows people to partake in the everyday lives of remote family members seamlessly and in a way that is not felt as an imposition or disruption: "These exchanges are designed to blend into the everyday life of a user." (Ibid.) By means of three sensors, it can transmit various audio signals, and thus the system can also be used to convey explicit messages. More important than this explicit mode, however, is the implicit nature of its use, which, with its unobtrusive participation, submits to the logic of media and the way in which they increasingly blend inconspicuously into our environments. They now do so seamlessly, unobtrusively, quietly, smoothly, and ubiquitously, and these qualities are redefining the ways that theorists should think about media in general. This would be a media theory that, freed from the paradigmatic idea that media are extensions or organic projections of humans, could instead be described with adjectives such as *ubiquitous*, *seamless*, and *calm* (cf. Weiser/Brown 1996). It would be a media theory that directs its focus toward the issues involved with making communication more intimate and embraces its own intimately charged objects. This trend toward developing things that can be laden with affect is only growing. It is driven by an identifiable agenda and not

by casuistry (cf. Choi et al. 2014; Kaye/Goulding 2004). Its basis—reduction—has become a program whose results will become a part of everyday life.

Fig. 4: Lamp (Angelini/Caon. 2015: n.p.)

4. Finis (hominis)

Children and cats aside, what all of this brings to light are the needs and venues of a sort of communication and collaboration that is designed to be asymmetrical and yet non-discriminatory. The applications presented above do not aim to optimize ways of dealing with technical environments but rather hope to provide alternative and less complex ways of using them (cf. Rieger 2019). Thus the view has also shifted away from the previous stubborn orientation toward a particular type of user (cf. Satchell/Dourish 2009). Two things remain to be said in closing: First, the children and cats, which I have introduced here as representatives of a broader phenomenon, are being put to functional use. What this comes down to is not an offer of minimization, such as that which defines rampant cat content, but rather the functional equivalent of a strategically pursued reduction of complexity. Among these pursuits are campaigns for acceptance that include special users and shift the focus of designs toward all possible forms of participation. One of the latter is the gesture of naturalization (cf. Andreas et al. 2018).

The second point concerns the question of who rules the network. To this question there is, at first glance, a simple answer, and it has nothing to do with the power of inconspicuous algorithms but rather with online content. It was none other than the deep-learning processes of Google Brain that brought to light the fact that it is cats that have, in quantitative terms, been dominating what is going on there (cf. Guerin/Vasconcelos 2008). Much to the amusement of those working on the project, their algorithms revealed that, indeed, the cat is the lord of the internet—a supposition that Alexander Pschera (2016) also plays with, though somewhat less jokingly, in an article devoted to the "internet of animals." For some time now, the internet has not belonged to people alone. This situation is now even reflected in puns that, as silly as they may be, nevertheless support the ethical arguments of participatory design: "Our work focuses on canine companions, and includes, *paw*ticipatory design, *lab*radory tests, and *canid* camera monitoring." (Mankoff et al. 2005: 253; cf. Trindade et al. 2015) Or, regarding cats in particular: "In the modern era of digital media, it is hard to deny that cats have clawed their way into the zeitgeist of the Internet." (Myrick 2015: 175)

The title that I have chosen for this essay—"Reduction and Participation"—takes the demands for including other species and forms of existence at their word. The aim of such demands is to expand the circle of those with agency and epistemic relevance. Multispecies communities will be home to new actors, new forms of communication and collaboration, new types of design and participation, new responsibilities and social forms: between humans and animals, plants and stones, artefacts and biofacts, machines and media, the living and the non-living, the real and the virtual, the augmented and un-augmented, the simulated and the modelled, the increased and the reduced (cf. Leistert 2017). It is therefore only consistent that, in this sphere of actors, algorithms might not find their peace but will certainly find their place.

Translated by Valentine A. Pakis

Images

Fig. 1: Surveillance versus Sousveillance (https://en.wikipedia.org/wiki/File:Sur SousVeillanceByStephanieMannAge6.png, accessed June 4, 2019)
Fig. 2: LumiTouch (Chang et al. 2001: 314)
Fig. 3: Family Planter (Itoh et al. 2002: 810)
Fig. 4: Lamp (Angelini/Caon. 2015: n.p.)

References

Anderson, Theresa Dirndorfer et al. (2017): "Data Play: Participatory Visualisation to Make Sense of Data." In: Proceedings of the Association for Information Science and Technology 54/1, pp. 617-18.

Andreas, Michael et al. (2018): "Unterwachen und Schlafen: Einleitung." In: Michael Andreas et al. (eds.), Unterwachen und Schlafen: Anthropophile Medien nach dem Interface, pp. 7-31.

Angelini, Leonardo/Caon, Maurizio (2015): "Towards an Anthropomorphic Lamp for Affective Computing." In: Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction, New York: ACM, pp. 661-666.

Aspling, Fredrik (2015): "Animals, Plants, People and Digital Technology: Exploring and Understanding Multispecies-Computer Interaction." In: Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology, New York: ACM, pp. 1-55

Aspling, Fredrik et al. (2018): "Understanding Animals: A Critical Challenge in ACI." In: Proceedings of the Tenth Nordic Conference on Human-Computer Interaction, New York: ACM, pp. 148-60.

Bennke, Johannes et al. (2018) (eds.): Das Mitsein der Medien: Prekäre Koexistenzen von Menschen, Maschinen und Algorithmen, Paderborn: Wilhelm Fink.

Braidotti, Rosi (2013): The Posthuman, Cambridge: Polity.

Bruns F. Wilhelm (1993): "Zur Rückgewinnung von Sinnlichkeit: Eine neue Form des Umgangs mit Rechnern." In: Technische Rundschau 29, pp. 14-18.

Bunz, Mercedes (2012): Die stille Revolution: Wie Algorithmen Wissen, Arbeit, Öffentlichkeit und Politik verändern, ohne dabei viel Lärm zu machen, Berlin: Suhrkamp.

Caon, Maurizio et al. (2018): "Towards Multisensory Storming." In: Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems, New York: ACM, pp. 213-18.

Chang, Angela et al. (2001): "LumiTouch: An Emotional Communication Device." In: Extended Abstracts on Human Factors in Computing Systems, New York: ACM, pp. 313-14.

Chisik, Yoram/Mancini, Clara (2017): "Of Kittens and Kiddies: Reflections on Participatory Design with Small Animals and Small Humans." In: Proceedings of the 2017 Conference on Interaction Design and Children, New York: ACM, pp. 753-56.

Choi, Yongsoon et al. (2014): "Ring*U: A Wearable System for Intimate Communication Using Tactile Lighting Expressions." In: Proceedings of the 11th Conference on Advances in Computer Entertainment Technology, New York: ACM, n. pag.

Davis, Heather (2016): "Molecular Intimacy." In: James Graham et al. (eds.), Climates: Architecture and the Planetary Imaginary, Zurich: Lars Müller, pp. 205-11.

Eisapour, Mahzar et al. (2018): "Participatory Design of a Virtual Reality Exercise for People with Mild Cognitive Impairment." In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, New York: ACM, n. pag.

Engemann, Christoph/Sudmann, Andreas (2018) (eds.): Machine Learning. Medien,Infrastrukturen und Technologien der Künstlichen Intelligenz, Bielefeld: transcript.

Foucault, Michel (2002 [1966]): The Order of Things: An Archaeology of the Human Sciences, New York: Routledge.

Gennari, Rosella et al. (2017): "The Participatory Design Process of Tangibles for Children's Socio-Emotional Learning." In: Simone Barbosa et al. (eds.), End-User Development: 6th International Symposium, Cham: Springer, pp. 167-82.

Golbeck, Jennifer/Neustaedter, Carman (2012): "Pet Video Chat: Monitoring and Interacting with Dogs over Distance." In: Extended Abstracts on Human Factors in Computing Systems, New York: ACM, pp. 211-20.

Guerin, Frank/Vasconcelos, Wamberto (2008) (eds.): The Reign of Catz and Dogs: The Second AISB Symposium on the role of Virtual Creatures in a Computerised Society, London: AISB.

Hall, Gary (2013): "Toward a Postdigital Humanities: Cultural Analytics and the Computational Turn to Data-Driven Scholarship." In: American Literature 85, pp. 781-809.

Haraway, Donna J. (2016): Staying with the Trouble: Making Kin in the Chthulucene, Durham, NC: Duke University Press.

Heller, Christian (2011): *Post Privacy: Prima leben ohne Privatsphäre*, Munich: C. H. Beck.

Hirskyj-Douglas, Ilyena et al. (2016): "Where HCI meets ACI." In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction, New York: ACM, n. pag.

Hirskyj-Douglas, Ilyena et al. (2018): "Seven Years After the Manifesto: Literature Review and Research Directions for Technologies in Animal Computer Interaction." In: Multimodal Technologies and Interaction 2/2, pp. 1-25.

Homepage Kobayashi, Hill Hiroyuki: http://hhkobayashi.com (accessed June 2, 2019).

Hornecker, Eva (2008): "Die Rückkehr des Sensorischen: Tangible Interfaces und Tangible Interaction." In: Hans Dieter Hellige (eds.), Mensch-Computer-Interface: Zur Geschichte und Zukunft der Computerbedienung, Bielefeld: transcript, pp. 235-56.

Hourcade, Juan Pablo/Bullock-Rest, Natasha E. (2011): "Universal Interactions: Challenges and Opportunities." In: Interactions 18/2, pp. 76-79.

Hourcade, Juan Pablo et al. (2018): "Child-Computer Interaction, Ubiquitous Technologies, and Big Data." In: Interactions 25/6, pp. 78-81.

Itoh, Yoshihiro et al. (2002): "'TSUNAGARI' Communication: Fostering a Feeling of Connection between Family Members." In: Conference on Human Factors in Computing Systems, New York: ACM, pp. 810-11.

Jones, Nicola (2014): "The Learning Machines." In: Nature 505, pp. 146-48.

Kammerer, Dietmar/Waitz, Thomas (2015): "Überwachung und Kontrolle: Einleitung in den Schwerpunkt." In: Zeitschrift für Medienwissenschaft 13, pp. 10-20.

Kaye, Joseph/Goulding, Liz (2004): "Intimate Objects." In: Proceedings of the Fifth Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, New York: ACM, pp. 341-44.

Kelty, Christopher (2016): "Participation." In: Benjamin Peters (eds.), Digital Keywords: A Vocabulary of Information Society and Culture, Princeton, New Jersey: Princeton University Press, pp. 227-41.

Kobayashi, Hill Hiroki et al. (2008): "Wearable Forest-Feeling of Belonging to Nature." In: Proceedings of the 16th ACM Conference on Multimedia, New York: ACM, pp. 1133-34.

Kobayashi, Hill Hiroki (2014): "Human-Computer-Biosphere Interaction: Beyond Human-Centric Interaction." In: Norbert Streitz/Panos Markopoulos (eds.), Distributed, Ambient, and Pervasive Interactions: Proceedings of the Second International Conference, DAPI 2014, Cham: Springer, pp. 349-58.

Kobayashi, Hill Hiroki (2010): Basic Research in Human-Computer-Biosphere Interaction, Doctoral Diss.: Tokyo University, n. pag.

Kortum, Philip (2008) (eds.): HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces, Oxford: Elsevier Science.

Kuribayashi, Satoshi et al. (2007): "Plantio: An Interactive Pot to Augment Plants' Expressions." In: Proceedings of the International Conference on Advances in Computer Entertainment Technology, New York: ACM, pp. 139-42.

Lee, Shang Ping et al. (2006): "A Mobile Pet Wearable Computer and Mixed Reality System for Human-Poultry Interaction through the Internet." In: Personal and Ubiquitous Computing 10/5, pp. 301-17.

Leistert, Oliver (2017): "Social Bots als algorithmische Piraten und als Boten einer techno-environmentalen Handlungskraft." In: Robert Seyfert/Jonathan Roberge (eds.), Algorithmuskulturen: Über die rechnerische Konstruktion der Wirklichkeit, Bielefeld: transcript, pp. 215-34.

Lindsay, Stephen et al. (2012): "Empathy, Participatory Design and People with Dementia." In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York: ACM, pp. 521-30.

Maaß, Susanne/Buchmüller, Sandra (2018): "The Crucial Role of Cultural Probes in Participatory Design for and with Older Adults." In: i-com 17/2, pp. 119-35.

Mancini, Clara (2017): "Towards an Animal-Centred Ethics for Animal-Computer Interaction." In: International Journal of Human-Computer Studies 98, pp. 221-33.

Mancini, Clara (2011): "Animal-Computer Interaction (ACI): Changing Perspective on HCI, Participation and Sustainability." In: Interactions 18/4, pp. 60-73.

Mancini, Clara (2018): "Animal-Computer Interaction (ACI): A Manifesto." In: Interactions 18/4, pp. 60-73.

Mankoff, Demi et al. (2005): "Supporting Interspecies Awareness: Using Peripheral Displays for Distributed Pack Awareness." In: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, New York: ACM, pp. 253-58.

Mann, Steve et al. (2015): "Declaration of Veillance (Surveillance is a Half-Truth)." In: 2015 IEEE Games Entertainment Media Conference, Piscataway, New Jersey: IEEE, pp. 1-2.

Mann, Steve (2016): "Surveillance (Oversight), Sousveillance (Undersight), and Metaveillance (Seeing Sight Itself)." In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Piscataway, New Jersey: IEEE, 2016), pp. 1-10.

Miyajima, Asami et al. (2005): "'Tsunagari-kan' Communication: Design of a New Telecommunication Environment and a Field Test with Family Members Living Apart." In: International Journal of Human-Computer Interaction 19, pp. 253-76.

Myrick, Jessica Gall (2015): "Emotion Regulation, Procrastination, and Watching Cat Videos Online: Who Watches Internet Cats, Why, and to What Effect?" In: Computers in Human Behavior 52, pp. 168-76.

Nijholt, Anton (2015) (eds.): More Playful User Interfaces: Interfaces that Invite Social and Physical Interaction, Singapore: Springer.

Norman, Donald A. (2010): "Natural User Interfaces Are Not Natural." In: Interactions 17/3, pp. 6-10.

Pongrácz, Alexandre et al. (2016): "A Dog Using Skype." In: Proceedings of the Third International Conference on Animal-Computer Interaction, New York: ACM, n. pag.

Pschera, Alexander (2016): "Das Internet der Tiere: Natur 4.0 und die conditio humana." In: Zeitschrift für Medien- und Kulturforschung 7, pp. 111-24.

Rehman, Arshia et al. (2019): "Automatic Visual Features for Writer Identification: A Deep Learning Approach." In: IEEE Access 7, pp. 17149-57.

Reißmann, Ole (2011): "Internet-Exhibitionisten 'Spackeria': 'Privatsphäre ist so was von Eighties'." In: Spiegel Online March 10, http://www.spiegel.de/netzwelt/netzpolitik/internet-exhibitionisten-spackeria-privatsphaere-ist-

sowas-von-eighties-a-749831.html (an interview with Julia Schramm; accessed May 26, 2019).

Rieger, Stefan (2019): "Virtual Humanities." In: Dawid Kasprowicz/Stefan Rieger (eds.), Handbuch Virtualität, Berlin: Springer.

Rieger, Stefan (2019): Die Enden des Körpers: Versuch einer negativen Prothetik, Wiesbaden: Springer.

Rieger, Stefan (2018): "Alles, was zählt: Observations by a Quantified Selfie." In: Psychosozial 152, pp. 47-56.

Ritvo, Sarah E./ Allison, Robert S. (2014): "Challenges Related to Nonhuman Animal-Computer Interaction: Usability and 'Liking'." In: Proceedings of the 2014 Workshops on Advances in Computer Entertainment, New York: ACM, n. pag.

Rodriguez, Catherine S. et al. (2019): "Two-Stage Deep Learning Approach to the Classification of Fine-Art Paintings." In: IEEE Access 7, pp. 41770-81.

Satchell, Christine/Dourish, Paul (2009): "Beyond the User: Use and Non-Use in HCI." In: Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group, New York: ACM, pp. 9-16.

Satterfield, Debra et al. (2016): "An Analysis of Data Collection Methods for User Participatory Design for and with People with Autism Spectrum Disorders." In: Marcus Aaron (eds.), Design, User Experience, and Usability: Design Thinking and Methods, Cham: Springer, pp. 509-16.

Seaman, Bill (2011): Neosentience: The Benevolence Engine, Bristol: Intellect.

Seyfert, Robert/Roberge, Jonathan (2017) (eds.): Algorithmuskulturen: Über die rechnerische Konstruktion der Wirklichkeit, Bielefeld: transcript.

Stahl, Bernd Carsten (2014): "Participatory Design as Ethical Practice: Concepts, Reality and Conditions." In: Journal of Information, Communication and Ethics in Society 12, pp. 10-13.

Stark, Emily et al. (2018a): "Medicine Has Gone to the Dogs: Deep Learning and Robotic Olfaction to Mimic Working Dogs." In: IEEE Technology and Society Magazine 37/4, pp. 55-60.

Stark, Emily et al. (2018b): "Odorveillance and the Ethics of Robotic Olfaction." In: IEEE Technology and Society Magazine 37/4, pp. 16-19.

Sudmann, Andreas (2017): "Deep Learning als dokumentarische Praxis." In: Sprache und Literatur 48/2, pp. 155-70.

Trindade, Rui et al. (2015): "Purrfect Crime: Exploring Animal Computer Interaction through a Digital Game for Humans and Cats." In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, New York: ACM, pp. 93-63.

van Dijck, José (2014): "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." In: Surveillance & Society 12/2, pp. 197-208.

Weiser, Mark/Brown, John Seely (1966): "The Coming Age of Calm Technology." In: Xerox PARC October 5, https://calmtech.com/papers/coming-age-calm-technology.html (accessed June 4, 2019).

Westerlaken, Michelle/Gualeni, Stefano (2016): "Becoming With: Towards the Inclusion of Animals as Participants in Design Processes." In: Proceedings of the Third International Conference on Animal-Computer Interaction, New York: ACM, n. pag.

The Political Affinities of AI

Dan McQuillan

Introduction

We need a radical politics of AI, that is, a politics of artificial neural networks. AI acts as political technology, but current efforts to characterise it take the form of liberal statements about ethics. Issues of bias in AI are treated as questions of fairness, as if society is already a level playing field that just needs to be maintained. Transparency and accountability are seen as sufficient to correct AI problematics (ACM FAT 2019), as if it is being introduced into well-functioning and genuinely democratic polities. The apparent refusal to see AI as political flies in the face of its promotion as a solution to austerity. In the UK, for example, discussions about the under-funded public healthcare system are peppered by senior level statements that "AI may be the thing that saves the NHS" (Ghosh 2018). Austerity is not a natural disaster but a political decision to prop up financial institutions at the expense of public spending. The hope of those decision makers is that machinic reasoning can solve the riddle of dealing with rising needs using sharply reduced resources. Meanwhile the operating characteristics of actual AI have other political impacts, such as the deracination of due process. The vast parallel iterations carried out by backpropagation cast an opacity over AI by making its optimisations very hard to reverse to human reasoning (Lipton 2016). Algorithmic judgements that affect important social and political decisions are thus removed from discourse.

The political dimensions of artificial intelligence cannot be divined in the abstract nor solved by philosophical ethics. They result from concrete technical operations, such as sums over vectors, in the context of specific social conditions. The idea of ethical AI is an information operation designed to calm public fears about algorithmic impacts, and to position it for market advantage (Hern 2018). The real hazards of AI emerge as it intermingles with the political currents of our time. A figure for the political entanglement of AI is a photograph taken at the recent World Economic Forum showing the populist and extreme right Brazilian politician Jair Bolsonaro seated at lunch between Apple CEO Tim Cook and Microsoft CEO Satya Nadella (Slobodian 2019). Artificial neural networks are in demand because the confluence of big data and processing power, in the form of GPUs, has

enabled them to produce uncanny results in fields like image recognition. However, the years in which AI is coming of age are also the years of neoliberal crisis and a global rise in far right politics. The urgent question is, how do the concrete technical operations of neural networks reinforce, enforce or extend these political currents, and how (if at all) they might instead serve the goals of social justice.

Boundaries

The attraction of deep learning is its ability to produce predictions from overflows of data. The weights in the layers are optimised by iterations that drive the loss function into a minima, while activation functions like ReLU act ruthlessly at each neuron to remove weaker signals (3Blue1Brown. n.d). The overall effect is the production of statistical certainty; a net of weights that will transform messy input data into unambiguous classifications. This is done by substituting correlations for any attempt to establish causal mechanism, and is not constrained by any wider framework of consistency. There is no element of 'common sense' in the mechanism that differentiates between guesses based on embodied experience of the world. Neural networks are neoplatonic; they claim a hidden mathematical order in the world that is superior to direct experience (McQuillan 2017). The politics enters in the way these orderings are entrained in wider mechanisms. Instead of constraining statistical authority based on a broader care for the human consequences, the current race to adopt AI is driven by the way its single-minded optimisation resonates with institutional goals of maximising efficiency or shareholder value. The operations of AI act in harmony with a neoliberalism that perceives the world as an atomised set of inputs into a market mechanism that will necessarily produce the optimum result.

The purpose of AI's mathematical regressions is to draw decision boundaries, such that an input is cleanly categorised as on one side or the other. Connecting this to matters of risk in the external world, even where the ends are supposedly benign, propels AI into being a system of control. Its calculative categorisations trigger chains of machine and human decisions with real consequences, involving the allocation or removal of resources or opportunities. Embedded in deep learning, obfuscated from due process or discourse, these numerical judgements have a law-like force without being of the law. Thus, the predictive boundaries of AI map outwards as continuous partial states of exception (Agamben 2005). The expertise to contest the calculations of machinic reason in their own terms is highly centralised in a few corporations and universities. For the rest of us, the calculative authority of machine learning leads to situations where personal testimony is devalued with respect to computational insights. AI becomes an engine for epistemic injustice, claiming insights that override lived experience (Fricker 2007).

Bureaucracy

Like bureaucracy in the twentieth century, AI is poised to become the unifying logic of legitimation across corporations and government. At the current time, the performance of deep learning is proportional to the amount of computing power used: between the AlexNet image recognition breakthrough of 2012 and the Google DeepMind system that beat the Go grandmaster, the required processing power grew by a factor of 300,000 to around 2000 Petaflops/s-day (Amodei/Hernandez 2018). The hardware and software pipelines of deep learning are becoming strategically important, and existing instances like Amazon Web Services are increasingly indistinguishable from critical national infrastructure (Konkel 2016). But although AI is materialised in the fenced-off anonymity of server farms, its leverage lies between thought and action. Deep learning applied to social decisions becomes the concrete manifestation of Bourdieu's habitus; structured structures predisposed to function as structuring structures (Bourdieu 1990). It is not that key decisions are delegated to machines with no human in the loop; rather, that people making pressured decisions are presented with empirical rankings of risk, who's derivation they have no way of questioning. AI encourages thoughtlessness in the sense described by Hannah Arendt; the inability to critique instructions, the lack of reflection on consequences, a commitment to the belief that a correct ordering is being carried out (Arendt 2006).

Through prediction, this ordering extends bureaucratic governmentality to the domain of intent or tendency, which it strives to preempt as a service or as an intervention. As well as classifications of pre-crime and the proliferation of forms of 'pre-extremism', as prototyped by the UK's Prevent Strategy (Sian 2017), there will also be classifications that claim benevolence and efficient resource allocation, such as pre-diabetes or pre-dementia (LaMattina 2016). The drive for preemption enfolds social fears and market interests with the aim of eliminating that which is undesirable. The problem is that AI is reductionist. It can only learn from those aspects of the context that can be mathematised, and it is given a singular goal to optimise on. Therefore attempt by AI to explain what is going on reduces the entire system to certain constituent elements and their interactions. Moreover, predictive deep learning applied to social questions implies that attributes are individualised and innate, while obfuscating the background of common social causes. It will extend an apartheid bureaucracy to any aspect of life touched by data.

Instability

AI should not be applied to any part of complex social and cultural problems, outside of extremely narrow and restricted aspects. This is not only because its mode of operation encourages thoughtlessness and reductionism, but because deep learning is literally out of its depth when it comes to social and political complexity. It is, in essence, simply a pattern finding technique which works surprisingly well at perceptual classification and in some other well-bounded applications such as game playing. However, even in these heartlands of AI there are signs of systematic problems. There are many adversarial examples where the addition of carefully chosen noise to an image, which appears to human perception as no more than a scattering of insignificant white dots, can force a neural network to wrongly classify an obvious image (Goswami et al. 2018). Perhaps more significantly, a recent paper shows that deep learning's image recognition often falls apart when confronted with common stimuli rotated in three dimensional space into unusual positions. In one of the examples, the network correctly recognises a school truck, but when it sees a real picture of one on it's side it mis-classifies it as a snow plough. The authors conclude that, while deep neural networks work well at image classification, they are still far from true object recognition, and their understanding of objects is quite naive. Their conclusion is that "deep neural networks (DNNs) can fail to generalize to out-of-distribution (OoD) inputs, including natural, non-adversarial ones, which are common in real-world settings" (Alcorn et al. 2018). This is an important but hardly surprising observation. No neural network has any understanding of anything, in the form of an abstract model or ontology that can be freely applied to novel situations. That is, neural networks are incapable of exactly the kind of adaptive and analogical thinking that characterises even young children. Statements from leading AI engineers that neural networks would either "now or in the near future" be able to do "any mental task" a person could do "with less than one second of thought" (Ng 2016) is not only laughable but actually dangerous. If deep learning can't recognize objects in non-canonical poses, we should not expect it to do everyday, common sense reasoning, a task for which it has never shown any facility whatsoever. Still less should we apply it in messy socio-political contexts and expect it to draw out insights that have previously been delegated to discourse.

However, being out of its depth is not the only reason we should keep deep learning clear of socially sensitive situations. The single-minded optimisation that makes AI resonate so well with a neoliberal perspective brings with it a fatal ethical payload. Utility functions, like deep learning's backpropagation, get in to ethical deep water when there are independent, irreducible objectives that need to be pursued at the same time. Ethicists have theorems that suggest it's impossible for an optimisation to produce a good outcome for a population without violating

our ethical intentions. For example, the mere addition paradox shows that, if optimising on a social welfare function over any population of happy people, there exists a much larger population with miserable lives that is 'better' (more optimised for total wellbeing) than the happy population (Eckersley 2018). Not surprisingly this paradox is also known as the repugnant conclusion. While this may seem to derive from an abstract, analytical logic of moral philosophy, let us remember that that is the point: through institutionalised neural networks we are applying an abstract and calculative logic to the social world. Similar ethical reasoning has produced a whole set of unappealing paradoxes such as the 'sadistic conclusion' and the 'very anti-egalitarian conclusion'. These suggest a basic incompatibility between different utilitarian objectives such as maximizing total wellbeing, maximizing average wellbeing, and avoiding suffering. Thoughtless pursuit of an objective function, as instrumentalised in AI, leads to ethically toxic consequences even when the initial function is apparently benign, let alone when it serves the capitalist goal of profit.

The possible

Thoughtlessness also enters at the start of the road to an AI solution. AI is always in the service of solving what Bergson called 'ready-made problems'. That is, machine learning is applied to problems which are based on unexamined assumptions, such as cultural biases and institutional goals, and those deeper prejudices which are embedded in language itself. The problem with a ready-made problem is that it presupposes a range of possible solutions which are coterminous with that particular expression of the problem. Bergson argued that if one accepts a ready-made problem "one might just as well say that all truth is already virtually known, that its model is patented in the administrative offices of the state, and that philosophy is a jig-saw puzzle where the problem is to construct with the pieces society gives us the design it is unwilling to show us." (Henri Bergson, La pensée et le mouvant, cited in Solhjdu 2015). To have agency, to be able to change a given reality, is instead a question of finding the problem and of positing it. This is different because "stating the problem is not simply uncovering, it is inventing." According to Solhdju, Isabelle Stengers expresses this as "the difference between the possible and the probable" where the probable is "that which with respect to the real only lacks one single thing, existence" (Solhdju 2015). It can already be constructed using the same conceptual scaffolding that was used to build the problem, and figuring it out is simply a matter of probabilistic deduction. The possible, on the other hand, is of something unpredictable and non-calculable; a creative act that is not merely rearrangement of existing truths. What's at stake is not the probable of current AI but the possible of political thought and action. We

need to approach AI in a way that enables us to take sides with the possible against probabilities.

Recreating the possibilities of machine learning means working with programming and politics as non-divisibles, solving engineering problems while sustaining a focus on social impacts. It requires both precision at a mathematical level and and openness towards the different possible realities that might be articulated. One approach to such a discipline might be offered by a feminist model of science, such as that described by Roy, Harding and Spanier (Roy 2004). That is, an expanded form of scientific methodology that includes the origination of the problem and the purpose of the inquiry. Those wishing to develop non-oppressive machine learning should not accept a problem as given, but should start by locating its origins, in other words the structural forces which have posited it and prioritised it. Uncovering the purposes of an inquiry with deep learning means going beyond accurately predicting the validation data by optimising hyperparameters. It means understanding this narrow technical purpose as part of a broader set of impacts, asking who's ends it will serve, who it might exclude, and how it would effect the wider wellbeing of society. Perhaps most radically for AI, a feminist approach establishes a relationship between the inquirer and their subject of inquiry, requiring us to purposefully put aside the onlooker consciousness that fuels AI's hubris. The most direct way to put this feminist method into practice with machine learning is through collective structures of research that include the 'target group' in the process of inquiry, through structures such as people's councils (McQuillan 2018). Such situated collectives of inquiry are well placed to re-invent the problem as lines of flight from the tyranny of the probable.

Political action

We must establish this alternative against the political currents that resonate with thoughtless AI. This will not be an easy task. AI being implemented as 'AI under austerity', that is, as neoliberalism's response to its own crisis. Everyday cruelties such as welfare cuts to the disabled are being increasingly obfuscated by machinic classifications (Alston n.d.). Neural networks could become engines of epistemic injustice and partial states of exception. Even more dangerously, the simplification of social problems to optimisation based on reductionist reasoning and innate characteristics echoes exactly the politics of the populist far right. Pointing out the inconsistencies in the claims of AI has no traction with this political tendency. Stupidity and hate don't require philosophical consistency, only an operational effectiveness that performs their ideological theatre of cruelty. The practice of AI must develop a politics that resists authoritarianism and asserts a care for our common humanity.

Thus the necessity of collective practices of AI is not only an epistemological necessity but a political one. The political forms of the people's council and the general assembly can return the questions of due process and justice to their proper place in discourse. Where algorithmic authority comes from privileging generalised abstractions, direct democracy can be reasserted by the mobilisation of situated knowledges. These need to be channelled into forms applicable to computational technologies. Ivan Illich, in his call for convivial technology, proposed 'counterfoil research' whose goal is to detect "the incipient stages of murderous logic in a tool" (Illich 1975) where a tool, for Illich, means a specific combination of technologies and institutions. Counterfoil research lays out a plausible programme for AI people's councils; that they should "clarify and dramatize the relationship of people to their tools", "hold constantly before the public the resources that are available and the consequences of their use in various ways" and identify "those classes of people most immediately hurt by such trends". This is not a negative programme but a positive one, to create conditions where people have the capacity for autonomous action by means of tools least controlled by others. The goals is to find appropriate limits for our tools. Limiting tools through the mechanism of assemblies also creates what Hannah Arendt identified as spaces for action, which only arises from face-to-face encounters and is that which happens "against the overwhelming odds of statistical laws and their probability" (Arendt 1998).

However, such spaces will not be freely given. Forms of resistance will be necessary to create them. One potential form of resistance is in worker self-organisation, both in the heart of AI engineering and in the other places of work which will be affected by it. There are small signs of the former in way employees of corporations like Google, Microsoft and Amazon have expressed dissent at the adoption of their technical creations by the military and security apparatus (Alba 2018, Conger 2018, Lee 2019) whereas the latter has so far been limited to those precarious workers, like Uber and Deliveroo drivers, who are already "below the algorithm" (Möhlmann/Henfridsson 2017). In workers self-organisation, too, the collective forms of the assembly and council have a key role to play, especially in advancing the ambition of workers to know that "by organizing industrially we are forming the structure of the new society within the shell of the old." In the mid-1970s workers in a major arms company used grassroots assemblies to generate a plan for restructuring their factories. Their programme, the 'Lucas Plan', would have not only converted the activity of the machinery away from arms production but included newly invented possibilities for products which, in retrospect, seem ahead of their time in terms of environmental impact (Open University 1978). Another potential form of resistance that may emerge by necessity is Luddism, where people oppose the predations of hegemonic technology through direct action. The historical Luddites' opposition to steam-powered machines of

production was based on the new social relations of subjection that they produced. Rather than some atavistic dislike of technology, the resistance of the Luddites was motivated by their alternative social vision (Binfield 2004). Their call was to 'put down all Machinery hurtful to Commonality'. A new Luddism is one way to characterise attacks on self-driving vans by residents in Arizona, fed up of the way Waymo is testing its autonomous AI in their communities and on the streets where their children are playing. Deep learning has proved again what the radicals of the 1970s claimed, that the domain of production has extended to everyday life, and that we live in the 'social factory' (Cuninghame 2015).

Historical Luddism was part of a wider uprising of workers and communities that deeply rattled the emerging industrial elites of its time. AI as it stands is the tool of a new technocratic elite. Whatever the strategies for restructuring AI, they clearly won't come about without engagement with the wider field of progressive politics. AI, in the form of neural networks, is an inherently political technology which must be acknowledged as such. Adopted without constraints it will tend to amplify the injustices of the status quo, or even become part of a shift to a darker normativity under the hostile environment of the far right. There is, however, the possibility of an AI that consciously aligns itself with ideals of social justice and egalitarianism. Not as autonomous decision making, but as part of a movement for social autonomy. This is AI as part of a wider structural renewal, supporting the withdrawal of power from hegemonic institutions and the creation of alternative structures of social organisation based on mutual aid (Landauer/Richard 2010). Reclaiming our own agency is not to attack AI as such but to challenge the system that produces AI in its own image. This is what it means to take sides with the possible against the probable. Retrieving our capacity to think collectively, learning from and with each other rather than relying on machine learning, we can counter thoughtlessness with practices of solidarity and collective care.

References

3Blue1Brown. n.d. "What Is Backpropagation Really Doing?" Deep Learning, Chapter 3. (https://www.youtube.com/watch?v=Ilg3gGewQ5U). Accessed 25 September 2018.

"A Month of Revolt in the Service Sector (#4.1)". n.d. Notes From Below. (https://notesfrombelow.org/issue/revolt-in-the-service-sector.) Accessed 12 March 2019.

ACM FAT (2019): ACM Conference on Fairness, Accountability, and Transparency (ACM FAT)'. (https://fatconference.org/). Accessed 11 July 2019.

Agamben, Giorgio (2005): State of Exception. Translated by Kevin Attell. 1 edition. Chicago: University Of Chicago Press.

Alba, Davey (2018): "Here's How Amazon Defended Its Facial Recognition Tech To Concerned Employees At An Internal Meeting". BuzzFeed News. 8 November 2018. (https://www.buzzfeednews.com/article/daveyalba/amazon-all-hands-facial-rekognition-ice). Accessed 11 July 2019.

Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen (2018): "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects". ArXiv:1811.11553 [Cs], November. (http://arxiv.org/abs/1811.11553).

Amodei, Dario, and Danny Hernandez (2018): "AI and Compute". OpenAI Blog. 16 May. (https://blog.openai.com/ai-and-compute/). Accessed 11 July 2019.

Alston, Philip. n.d. "OHCHR | Statement on Visit to the United Kingdom, by Professor Philip Alston, United Nations Special Rapporteur on Extreme Poverty and Human Rights". (https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23881&LangID=E.) Accessed 9 January 2019.

Arendt, Hannah (1998): The Human Condition. Chicago & London: University of Chicago Press.

Arendt, Hannah (2006): Eichmann in Jerusalem: A Report on the Banality of Evil. 1 edition. New York, N.Y: Penguin Classics.

Bourdieu, Pierre (1990): The Logic of Practice. Stanford University Press.

Conger, Kate (2018): "Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program". Gizmodo. 1 June 2018. (https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620). Accessed 11 July 2019.

Cuninghame, Patrick (2015): "Mapping the Terrain of Struggle: Autonomous Movements in 1970s Italy". Viewpoint Magazine. 1 November 2015. (https://www.viewpointmag.com/2015/11/01/feminism-autonomism-1970s-italy/). Accessed 11 July 2019.

Eckersley, Peter (2018): "Impossibility and Uncertainty Theorems in AI Value Alignment (or Why Your AGI Should Not Have a Utility Function)". ArXiv: 1901.00064 [Cs], December. http://arxiv.org/abs/1901.00064.

Fricker, Miranda (2007): Epistemic Injustice: Power and the Ethics of Knowing. Oxford, New York: Oxford University Press.

Ghosh, Pallab (2018): "AI Could Save Heart and Cancer Patients", 2 January 2018, sec. Health. (https://www.bbc.com/news/health-42357257). Accessed 11 July 2019.

Goswami, Gaurav, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa (2018): "Unravelling Robustness of Deep Learning Based Face Recognition Against Adversarial Attacks". ArXiv:1803.00401 [Cs], February. http://arxiv.org/abs/1803.00401.

Hern, Alex (2018): "Google 'betrays Patient Trust' with DeepMind Health Move". The Guardian, 14 November 2018, sec. Technology. (https://www.theguardian.

com/technology/2018/nov/14/google-betrays-patient-trust-deepmind-health care-move). Accessed 11 July 2019.

Illich, Ivan (1975): Tools for Conviviality. Glasgow: Fontana.

Konkel, Frank (2016): "The CIA's Classified Cloud Is Reducing Tasks from Months to Minutes". Defense One. 15 December 2016. (https://www.defenseone.com/technology/2016/12/cias-classified-cloud-reducing-tasks-months-minutes/133925/). Accessed 11 July 2019.

LaMattina, John (2016): "Is Prediabetes An Epidemic Or A Creation Of Drug Companies?" Forbes, 4 August 2016. (https://www.forbes.com/sites/johnlamattina/2016/08/04/is-prediabetes-an-epidemic-or-a-creation-of-drug-companies/#4302d76779c2). Accessed 11 July 2019.

Landauer, Gustav, and Richard Day (2010): Revolution and Other Writings: A Political Reader. Edited by Gabriel Kuhn. Oakland, CA: PM Press.

Lee, Dave (2019): "Microsoft Staff: Do Not Use HoloLens for War", 22 February 2019, sec. Technology. (https://www.bbc.com/news/technology-47339774). Accessed 11 July 2019.

Lipton, Zachary C. (2016): "The Mythos of Model Interpretability". ArXiv:1606.03490 [Cs, Stat], June. http://arxiv.org/abs/1606.03490.

McQuillan, Dan (2017): "Data Science as Machinic Neoplatonism". Philosophy & Technology, August, 1-20. https://doi.org/10.1007/s13347-017-0273-3.

McQuillan, Dan (2018): "People's Councils for Ethical Machine Learning". Social Media + Society 4 (2): 2056305118768303. https://doi.org/10.1177/2056305118768303.

Möhlmann, Mareike, and Ola Henfridsson (2017): "Uber Drivers Are Gaming the System and Even Going Offline En Masse to Force 'Surge' Pricing." 2 August 2017. (https://warwick.ac.uk/newsandevents/pressreleases/uber_drivers_are/). Accessed 11 July 2019.

Ng, Andrew (2016): "What Artificial Intelligence Can and Can't Do Right Now". Harvard Business Review, 9 November. (https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now). Accessed 11 July 2019.

Open University (1978): Lucas Plan Documentary.(https://www.youtube.com/watch?v=opgQqfpub-c). Accessed 11 July 2019.

Roy, Deboleena (2004): "Feminist Theory in Science: Working Toward a Practical Transformation". Hypatia 19 (1): 255–79.

Sian, Katy (2017): "Born Radicals? Prevent, Positivism, and 'Race-Thinking'". Palgrave Communications 3 (1): 6. https://doi.org/10.1057/s41599-017-0009-0.

Slobodian, Quinn (2019): "The Rise of the Right-Wing Globalists". The New Statesman, 31 January 2019. (https://www.newstatesman.com/politics/economy/2019/01/rise-right-wing-globalists). Accessed 11 July 2019.

Solhdju, Katrin (2015): "Taking Sides with the Possible against Probabilities or: How to Inherit the Past". In. ICI Berlin. (https://www.academia.edu/19861751/Taking_Sides_with_the_Possible_against_Probabilities_or_How_to_Inherit_the_Past.)

Artificial Intelligence
Invisible Agencies in the Folds of Technological Cultures

Yvonne Förster

1. Introduction

Democracy is all about transparency, visibility, and public engagement. In the Greek polis, political decisions were discussed in the *agora*, a public place where all citizens (in that case only free men older than 30) could listen and engage. Representational democracy today is less public, but transparency of decision processes is of the utmost importance. If a government cannot make its decisions transparent enough, it runs the risk of losing the people's trust. Transparency in a political sense implies rules, visibility, and the readiness to argue and give reasons. With the emergence of AI applications not only in the political sphere but in basically every aspect of social and private life, we are faced with new forms of opacity and nonconscious cognition, which strongly impact human decision making, behavior, movement, and communication. The central problem is that AI applications act without being able to give an account of the underlying reasons and even the underlying causal processes remain opaque (black box). If an AI used for analyzing credit rating denies credit, this decision can ruin a private life. If then reasons are not given or possibilities explained, this alone might shake people's trust in civil society. Agency based in nonconscious cognition is becoming a ubiquitous phenomenon and thus calls for ethical and phenomenological reflection. In this essay, I aim at understanding the way in which AI is experienced in terms of visibility and transparency. Toward this end I will combine phenomenological considerations with Martin Heidegger's reflections on the nature of technology.

One of the features that elicits speculation about artificial intelligence at stake here is the fact that at least for the user it is nearly impossible to understand how AI arrives at its outputs. AI applications are often characterized as black boxes (cf. Sudmann 2018a). Even if the math behind self-learning algorithms is quite straightforward, the causal processes leading from input to output are not really transparent (cf. Sudmann 2018b: 63). Obscurity is usually conceived of as a threat and potential danger. This leads to the central question of this article: Should AI

be regarded as a threat to democracy because of its invisibility? As I will argue, this is true at the surface, but I will also show that technology always comes with a certain form of invisibility. The question is whether this reaches a new level with AI. In a first step, I will define what I mean by visibility/invisibility from a phenomenological perspective. I introduce this view because it relates perception, experience to technology. Then I will clarify how this applies to the relation of human and artificial intelligence. The last part of the paper discusses the issues of the disappearance of technology and the complex relation of transparency and opacity with regard to technology. My aim is to show how AI systems introduce a new kind of invisibility or opacity to the ecological structures of the life-world.

There are at least three different layers in the interplay of visibility and invisibility involved: One goes for every object of perception: Perception is perspectival and thus invisibility is a necessary part of it. Invisibility therefore is a constitutive part of every form of perception and cognition. In the case of technology, I follow Heidegger in the diagnosis that there is a higher order form of invisibility. This is the essence of technology, which is itself not technological, but a fundamental style of thinking or revealing. This analysis of technology has a parallel in the analysis of consciousness, which is in its constitution also opaque to the conscious subject. To this extent there is nothing groundbreaking or new in terms of technology. With AI a third layer of opacity enters the stage: This is nonconscious agency—an agency that cannot give reasons but shapes lives in a very profound way. Although nonconscious agency is present also in humans and animals, technological nonconscious agency is new because it essentially shapes social and political life now and in the future. The combination of these aspects of invisibility and opacity makes up for the widespread uneasiness with AI. My aim is to give an idea how the different forms of visibility, transparency, and opacity influence the potential of AI to endanger or enable democracy.

2. Conditions of Appearance: Visibility and Invisibility

In his essay, *The Question Concerning Technology*, Martin Heidegger describes technology as a way of revealing, of bringing the concealed into unconcealment (cf. 11f.). This view is more profound than the usual instrumental view of technology as a means to an end. The character of technological artifacts is not understood adequately according to Heidegger, if this is conceived of as a tool that simply helps humans achieve particular ends. Furthermore, Heidegger also claims that seeing technology as a human doing does not capture it fully. Both notions of technology as instrumental or anthropological are not wrong. They capture technology in terms of how it is usually experienced and used. Nevertheless, they do not get to the essence of what technology is. But what is the essence of something?

Is it the thingness of a thing, that through which a thing is a thing? Is it something that does not change, while other parts or aspects may do so? In fact, it is hard to specify conceptually what the essence of something actually means.

In Heidegger's writings, at least two notions of the concept are at stake: First, the ancient Greek notion *what* something is (Heidegger 1977: 4); and, second, that of "enduring as presence" (Heidegger 1961: 59). Both aspects are relevant in his essay. The quest to understand *what* technology is determines the whole text. Heidegger is convinced that the answer to this question will not point towards an entity that is of a technological character. The essence of technology is not itself technological (cf. Heidegger 1977: 4). That means that the essence of technology is not a thing; it itself is not a physical entity. Furthermore, he holds that the essence of technology is an activity: revealing or bringing something into *unconcealment*. His claim is that it is only as a basic process or activity that technology endures.

The current discussion around AI is characterized by a similar tension. On the one hand side, intelligent technologies are conceived of simple means to ends. Processes in automation, robotics or speech recognition, to name only a few, are AI-based. These complex tasks require the ability to learn. Self-learning programs seem uncanny from the outside, but maybe not so much from the inside. Creators of such AI's usually hold that there is not much intelligence hidden in the programs. Rather it is a technological agency that reaches quite a level of sophistication, but is far from being creative beyond the limits of its training. This task-oriented functional intelligence is to be sure continually evolving, but as of now only within certain limits and on the basis of the input the AI is trained with (cf. Pontin 2018).

Public discourse, on the other hand, is fueled by threatening scenarios of a singularity transcending human powers or, less futuristically put, threats of AI erasing jobs and manipulating human behavior (e.g., targeted personalized marketing). These issues arise from AI being generally opaque (ibid.), even if it is possible to develop applications to observe AI learning processes (Sudmann 2018a). Also, the envisioned ubiquity of AI applications elicits broad discussions of the consequences for labor cultures (AI for optimizing work processes and automation) and social environments (sensor-based observation systems).

These preoccupations are related to Heidegger's discussion of the essence of technology. What might be lying at the core of our preoccupations with AI is the fact that they are (or at least are envisaged) as *world-making* technologies. Technology according to Heidegger is not the sum of physical devices but above all a style of thinking and revealing entities. This aspect is made more and more explicit within the realm of future technologies.

When we take a closer look at Heidegger's words to describe the essence of technology, the relation to visibility and invisibility is undeniable. Describing the essence of technology as something that is itself not technological gestures toward

an invisibility. The transcendental conditions of technology are themselves not of a technological or objective character. Heidegger arrives at the idea that technology is essentially a way of *world-making*. The logic of *enframing* (*Gestell*) conceives of the world as *standing-reserve* (*Bestand*), i.e., a *constellation of resources that is at disposal at all times*. He finds this logic at work already long before modern technology even emerged. While history tends to view modern physics as the enabler of modern technology, Heidegger holds that the structure or logic of technology already governs the development of modern physics (ibid.: 22 f.). The reason he gives for this claim is that modern physics as such is based on the belief that the world must be observable, measurable, and rendered predictable (ibid.: 172). Predictability is necessary in order to treat the environment as *standing-reserve*. The interplay of needs and resources is a future- and hence prediction-based endeavor. Modern physics was already driven by the goal to tame the physical world through prediction and calculability, which is most explicit in the use of AI (e.g. for facial recognition used in border control or urban CCTV applications, and predictions of consumer behavior or optimizations of workflows through management AI). This means current usages of AI expand the potential to uncover *standing-reserves* beyond the exploitation of natural resources and thereby far into the depths of human behavior. The extent of this process is not yet clear, much less its consequences and ethical challenges.

When technology constitutes the intelligibility of the world that reveals it as *standing-reserve*, as being always at disposal for our use, it also at the same time hides or conceals something. The way technology (or rather its essence, the process of *enframing*) insidiously compels humans to conceive the world as intelligible generally in terms of technology is tainted by the logic of instrumental thinking, of means and ends. It thus hides the character of objects as what stands over against subjects: "Whatever stands by in the sense of standing-reserve no longer stands over and against us as object." (Heidegger 1977: 17) The process of revealing or making visible as described by Heidegger is perspectival, and a perspective also necessarily hides other or background aspects of the perceived objects.

Visibility and invisibility condition each other in more than one aspect: In the case of technology, this interrelatedness or, to speak with Merleau-Ponty (1969), the *chiasm* (entanglement or intertwining) of visibility and invisibility goes deeper than in the case of perception. Perception is always situated and hence perspectival. There is no perception without a perspective. And that means there is no visibility without the invisible. The dialectic of visibility and invisibility constitutes perception in general.

Beyond the perception of technology as material objects/devices, which is an important topic in its own right (cf. Verbeek 2005), Heidegger sees a causality at work that is not exhausted by the instrumental definition of technology. Through technology we see the world *as* standing-reserve. Thus, technology *produces* visi-

bilities (the life-world as *standing-reserve*) rather than just adding (visible) objects to the world. As mentioned above these visibilities, or rather the all-encompassing style in which technology compels the world to appear as technological in general, also hides something, i.e., makes something invisible: namely, the objective character of things as *Gegenstände*. This opens up another aspect within the broad topic of visibility. What is a thing when its thingness or *Gegenständlichkeit* is hidden?

This is what happens when a tool like a hammer is used: The skilled user is not aware of the hammer as an object. Rather, the hammer becomes a prolongation of the body during usage. As long as the use remains frictionless, the hammer as object will not draw attention. It remains unthematic and its character as an object transparent. Such a use of things as tools is what Heidegger calls throughout his works "*readiness-to-hand*" or availability (*Zuhandenheit*): a description of a certain comportment toward things as being ready to use, being at our disposal. The instrumental attitude of technology makes things appear as means and hides their being as objects.

3. Transparency and Opacity in Technological Objects

If we translate this Heideggerian view of technology into a more common terminology, we arrive at a different form of visibility: namely, *transparency*. A tool or a technological device can be transparent in the sense that the user experience is smooth. Such a smooth user experience (or so-called "*frictionless UX*") has become the gold standard in technology design and AI is one of the means to achieve this goal. A self-learning software can ideally learn from the user what it means to function smoothly. Any disruptions within the use of applications can further serve as materials from which it can learn and then create smoother functional processes that flow without disruptions.

From a phenomenological perspective disruptions break the everyday attitude of smooth functioning and reveal the thingness and the character of objects as that which stands over against us (*Gegen-stände*). Only then will users have or find a reason to actually reflect on the technology. This also opens up the following possibility: In order to develop a critical attitude, disruption or friction is a necessary component. In neuroscience and philosophy of mind disruption or *prediction errors* is integrated in the model of neural activity as *prediction processing*: "In predictive coding schemes, sensory data are replaced by prediction error, because that is the only sensory information that has yet to be explained. (Feldmann & Friston 2010, p. 2)." (Cited by Clark 2015: 4)

Conversely, this also reveals that functional transparency is at the same time associated with being *opaque*. The constitutive processes of a functioning technology and hence a smooth user experience has to stay hidden in order to perform

this job. In that sense, technological processes are supposed to be *opaque*: They remain hidden throughout the process of usage when they function smoothly. Transparency and opacity are manifest themselves like visibility and invisibility. The difference between the two pairs of concepts is that the case of visibility/invisibility is a more neutral way to describe the givenness of objects in perception.[1] Transparency and opacity tend to have a meaning that includes a *normative* aspect. At least this is the case when we broaden the perspective toward questions of democratization or the potential of AI to foster democracy.

To explain this train of thought in more detail, let me draw a line from the phenomenological use of the concept of transparency to its application in technology. Jean-Paul Sartre uses the concept of transparency in order to describe consciousness or, more narrowly, the imaginary, i.e., modes of consciousness related to images and phantasy. Consciousness constitutes perceptions in various modes without making the constitutive process itself perceptible. It remains transparent in its functionality, meaning that it does not become part of the object presented as perceived, remembered, or anticipated. By analogy, an AI application does not itself become an item of awareness when it functions smoothly.

This becomes clear, when comparing different forms of givenness. For example, just now there is a cup of tea sitting on my desk. My act of seeing the cup of tea is an act of consciousness, an act of visual perception. This is one mode of how consciousness can present a thing: as given to vision, physically being there, within my reach. But the act of perceiving itself is not thematic, is not part of the intentional consciousness of the cup. The workings of consciousness remain transparent and they should do so, because otherwise something could be wrong with our eyesight or the overall state of health. If I remember the cup of tea later on, I will reproduce the visual characteristics of the cup through memory. The correlate of my memory is one produced by my imagination, which gives the cup to my consciousness as if I saw it. Again, the intentional act is perceived as an act of memory, but how this memory is constituted is not thematic in the memory itself. The workings of consciousness remain transparent. They are not thematic

1 Edmund Husserl describes perception in his lectures on *Thing and Space* [1907] as being necessarily inadequate in the sense of necessarily involving aspects that are not directly perceived. Perception of a thing in space is always partial, being enriched step-by-step by changes in perspective and the simultaneous quasi-perception (adumbration) of the hidden sides of the thing: "We see that the continuity of the corporeal thing presupposes 'inadequate' perception, perception through adumbrations that are always capable of enrichment and more precise determination." (Husserl 1997: 101 [121]) This notion of perception necessarily includes perception of the non-perceived. That means human perception does not only conceive of things through adding perspectives consecutively to each other. Rather we are acquainted with spatial and temporal things in such a way that the hidden sides are perceived implicitly. This is what Husserl and Merleau-Ponty call "apperceptions": The perception of the non-perceived.

in the process of cognition. In that sense, these processes are also opaque for the exploring mind. We have no conscious access to the inner workings of the mind. And this usually poses no problem.

In the case of AI, however, it is different: Not knowing how an algorithm arrived at a solution can be highly problematic. If, for example, medical data are analyzed through an AI in order to identify a disposition for cancer, it is necessary to know on which grounds a diagnosis has been generated. Only on these grounds can a decision for preventive treatment be made. The problem is that an AI can generate predictions without being able to give a *reason* for the outcome, the choice of samples, or the method used. There is a categorical difference between the causal processes leading to a mental state or an output of a program, and the ability to give reasons and reflect on mental states, as it is discussed within philosophy of mind.

One can, for example, analyze the modes of consciousness through methods of phenomenological analysis and reflect on the different modes of intentionality in a given situation. Then consciousness as a process loses its transparency. The unthematic act of remembering or imagining becomes itself object of a higher order reflection. But then also a higher order of transparency emerges, namely the focus on constitutive processes of mental states becomes itself an object of perception and hence must itself be constituted. The infinite regress looms large here. The lesson to be learned from Edmund Husserl's analysis of intentionality is that there is always a layer of consciousness that cannot itself be conscious because it itself constitutes a lower level or aspect of consciousness. Consciousness of temporal change, for example, cannot itself be temporal, at least not in the same way as the experience of time is:

> But we should seriously consider whether we must assume such an ultimate consciousness, which would be necessarily an 'unconscious' consciousness; that is to say, as ultimate intentionality it cannot be an object of attention [...], and therefore it can never become conscious in this particular sense. (Husserl 2008: 394)

Consciousness, therefore, is not only transparent as a medium of perception, it must also in some constitutive aspects remain opaque. We cannot understand consciousness simply by being conscious.

Human consciousness is deeply influenced by technology and today in particular by AI (cf. Hansen 2012, Hayles 2012, Stiegler 1998). The *technogenesis* of human consciousness, as Katherine Hayles puts it, opens up another dimension of transparency/opacity. AI is a form of nonconscious cognition (cf. Hayles 2014) that becomes more and more ubiquitous. There is no online-shopping without suggestions generated by an AI; every social media news feed is individualized by algo-

rithms and even airfares are adapted to time, location, and devices. The virtual world is highly personalized through more or less sophisticated AI applications.

Not only are the workings of the devices opaque in the sense that the user does not perceive the actual computational processes and even less so the data gathering that goes along with these processes. Even more so the output generated by AI applications does not necessarily reveal the underlying personalization processes. The Internet is only to a very limited extent a shared world. Most of the contents are shaped through user-AI interactions, though the user is not consciously aware of these interactions. Regarding technology in general, one can observe changes in human behavior and cognition with every new invention. The invention of writing, for example, has deeply altered how people memorize contents and how cultures preserve their traditions. The rise of smartphones has altered completely human ways of communicating. One simple example is communication through messaging devices and social networks: "tele-communication [...] entails a hiddenness of the face, a disappearance of the voice with its tonalities, the assuming of quasi-identities that do not authentically emanate from the concreteness of our being-in-the-world-in-the-flesh." (El Bizri 2018: 130) One could find countless examples of how new visibilities and at the same time opacities are generated through emerging technologies.

The eerie twist comes with AI. Two factors are relevant: The temporal microscale of computational processes and the predictive coding. The first factor, namely, the speed of computational processes that makes them inaccessible for human cognition, generates a scenario in which the second factor, namely, how the predictive coding turns into a preemptive force on human perception. As Mark Hansen writes in considering how computational processes that become a central element in the tissue of the life-world function on temporal microscales beyond our awareness:

> through the distribution of computation into the environment by means of now typical technologies including smart phones and RFID tags, space becomes animated with some agency of its own. One crucial feature of this animation is its occurrence largely outside—or beside—the focal attention of actants within smart environments. For this reason, the intelligent space of contemporary life offers a kind of affordance—an unperceived or directly sensed affordance—that differs fundamentally from affordances as they have been theorized, following upon the work by James Gibson, in relation to media. When "we" act within such smart environments, our action is coupled with computational agents whose action is not only (at least in part) beyond our control, but also largely beyond our awareness. (Hansen 2012: 33)

This description rests on the assumption that human cognition is constituted in relation with or by means of embeddedness in an environment. Hansen coins this as our "environmental condition" (ibid.), which describes the coupling of the individual and its environment. This coupling is not a static relationship, but a very dynamic one—a constant process of becoming. This refers to process ontologies, which either hold that consciousness emerges from being embedded in an environment (cf. Merleau-Ponty 1969, Thompson 2013), or that consciousness even extends into the environment (a version of panpsychism, cf. Chalmers 2013, Whitehead 1929). Without delving into the environmental/ecology debate, I want now to transpose these thoughts into the context of smart environments and AI driven ecologies.

Let me briefly summarize the train of thought leading up to this current juncture. I started out with Heidegger's notion of the essence of technology as enabling condition of visibility or, more concretely, rendering the world perceptible as *standing-reserve*. This aspect of technology is itself not technological; rather it is the constitutive structure of technological thinking and thus underlies and makes possible the visible materiality of technological device. From there I took a detour into how human perception is constituted and showed that visual perception is always situated and hence perspectival. That means aspects of invisibility are a constitutive part of vision or perception in general. The next step of my argumentation transposed the relation of visibility and invisibility into technological artifacts, where we speak of transparency and opacity, rather than of visibility and invisibility. Technological devices become transparent during use just as human consciousness is transparent in perception (the process of the constitution of perception, for example, is not itself object of perception). Technological device function smoothly if there is no disruption and thus no reflection on process of usage required. This transparency is always accompanied by opacity. Although the mechanisms produce functionality, the computational processes remain hidden, which is why digital technologies is often described as black boxes. This gets even more poignant with self-learning algorithms, which are not even fully understood by their programmers.

My aim is to show in the remaining sections how transparency/visibility and opacity/invisibility intertwine and establish new affordances. At this point I will go on with a reflection on smart environments. Smart environments or houses that are turned into an Internet of Things (IoT) exemplify a technology that is governed by self-learning AI, whose main function is prediction. Prediction is necessary because the IoT within a household, for example, is a highly dynamic compound of interlinked processes that has to be adaptive for all kinds of situations and changes. Ultimately, I will argue that AI-driven smart environments differ strongly from low-tech environments for two reasons: (1) predictive responsiveness has not been a common feature of environments before and (2) the prediction

and hence preemptive functionality is modelled around a conception of the ideal human/human behavior. It is here that the political discussion needs to start.

4. Smart Environments: Technologies in the Tissue of the Life-World

Intelligent technologies are being woven into the tissue or the *flesh* (Merleau-Ponty 1969, Rabari, Storper 2015, Förster 2018a) of the life-world, and it is important to understand that this is decidedly *not* a metaphor: Urban spaces consist of countless sensors, cameras, and monitors. Especially megacities like Seoul, Tokyo, London, or New York City have CCTV in literally every corner of the city. Displays are present wherever you look and sensors measuring air quality, light intensity, or listening into the noise of the city go unnoticed, even if you start looking for them. The growing density of connected devices within smart environments creates a growing demand for very small hardware, integrated devices, and high-speed data nets.

While urban spaces, work, and private spaces become more and more technological, hardware in turn becomes less visible. Sensors see without being seen, and hear without being heard. This peculiar phenomenon makes up for the narratives of future life-worlds, especially in contemporary science fiction movies. What is currently advertised or else emerging under the label of IoT or Internet of Everything (IoE) extends AI and thus nonconscious cognition into the last corners of the life-world. The topos of the vanishing of the hardware adds to the functional opacity of AI applications. Users have barely any chance to understand how AI is incorporated in devices when it is actually at work or how it shapes the process or experience of use. On top of this the physical implementation is no longer easy to locate. This means that technological environments are turned into a sensory, responsive surface with nonconscious cognition. Dealing with responsive AI driven environments requires, therefore, new forms of knowledge and behavior, such as an understanding of technological agency. Nonconscious cognition and agency make up for fairly new affordances in daily life. On the one hand side, human behavior and movement needs to be adapted to the technological systems in order for them to work properly. On the other hand side, humans need to reflect on how they want these new technologies to be integrated in their life-worlds. This is precisely the point where an active engagement with new affordances and hence novel cultural structures needs to take place.

The AI's integrated in smart environments actively shape perception, movements, emotions, and rational choices (e.g., elections, ethical choices, etc.). One of the central problems is that AI's exert their influence predominantly on the level of affects (cf. Parisi 2018). This adds a third level of opacity or invisibility: the nudges generated by AI applications are not always perceivable as such. Recom-

mendations in shopping apps are quite straightforwardly nudges. The underlying structures of newsfeed generation are much less obvious. The way we retrieve information from the Internet is always tainted through predictions of underlying learning algorithms. Thus, the world presented through a news feed is a personalized world, generated by an AI that seems to know the user, while the user does not know how the program generates its output. The opacity of nonconscious cognition and agency, as it is operative in AI applications, creates uncertainties concerning current and future social life. Current science fiction movies are symptomatic for a more nervous human condition (cf. Förster 2016). There technology tends to be portrayed as a hidden force that goes through a cognitive evolution and eventually overpowers or leave humanity behind as an outdated life form (e.g. the movies *Her* (USA, 2013) and *Transcendence* [USA, 2014]).

It is an undeniable fact that technology is becoming more and more invisible or at least smaller and more integrated within everyday objects and urban surfaces. Even the skin as a limit is slowly breaking down. Sensors integrated in the body become increasingly more normal, even though the ethical dimension of this is debated. In Sweden, for example, some 3000 people already had such sensors implanted under their skin to replace keys, credit cards, or train tickets. There are two salient characteristics of distributed AI systems today: they become part of the environment (merging in tendency with everyday objects and surfaces, such as refrigerators, surveillance cameras, or Alexa voice assistant), or else parts of devices that function in close proximity to the body or become integrated within the body (clothes with smart fabrics, jewelry, or smart implants). One could say that technology becomes naturalized, if there ever was a clear-cut distinction between the artificial and the natural to begin with.

Smart environments are largely governed by AI because the sheer amount of data generated by the distributed net of devices needs to be digested and made useful. At this stage, we are faced with a complex structure of visibilities and their counterparts. Technology as hardware starts to disappear while its function-potential increases evermore exponentially. This tendency toward invisibility generates a second- (or even third-)order transparency: Not only is technology in its usage transparent, but it becomes transparent as an object. If technology had lost or hidden is object-character already according to Heidegger's consideration, we are now reaching a higher level of *enframing* or *Gestell*. In Heidegger's view technology obstructs our view of the world as object because it compels us to conceive of the world as *standing-reserve*. This basic characteristic also holds for technology: It does not appear as an object that stands over against a subject and, in this respect, transcends human aims. It is a means to an end that makes the environment appear as a predictable, calculable reservoir of potentialities. This new layer of transparency comes into play through the disappearance of technological devices and the emergence of distributed AI. This makes a difference because most of what AI

today has accomplished is predictive and shapes functional processes according to those predictions. For the users, the unknown factors are huge: Users cannot know how exactly the system (e.g., IoT) or even one single device works. Moreover, they cannot know or actively experience better which, when, and how much data are gathered from the usage and behavior relating to these devices. Much less can be known of the use that is made of the collected data. The Big Data problem is becoming discussed ever more widely, and the complexity grows with the increasing use of devices by the hour.

From a phenomenological perspective, the decreasing visibility (and opacity) of smart technologies and their increased potential for agency is problematic for a democratization of AI. And the problem is not AI itself, whose actual intelligence is amazing but also constantly overrated. AI does not have the intention to build a better self, a better society, or a better future. Human beings aim for that. Philosophy is not a stranger to such mostly exaggerated goals. One of the obstacles to a transparent use of AI is this striving towards perfectibility, which is more or less an economic vehicle. Smart technologies have the potential to be useful and maybe even create a better future, but only if a culture of critique and open discourse can be established and sustained. How does this point relate to the topic of visibility? Let me refer to Heidegger one last time. He argues that technology lets the world appear as *standing-reserve*. Today we should ask how human lives appear through technology. How do humans paint an image of human life by creating an environment that is machine-friendly? Do we have the means to make the hidden ratio of what it means to be human flourish in smart environments? How can we create enough freedom and potential for creative agency that allows for an active and critical engagement with existing technologies? That would imply experimenting, tampering, and first and foremost, conducting a critical discourse with industries relying heavily on predictions like retail and insurance businesses. The image of a "good" human life should be scrutinized (also with regard to the concept of the anthropocene). We need, therefore, a close observation of how nudges, prediction, and preemption influences everyday behavior—how we speak, move, and, indeed, smile or love. To do this successfully, humans in their whole range of diversity need to become visible and present as voices in public and in the industries that rely heavily on AI. The *political* dimension of AI is very much a human one. The *human* image built into intelligent technologies needs to be made visible. Only then can an ethical discussion properly take place.

References

Binfield, Kevin, ed. (2004): Writings of the Luddites. Baltimore: Johns Hopkins University Press.

Chalmers, David J. (2013): "Panpsychism and Panprotopsychism." In: *The Amherst Lecture in Philosophy* 8 (2013): 1-35 (www.amherstlecture.org/chalmers2013/).

Clark, Andy (2015): "Embodied Prediction." In: Thomas Metzinger & Jennifer M. Windt (Eds), *Open MIND:* 7(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570115

El-Bizri, Nader (2018), "Phenomenology of Place and Space in our Epoch: Thinking along Heideggerian Pathways." In: Erik Champion (ed.), *Phenomenology of Real and Virtual Places*, London: Routledge, 123-143.

Förster, Yvonne (2016): "Singularities and Superintelligence: Transcending the Human in Contemporary Cinema." In: *Trans-Humanities*, Seoul: Ewha Institute for the Humanities (EIH), 33-50.

Förster, Yvonne (2018a): "From Digital Skins to Digital Flesh: Understanding Technology through Fashion." In: *Popular Inquiry* (2), Aalto University Helsinki (www.popularinquiry.com/blog/2018/8/30/yvonne-foerster-from-digital-skins-to-digital-flesh-understading-technolgy-trough-fashion).

Förster, Yvonne (2018b): "Wenn künstliche Intelligenz laufen lernt. Verkörperungsstrategien im *Machine Learning*." In: Christoph Engemann und Andreas Sudmann (Eds.), *Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*, Bielefeld: transcript, 325-340.

Hansen, Mark B. N. (2012): "Engineering Pre-Individual Potentiality. Technics, Transindividuation, and 20th-Century Media". In: *SubStance* 41/3 (Issue 129), 32-59.

Hayles, Katherine (2012): *How we Think. Digital Media and Contemporary Technogenesis*, Chicago, London: University of Chicago Press.

Hayles, N. K. (2014): "Cognition Everywhere: The Rise of the Cognitive Nonconscious and the Costs of Consciousness". *New Literary History* 45(2), 199-220.

Heidegger, Martin (1961): *An Introduction to Metaphysics*, New York: Doubleday.

Heidegger, Martin (1967): *Being and Time*. Oxford, Cambridge MA.: Blackwell.

Heidegger, Martin (1977): "The Question Concerning Technology". In: *Essays*, New York, London: Garland, 3-35.

Husserl, Edmund (2008): "On the Phenomenology of the Consciousness of Internal Time (1893-1917)". In: Rudolf Bernet (ed.), Edmund Husserl. *Collected Works IV*, Dordrecht: Springer.

Merleau-Ponty, Maurice (1969): *The Visible and the Invisible*, Evanston: Northwestern University Studies.

Parisi, Luciana (2018): "Automated Cognition and Capital". In: Warren Neidich (Ed.). *The Psychopathologies of Cognitive Capitalism: Part Three*, Berlin: Archive Books.

Pontin, Jason (2018): "Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning". In: *Wired.com*: www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/ (26.03.2019).

Rabari, Chirag, Storper, Michael (2015): "The digital skin of cities: urban theory and research in the age of the sensored and metered city, ubiquitous computing and big data." In: *Cambridge J Regions Econ Soc* 8 (1), 27-42. https://doi.org/10.1093/cjres/rsu021

Stiegler, Bernard (1998): *Technics and Time, 1: The Fault of Epimetheus*, Stanford: Stanford University Press.

Sudmann, Andreas (2018a): "On the Media-political Dimension of Artificial Intelligence. Deep Learning as a Black Box and Open AI." In: Mathias Fuchs, Ramón Reichert (eds.): *Digital Culture & Society (DCS)*, Vol.4/1, *Rethinking AI: Neural Networks, Biometrics and the New Artificial Intelligence*, Bielefeld: transcript, 181-200.

Sudmann, Andreas (2018b): "Szenarien des Postdigitalen. Deep Learning als MedienRevolution". In: Christoph Engemann, Andreas Sudmann (eds.): *Machine Learing. Medien, Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, 53-73.

Thompson, Evan (2007): *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Cambridge MA, London: Harvard University Press.

Verbeek, Peter-Paul (2005): *What Things Do: Philosophical Reflections on Technology, Agency, and Design*, University Park: Pennsylvania State University Press.

Whitehead, Alfred North (1929): *Process and Reality an Essay in Cosmology; Gifford Lectures Delivered in the University of Edinburgh During the Session 1927-28*, New York: Free Press.

Race and Computer Vision

Alexander Monea

Introduction

Any analysis of the intersection of democracy with AI must first and foremost engage the intersection of AI with pre-existing practices of marginalization. In the United States, perhaps no intersection is more salient than that of AI and race. As AI is increasingly positioned as the future of the economy, the military, state bureaucracy, communication and transportation across time and space, in short, as the bedrock of humanity's future, questions of how AI intersects with pre-existing practices of racial marginalization become central. These questions are particularly difficult to answer given the black boxed nature of most contemporary AI systems. While it is certainly a worthwhile endeavor to push for increasing transparency into the datasets and algorithms powering AI systems, that transparency lies in an anticipated future and cannot help us now to analyze the operations of current AI systems. This picture is only complicated by the fact that AI systems, particularly those operating at web scale, are difficult for even their engineers to understand at later stages in their operation. For instance, a programmer may be able to easily describe the seed data and the machine learning algorithm that she started with, but may be completely unable to explain the rationale behind the subsequent classifications that the system learns to make. Again, it is certainly worthwhile to call for AI explicability—namely, requiring AI programmers and engineers to develop systems that can explain their decision-making processes or, in the most extreme case, *only make decisions that can be explained clearly to a human*—this again is an anticipated future that is of little use to answering the immediate question of race and AI which already has dire consequences at this very moment.

So how can an outsider go about critically analyzing the intersection of race and AI in the contemporary moment? This paper will utilize an interdisciplinary methodology that I am calling 'speculative code studies', which combines archival research into press releases, company blog posts, science and technology journalism, and reported instances of technological irregularities with critical code studies research into the available datasets that machine learning algorithms are trained on, analyses of open-sourced variants of black boxed code, and empirical

studies of the outputs of black boxed systems. The goal of such a study is to produce a rigorous, but speculative, analysis of black boxed code. This analysis must always remain speculative, as the actual systems are obfuscated from direct analysis, but the methodology ensures that the analysis is as rigorous as possible given the peripheral materials, segments of code, and inputs/outputs that are available. In this chapter I use this method to probe the myriad ways in which racial biases that are present both in the boardrooms and research and development wings of technology companies *and* in the broader socio-cultural milieu get hardcoded into the datasets and, subsequently, the machine learning algorithms built atop them. It will be my argument that many of these algorithms constitute a material manifestation of racial bias.

This paper will primarily be concerned with visual or optical media, and computer vision algorithms in particular. I will argue that within this context blackness falls into the notorious dialectic of hypervisibility and invisibility—blackness is too often rendered in stereotypes, at times even visually cartoonish, or it is rendered as systemically invisible. However, it is important to note that while the context of my analysis is systemic racism against predominantly African Americans, these systems have global impacts for people of color—or, to speak more precisely, those darker-skinned individuals who fall within types V or VI of the Fitzpatrick scale (cf. Fitzpatrick 1975, 1988). When I use terms like 'black,' 'person of color,' or 'dark skin,' it is meant to indicate that the problems I am identifying are of global concern and have high stakes impact on people at the darker end of the Fitzpatrick scale across the planet, even though my analysis is contextualized within the history and culture of the United States. It is outside the purview of this chapter to extend this analysis to other conjunctures, but I sincerely hope others will help me to do so by extending, revising, and challenging this work. In the first section of this chapter I will draw on critical race theory to demonstrate how this dialectic is problematic from the perspective of egalitarian democracy. In the next section I will offer a brief overview of the history of racial bias in visual media within the context of the United States that perfectly illustrates this hypervisible/invisible dialectic of blackness. In the following two sections I will look at the hypervisibility and invisibility of blackness in contemporary AI systems and will try to demonstrate the enormity of the stakes of this conjuncture. In conclusion I offer some preliminary thoughts about where we can go from here.

The Hypervisibility and Invisibility Dialectic of Racial Difference

> "Black is … 'an black aint."
> *Ellison 1989, p. 9*

In their theory of racial formations, Michael Omi and Howard Winant (2015) have argued that humans essentially use stereotypes to make sense of the world, even though these stereotypes are constantly changing. People make use of fundamental categories of difference, like race, gender, class, age, nationality, and culture to navigate society, all of which imply a certain politics of "othering" that produces structural marginalization, inequality, exploitation, and oppression. We might productively understand machine learning as engaging in a very similar behavior with similar political stakes. As I have argued elsewhere in the context of machine learned semantic labels, these algorithms engage in an iterative process of learning stereotypical differentiations to categorize the various data that they encounter (Monea 2016, 2019). However, race functions differently than these other stereotypical categorizations because, as Omi and Winant explain, race is crucially 'corporeal' and 'ocular'. What they mean by this is that racial distinctions take hold of a set of phenotypic differences—most noticeably morphological differences like skin tone, lip size, and hair texture in the case of blackness in the United States—and essentializes them, as if they were physical markers of an essential difference of kind (2015: p. 13). It is thus othering, as it establishes the border between an 'us' and a 'them,' and reifies that border by making it appear as a fundamental law of nature, a scientific fact, a marker of a different kind of being. There are two unique aspects of this process, which Omi and Winant refer to as 'racialization'. First, these phenotypic differences are arbitrarily selected, are not understood as having the same denotations and connotations across space and time, and often were previously unconnected to any racial classification. Second, they are written on to the body through morphological distinctions in such a way that racial difference is legible on sight alone.

This latter aspect of racialization has been a core component of critical race theory for decades, and was perhaps most notably articulated by Frantz Fanon (1967) in his concept of 'epidermalization'. For Fanon, epidermalization is a process by which black people realize their identification as the Other for white people as they encounter the white gaze that dissects and analyzes their body, without permission, to classify them. Fanon writes:

> I am overdetermined from the outside. I am a slave not to the "idea" others have of me, but to my appearance. […] The white gaze, the only valid one, is already dissecting me. I am *fixed*. Once their microtomes are sharpened, the Whites objectively cut sections of my reality. (1967: p. 95)

Stuart Hall succinctly defines Fanon's idea of epidermalization as "literally the inscription of race on the skin" (1996: p. 16). Hortense Spillers similarly writes about the 'hieroglyphics of the flesh', wherein black subjects are transformed into flesh through "the calculated work of iron, whips, chains, knives, the canine patrol, the bullet" (2003: p. 207). For Spillers, Western humanism is built atop these hieroglyphics, as the liberated Man requires definitionally that an other be designated as not fully human. This legacy is passed down through the generations even after black subjects were granted possession of their own bodies and continues to structure our social lives. As Alexander Weheliye describes it, racial categories "carve from the swamps of slavery and colonialism the very flesh and bones of modern Man" (2014: p. 30). Sylvia Wynter (2001) has similarly shown how this happens in her arguments about sociogeny, where a focus on phenotypical differences is just a ruse to essentialize racial difference and divide the Homo sapiens species into humans and nonhuman beings.

We can understand this corporeality, ocularity, epidermalization, and fleshiness of race as a fundamentally *visual* component, and one that makes race hypervisible by stressing phenotypes—especially morphological features like skin tone, lip size, and hair texture—and connoting racial stereotypes that help bolster racial marginalization. Lisa Nakamura (2007) has shown how this hypervisibility of race is perpetuated today by computation and digital visual culture, leading to the production of 'digital racial formations'. John Cheney-Lippold (2017) has similarly shown how algorithms have digitized race into 'measurable types', or statistical probabilities based on user data. Here I'd like to introduce the term 'users of color' to replace the term 'people of color' for this digital context. Users of color are digitally racialized based on algorithmic analysis of big data, which is reifying some old phenotypic stereotypes of racial difference at the same time that it is producing new ones. We might also follow Simone Browne (2015) and think of this process as a 'digital epidermalization'.

Now, we can easily see that even for technology companies that may have little ethical commitment to egalitarian democracy, the public relations nightmare alone of being seen as reifying racial stereotypes in digital culture is a huge deterrent. As we'll see throughout this piece, the most frequent response to criticism of algorithmic racialization is to make race invisible. Rendering blackness invisible is always the flipside of the coin, in dialectical tension with racial hypervisibility. Both options are unsatisfactory, as they both provide safe haven for racism, albeit in different ways. The most recent and visible example of rendering race invisible is the 'color blind' policies that have been pursued in the United States since the 1960s. These policies have been near universally condemned by scholars in critical race studies and related disciplines (e.g. Bonilla-Silva 2017; Brown et al. 2003; Omi/Winant 2015). Color blindness delegitimates affirmative action and similar programs striving for racial equality, allows racism to operate unchecked

provided it uses dog whistles and other careful language to obscure racial malice, and obscures important racial data trails that might otherwise have been used to uncover statistical trends of racism (e.g. in policing, court sentencing, allocation of welfare benefits, etc.). In addition, it has made speaking about race so taboo that the term 'white fragility' was coined to describe contemporary white people's inability to openly talk about race and racism (Dyson 2018). Thus, rather than deal with racialization's roots in colonialism and slavery and doing the hard work of moving towards actual egalitarian democracy, blackness is alternately rendered as hypervisible or invisible, both of which leave much to be desired. In short, as Ralph Ellison put it, "Black is … 'an black ain't" (1989, p. 9).

Accounting for the Visibility of Race in Visual Media

> "Photography is a weapon"
> —*Oliver Chanarin (PhotoQ 2015)*

The United States has a long history of embedding racial stereotypes in its visual media and communications technologies (cf. Dyer 1997, hooks 1992, Nakamura 2007). This legacy spans from analog to digital photography and, as we'll see, continues to impact a number of computer vision applications. For example, Lorna Roth (2009) has shown in detail that Kodak optimized its entire suite of products for white skin. Kodak produced a long series of "Shirley cards", named after Shirley Page, the first studio model for the photos Kodak sent out with its new products. These Shirley cards were marked "normal" and used as test cards for color balancing film stock and printers. The optimization for white skin was to the detriment of people of color, whose features increasingly disappeared in direct correlation with how dark their skin was. One legendary result was French film auteur Jean-Luc Godard's refusal to shoot on Kodak film for an assignment in Mozambique because of the racial bias hardcoded into the film—it literally would not work in Africa. It was not until complaints were made by companies trying to photograph dark objects for advertisements that Kodak developed film that could capture the details of black flesh, a project that was kept quiet at Kodak and referred to via the coded phrase "To Photograph the Details of a Dark Horse in Low Light" (Broomberg & Chanarin n.d.).

Even when film was made with the explicit intent of rendering black skin visible—rather than for rendering chocolate bars and wooden furniture visible in advertisements—deeply rooted racial problems cropped up. Take, for instance, Eric Morgan's (2006) story of the Polaroid ID-2 camera. The ID-2 was designed to take two photos per self-developing print, one portrait and one profile image, of a subject 1.2m from the lens. This was, in essence, a streamlining of Alphonse Bertillon's anthropometric identification system for state policing, which has always intersected in complex

ways with existing practices of racial and gendered marginalization (Browne 2015; Fair 2017; Lyon 2008; Wevers 2018). In the case of the ID-2, this was manifested in a special "boost" button for the flash that would make it around 40 percent brighter, the same amount of light that darker skin absorbs. In and of itself, this feature is rather innocuous. It is actually a step forward in the sense that it allowed the camera to capture the features of dark-skinned people more clearly, although still made it problematic to capture both black and white skin in the same picture. However, in 1970, a Polaroid chemist named Caroline Hunter uncovered evidence that Polaroid was making a lot of money selling ID-2 cameras to the South African government, which used them to make the passbook's that black citizens were forced to carry with them at all times within white areas (see Savage 1986).

In their exhibit titled after Kodak's coded phrase "To Photograph the Details of a Dark Horse in Low Light," South African artists Adam Broomberg and Alex Chanarin argue that the ID-2 was designed for the purpose of supporting apartheid (PhotoQ 2015; Smith 2013).[1] It is unlikely that this is the case, but it is certainly true that a certain number of Polaroid executives had a damning amount of knowledge of the trafficking of ID-2s in South Africa and worked to intervene too late and with too little energy to actually prevent the use of their new technologies for the support of Apartheid. Neither of these stories is meant to minimize the technical difficulties of capturing darker features and objects on film. These difficulties are inherent to optical media and would likely exist no matter the sociocultural context within which photographic technologies arose. What they do expose, however, is how absent and unimportant the dialectic of black invisibility and hypervisibility was to these companies, a fact that was made materially manifest in their research and development paradigms, the products they took to market, the discourse they used to position these products, and their responses to criticism from the public. It is not that these devices themselves are racist, but instead that racial biases from the context in which they are developed inflect how the companies approach research and development, imagine certain products and not others, prioritize some more highly than others, assess some potential bad applications of their technology as negative press coverage to be accounted for in the design process, etc. All of this is materially manifest in the technology itself, the way it is positioned discursively, and the myriad uses it gets put to in society.

It wasn't until the '90s that American companies really began to take seriously the need to better capture black features on film. Kodak released 'black', 'Asian', and 'Latina' Shirley cards and began marketing Kodak Gold Max film, which had an 'improved dynamic range'—it could finally 'photograph the details of a dark horse in low light'. These achievements were made by new innovations in the chemical composition of film negatives to make them more reactive with darker pigments and modi-

1 For more on Broomberg and Chanarin's exhibit, see O'toole 2014.

fications to color balancing techniques to optimize for different skin tones. They all required an increasing focus on the visibility of people of color that spanned from the consumer base to the research and development teams at companies like Kodak. This change in priorities also quickly manifested in digital camera and camcorder technologies. For example, in 1994 the U.S. Philips Corporation filed for a patent on a new 'tint detection circuit' that could automatically adjust the tint in a digital image to white balance for both light and dark skin tones at once. In this instance (see Figure 1 below), there are literally two components soldered onto the board, one for 'SKINR' and one for 'SKINB' that lead to an 'AND CIRCUIT' that combines their outputs to optimize for both. While this solution is not ideal, as the color balancing will be off for both light-skinned and dark-skinned people if they are in the same shot, it is a step in the right direction. It will work equally well for any skin type in isolation and when in the same frame will average towards the middle, rendering both poles of the Fitzpatrick scale equally less visible. We might also take this historical arc as emblematic of the level of intervention necessary if we are to make any given technology democratic as it pertains to racial marginalization. Just as we need to literally hardcode anti-racism measures into the circuit board to get an egalitarian camera, we will need to literally hardcode anti-racism measures into our machine learning algorithms to get egalitarian AI.

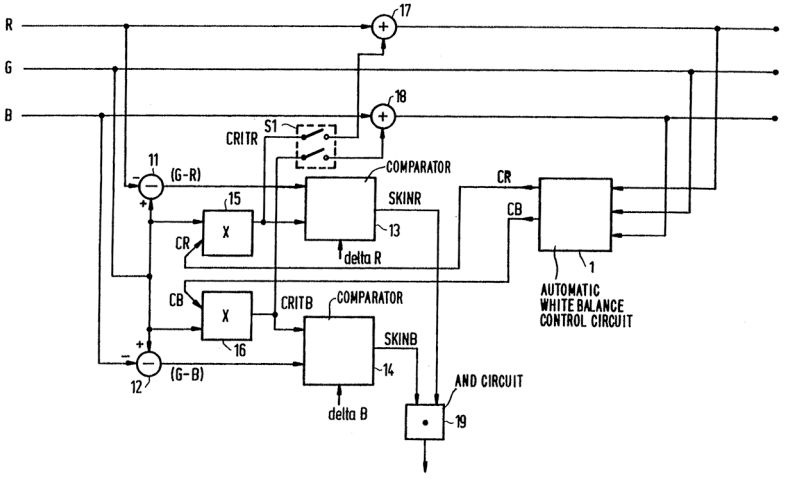


Fig. 1: U.S. Patent No. 5,428,402 (1995)

The Hypervisibility of Race in AI Systems

In a 'humorous' story read for *This American Life*, David Sedaris describes the strange cultural differences surrounding Christmas between the United States and the Netherlands, noting to refined laughter that in the Netherlands, Saint

Nicholas is accompanied by six to eight black men who prior to the mid-1950s were characterized as his personal slaves (Sedaris 2001; cf. Sedaris 2004). In Dutch tradition, these black helpers eventually stabilized into the image of Zwarte Piet, or 'Black Pete', with Dutch men and women dressing up in black face—black face paint, red lipstick, curly black-haired wigs, a golden hoop ear ring, and colorful Spanish/Moorish outfits—to lead Sinterklaas, or Saint Nicholas, in parades, distributing candy and kicks to good and bad children respectively. The majority of Dutch citizens have very positive attitudes about Zwarte Piet, and often downplay the connections between him and black face by noting alternately either that his face is colored such because he is a chimney sweep that now crawls through chimneys to deliver candies to good children who leave their shoes out or because he is a Moor that was adopted by Sinterklaas, who lives in Spain in the offseason rather than at the North Pole, as in American traditions. As Allison Blakely notes, both of these explanations are rather unconvincing (2001: pp. 47-48).

This desperate attempt to preserve a deeply problematic tradition are only complicated by the fact that Zwarte Piet is the name for the Devil in Dutch folklore, who is caught and chained for the celebration every year, and by the Netherland's complicated history with colonialism, the slave trade, and slavery—for example, the term 'apartheid' comes from the Dutch and arose particularly in the context of their colonial occupation of Surinam (Blakely 1980: p. 27). While the Dutch position themselves as a nation apart from colonial and racial marginalization, this has never been the case, especially since a number of Surinamers relocated to the Netherlands after the 1950s and have faced systemic marginalization based on race (Blakely 1980). It is no wonder that these same Surinamers are increasingly unenthused with the Zwarte Piet tradition and argue that it is insulting, and especially damaging to children of color who subsequently face bullying at school (Blakely 2001: p. 48).

David Leonard has argued that a more appropriate metric for determining culpability in instances of blackface is not whether the person dressing in blackface meant to offend people, but whether that person is causing harm, either to individuals or to society (Desmond-Harris 2014). This is a much smarter way of analyzing the situation, as it not only rids us of complex interrogations of intentionality, but also opens up the analysis of how photographic or videographic media might extend the range of impact of that harm. For instance, barring for the moment the issue of blackface in the Netherlands, it is certainly the case that it is extremely damaging in the context of the United States. As C. Vann Woodward (2001) has shown the campaign of pseudo-scientifically legitimated racism, dehumanization, segregation, disenfranchisement, and terror waged against African Americans that we know as "Jim Crow" was named after a blackface minstrel routine. Blackface thus encapsulates rather neatly the logic of American racism, in that it literally denotes and makes hypervisible phenotypes—and in still images

these are primarily morphological traits—while at the same time always preserving their connotation of inferiority in connection with white supremacy.

If we look at the case of ImageNet, we can clearly see how these Zwarte Piet images escape their context and cause social harm. ImageNet is a large dataset of labeled images first launched out of Princeton University in 2009. The dataset originally drew 80,000 labels from the semantic database WordNet—each label is referred to as a 'synset' which is a set of synonymous terms—with the goal of populating each of these labels with 500-1,000 clean and full resolution images (Deng et al. 2009). In 2010, ImageNet launched its annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has since served as an industry benchmark for success in computer vision applications (Russakovsky et al. 2015). This centrality was cemented by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton's (2012) groundbreaking success in using neural networks to win the competition to produce an algorithm that could learn to classify images in the ILSVRC (cf. Sudmann 2016, 2018). What is important to know is that ImageNet not only serves as the performance benchmark for nearly all computer vision systems, but that because those systems are trained on and optimized for the ImageNet dataset, any biases in the ImageNet dataset have wide-ranging repercussions since they subsequently become hardcoded through machine learning into a large portion of computer vision systems.

Fig. 2: Zwarte Piets and Sinterklaas (Splinter 2010[2]*)*

2 N.b. I could not establish the licensing for any of the Zwarte Piet images contained in ImageNet, but any search on Flickr for the term returns images like this one which are illustrative.

ImageNet gathers images for a synset that contains the terms "Black", "Black person", "blackamoor", "Negro", and "Negroid", which it defines as "a person with dark skin who comes from Africa (or whose ancestors came from Africa)". This synset is interesting for a number of reasons. First, just from browsing it, one can tell that it contains many fewer images as a percentage of the total images than other synsets that would contain useful visual details for building out a classifier. What I mean by this is that an inordinate number of photos are low resolution, don't show facial details, have black people's bodies positioned further away from the camera, and inordinately feature celebrities (about 1 per cent of the entire dataset is pictures of Barack Obama) and memes. While this certainly isn't a smoking gun for racial bias, let alone intentional racial bias, it does reveal that the capacity to accurately identify black facial features is not prioritized by default for any computer vision algorithms trained on ImageNet's data. What is perhaps closer to a smoking gun is that of the 1,286 images for this synset that are still available online, a full 79 of them are of people in blackface.[3] All but one of these images are of people dressed as Zwarte Piet. Thus, this odd Dutch phenomenon has the exact consequences that are feared in critiques of blackface: regardless of the intentions of those wearing blackface, when it enters public discourse, this signifier of blackness quickly escapes any contextualization and instead reifies racism through its connotations of white supremacy and its denotations of blackness being reducible to phenotypical difference. We can see this quite literally in the case of ImageNet, where blackface images have escaped their context to compose just over 6 per cent of the entire dataset, a dataset developed on a continent where the vast majority have never even heard of Zwarte Piet. Even if we allow the extremely dubious argument that these images are harmless in the Netherlands, a tradition beloved by fewer than 20 million has helped to bias a fundamental piece of infrastructure for computer vision.

Take, for example, Google's use of the ImageNet dataset. In 2014, Google researchers won the ILSVRC challenge with their 'Inception' algorithm (also known as 'GoogLeNet'), a 22-layered convolutional neural network (Syzegedy et al. 2015). This CNN was trained on the ImageNet dataset, and thus internalized any biases present in that data.[4] At Google I/O 2015, Google announced the launch of its new

3 At this time, I can only work from the publicly accessible list of links to images for given synsets. This means that I cannot access images that have since been taken down from the web, which makes it impossible to access the full set of 1,404 images that make up the dataset for that synset. I have repeatedly submitted requests to register for an account with ImageNet so that I might access the full dataset and have sent multiple emails to ImageNet's contact address, but have yet to get either access or a response.

4 While there is no comparable data for racial biases, it has been demonstrated that in the case of gender biases, neural networks not only internalize the biases in their datasets, but *amplify* them (Zhao et al. 2017).

Google Photos software. Google argued that humans were now taking over a trillion images a year and at this rate would need a second lifetime to label, organize, and revisit their photos. Google Photos was the solution we had all been waiting for:

> Google Photos automatically organizes your memories by the people, places, and things that matter. You don't have to tag or label any of them, and you don't need to laboriously create albums. When you want to find a particular shot, with a simple search you can instantly find any photo—whether it's your dog, your daughter's birthday party, or your favorite beach in Santa Barbara. And all of this auto-grouping is private, for your eyes only. (Google 2015)

The new Google Photos software was primarily powered by the Inception/GoogLeNet algorithm that was trained on ImageNet data, though it supplemented the image patterns it learned from ImageNet with a huge database of photos and nearby text from websites it had crawled and a few other indicators, like looking at the place and time stamps of both the user during the search and the images via their metadata (Brewster 2015). The centrality of ImageNet is no wonder, as it not only serves as the benchmark for computer vision algorithms and is standard across a large portion of the industry, but Google Photos was developed under Bradley Horowitz, Google's Vice President of Streams, Photos, and Sharing, who previously saw the value in Flickr's Creative Commons licensed images when he helped purchase the company as an executive at Yahoo (Levy 2015).

The point to be taken from all this is that Google Photos was designed by people who viewed ImageNet as an unquestioned industry standard and who placed strong faith in the utility of Flickr images. They thus were ill positioned to foresee the racial biases inherent in the visual data that their machine learning algorithms had used to develop their classifiers. This problem came to the fore just a month after Google Photos was released when in June 2015 a black software engineer named Jacky Alcine posted a set of images run through Google's photo tagging software to Twitter in which images of him and his girlfriend were labeled as photos of 'gorillas' (Alcine 2015).[5] The case clearly hit a nerve, as dozens of articles were published calling Google's algorithms racist within days and it has since been one of the most frequently cited examples of algorithmic bias in technology journalism. This comparison is only possible based on an arbitrary definition of certain phenotypical differences as the sole markers of an essentialized differ-

5 It is worth noting that while Google is the highest profile instance of this happening, it is by no means the sole incident. Just a month prior to the Alcine incident Flickr made news for mislabeling a black man (and a white woman) as "animal" and "ape", and also labeled photos of Dachau concentration camp as "jungle gym", "sport", and "trellis" (Hern 2015).

ence. In short, when we think of skin tone, lips, and hair as the cornerstone of racial difference, this slippage between classifiers is opened up. And further, when we essentialize these differences and connect them to racist stereotypes, the connotation of this slippage becomes unbearable. It calls into question not only people's intelligence, but their very humanity. This is a problem of the highest stakes for all users of color.

As can be seen from Google's response, the company similarly understood this instance to be a serious problem and one that might threaten the future of their computer vision platforms. Within hours, a Google engineer named Yonatan Zunger was responding to Alcine's tweet asking for permission to examine his Google Photos account to figure out what had gone wrong. The next day, Zunger tweeted that Google had not recognized a face in the images of Alcine and his girlfriend at all and noted, "We're also working on longer-term fixes around both linguistics (words to be careful about in photos of people [lang-dependent]" and in "image recognition itself. (e.g., better recognition of dark-skinned faces) (Zunger 2015). Zunger promised that Google would continue to work on these issues, which included developing systems that could better process the different contrasts for different skin tones and lighting. A few days later a Google spokesperson told the BBC that, "We're appalled and genuinely sorry that this happened. We are taking immediate action to prevent this type of result from appearing" (BBC 2015). Yet, as we'll see in the next section, Google has yet to discover a solution for this problem. The datasets its algorithms are trained on make it so that race must be rendered either as a hypervisible emphasis on phenotypes, which, without a heavily curated new dataset will continue to produce slippage between users of color and apes, or as invisible.

The Invisibility of Race in AI Systems

While the speed and sincerity of Google's initial response seemed promising, after more than two years *WIRED* reported that all Google had managed to do was remove potentially offensive auto-tags for terms from its Photos software (Simonite 2018). In that same report, *WIRED* noted the results of a series of experiments they had done with Google Photos. First, they ran a collection of 40,000 images well-stocked with animals through the system and found that it did not locate any results for the terms "gorilla", "chimp", "chimpanzee", and "monkey". In a second experiment they tried uploading pictures solely of chimpanzees and gorillas and found that it still would not recognize the offending set of terms. In a third and more damning test, *WIRED* uploaded a collection of over 10,000 images used for facial-recognition research. Searching these photos for auto-tags of "black man", "black woman", or "black person" only delivered photos that were in black-and-white, which did correspond to the gender specified, but did *not* sort

people by race. In short, in response to this public relations disaster, blackness has become invisible in Google Photos.[6] This color blindness also extends to Google's Open Images dataset, which contains "30.1M image-level labels for 19.8k concepts, 15.4M bounding boxes for 600 object classes, and 375k visual relationship annotations involving 57 classes" (Kuznetsova et a. 2018). None of these millions of labels, thousands of concepts, or hundreds of classes, from what I can gather after examining the database, explicitly label race. Users of color are only identified by their absence.

The issue of black invisibility has a long history in systems that process visual data for applications like facial recognition and motion-sensing. Take, for example, the viral 2009 YouTube video of an HP laptop designed to use facial recognition to track users' faces and follow them as they move with the webcam that failed to register the movements of 'Black Desi' at all, despite easily following the motions of his white coworker (wzamen01 2012). *Consumer Reports* (2009) tried to debunk the argument that this was a racial bias by arguing that it instead is a factor of lighting conditions, and while they present their results as if the system would work the same for lighter and darker skin tones in the same lighting, in their video it is clear that this is not the case, as they have to better light their user of color's face before the system starts to track him. For another example, take Xbox Kinect, which in 2010 was reported to have trouble recognizing the faces of users of color (Ionescu 2010). This primarily effected their ability to automatically log in to their avatars, as Kinect gameplay largely functions on skeletal movement. In other words, Kinect is capable of seeing black bodies but not black faces, and can facilitate their gameplay so long as it doesn't need to recognize their face, which some games do. *Consumer Reports* (2010) similarly argued that this was merely a lighting issue and claimed to have 'debunked' the idea that Kinect is 'racist'.

It is certainly the case that these machines themselves are not intentional agents engaged in prejudicial thinking, but to merely wave away the claims of racism as 'debunked' after demonstrating that failure to function appropriately when utilized by people of color requires an uncomfortable amount of hubris and near total lack of empathy. There is something clearly going on here and it has a felt impact on racialized bodies so clear in the videos of people attempting and failing to utilize facial recognition and motion-sensing technologies. At the very least, these instances are material embodiments of some combination of the lack of forethought in the research and development phase and a lack of concern over

6 Interestingly, this invisibility plays out differently in Google's less publicly visible computer vision projects. Google's Cloud Vision API launched in 2016 as a new tool to make its computer vision algorithms available particularly to developers (Google 2016). Cloud Vision API still uses labels like "chimpanzee" and "gorilla", though I've found no studies to date of whether the same racial hypervisibility of Google Photos persists on Google's Cloud Vision API.

going to market with a product that would fail to operate for a large and protected minority class of citizens. And further, the common arguments that all technologies fail, that these technical constraints are unavoidable, and that it is common sense to design products based on your majority market (and in most instances this is code for 'white' people) are all inadequate at best, and deeply offensive at worst.

Take, for another example, the frequent instances of motion-activated devices like soap dispensers in public restrooms failing to recognize users of color (e.g. Fussell; Plenke 2015). Bharat Vasan, the COO of Basis Science, explained to *CNET* that there are systems that can avoid this problem by detecting the darkness of objects beneath their sensors and adjusting a spotlight to match, but these systems are too expensive for many of these lower-end motion-activated devices (Profis 2014). It is here that chronic lack of consideration for users of color becomes most apparent. These systems are designed for *public* use, and thus by default require consideration of users of color. Further, many aspects of the lighting conditions can be predicted in advance (e.g. fluorescent overhead lighting, often with the shadow of the motion-sensing object itself falling over the object to be detected). This picture only gets more complicated when we consider that the principle site of research and development for automated restroom innovation is in prisons, which are disproportionately comprised of people of color in the United States (Edwards 2015).

This is a much more significant problem than simply having automated public utilities that fail to operate for users of color. These same problems extend to medical technologies and limit the effectiveness of new wearable technologies touted as breakthrough technologies for everyday medical monitoring. For instance, Pulse Oximetry, which optically measures arterial hemoglobin oxygen saturation is demonstrably less effective in people with darker skin tones (Bickler/Feiner/Severinghaus 2005). While the FDA requires these devices to meet certain accuracy thresholds before they can go to market, they do not specify where on the Fitzpatrick scale the test subjects must fall. There is thus a financial incentive to use lighter-skinned test subjects, as utilizing and designing for darker skin types slows their path to market. New non-invasive neuroimaging techniques like functional near-infrared spectroscopy (fNIRS) are being used to study and potentially treat medical issues like Alzheimer's disease, Parkinson's disease, epilepsy, traumatic brain injury, schizophrenia, mood disorders, and anxiety disorders (Irani et al. 2007). fNIRS measures brain activity through blood oxygenation and volume in the pre-frontal cortext and is similarly less effective with darker pigmentation and darker, thicker hair (Saikia/Besio/Mankodiya 2019). These same problems of black invisibility extend to the optical heart trackers installed in many contemporary wearable technologies, like FitBit and the Apple Watch (Kim 2017; Profis 2014).

Perhaps the ultimate example of the algorithmic invisibility of users of color though can be found in lidar-based state-of-the-art object detection systems like those used in autonomous vehicles or 'self-driving cars'. Autonomous vehicles have to engage in grisly cost-benefit calculations in crash scenarios to determine how to kill or injure the fewest people (e.g. Roff 2018). New research has come out demonstrating that such systems are statistically less likely to identify darker-skinned pedestrians as humans to be factored into these calculations (Wilson/Hoffman/Morgenstern 2019). The researchers found that this remained true even when you factor in time of day (i.e., lighting, the go-to excuse for any technological failure to recognize black faces and bodies) and visual occlusion. This could have been predicted, as the training data used for the system they analyzed contained nearly 3.5 times as many images of light-skinned people as dark-skinned people (ibid.: p. 1). Black invisibility is thus not merely a matter of identity politics, but instead can literally have life or death stakes for users of color in our increasingly AI-driven future.

Conclusion

The hypervisibility/invisibility dialectic has historically sheltered the worst forms of racism in the United States and abroad. It has been endemic to visual media since their inception and is currently being cemented into the AI paradigms that we increasingly believe our going to usher in the next stage of human civilization. The approach that AI design takes to this dialectic thus ought to be a central battleground for anyone working towards the democratization of AI. We cannot have egalitarian or democratic technology if we hardcode pre-existing regimes of marginalization into our AI systems. While it is beyond the scope of this chapter to fully articulate an alternative to the hypervisibility/invisibility dialectic in AI systems, I think we look to some of the repeated refrains that have still yet to be initiated in Silicon Valley as a way to at least open the space where we might imagine a better alternative. In her *WIRED* article "How to Keep Your AI from Turning into a Racist Monster," Megan Carcia (2017) offers some common-sense approaches that might be advocated for and instituted broadly to great effect. These systems ought to better empower users to analyze, debug, and flag problematic components of AI systems. This process not only crowdsources the labor, making it much more appealing to tech companies beholden to their shareholders, but also in the process educates users about the importance of this labor as well as the technological functioning of the systems they are helping to analyze. Second, Silicon Valley needs to hire more diverse computer programmers immediately and without making any more excuses. More diverse development teams are much more likely to recognize hardcoded marginalization prior to release and are sta-

tistically likely to generate greater profits. Third, empower more third-parties to engage in high-tech auditing of AI systems. Even if trade secrecy and fear of spam prevent the open sourcing of all code, certain trusted third parties more dedicated to advancing democracy and eradicating marginalization ought to be allowed access to these systems to help monitor hardcoded biases. Lastly, we ought to further develop a public discourse that demands egalitarian AI and institute an inter-company set of tools and standards to help better motivate companies and hold them accountable. While these are Carcia's ideas, they are also ours, writ large, as they somehow echo across internet discourse ad nauseum without ever really being instituted in Silicon Valley. We might need to supplement this with more traditional grass-roots organizing and activism tactics to turn this polite request into a demand. Democracy often requires revolutionaries.

References

Alcine, Jacky (2015); Retrieved from https://twitter.com/jackyalcine/status/615329515909156865 (accessed July 9, 2019).

BBC (2015): "Google Apologises for Photo App's Racist Blunder". In: BBC News. Retrieved from https://www.bbc.com/news/technology-33347866 (accessed July 9, 2019).

Bickler, Philip E./Feiner, John R./Severinghaus, John W. (2005): "Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation". In: Anesthesiology 4/102, pp. 715-719.

Blakely, Allison (2001): Blacks in the Dutch World: The Evolution of Racial Imagery in a Modern Society, Bloomington: Indiana University Press.

Blakely, Allison (1980): "Santa's Black Aide: A Glimpse of Race Relations in Holland". In: New Directions 7/2, pp. 27–29.

Bonilla-Silva, Eduardo (2017): Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America (Fifth Edition), Lanham: Rowman & Littlefield.

Brewster, Signe (2015): "How Google's New Photos App Can Tell Cats From Dogs". In: Backchannel. Retrieved from https://medium.com/backchannel/how-google-s-new-photos-app-can-tell-cats-from-dogs-ffd651dfcd80 (accessed July 9, 2019).

Broomberg, Adam/Chanarin, Oliver (n.d.): "To Photograph a Dark Horse in Low Light". Retrieved from http://www.broombergchanarin.com/to-photograph-a-dark-horse-in-low-light-1-1/ (accessed July 9, 2019).

Brown, Michael K./Carnoy, Martin/Currie, Elliott/Duster, Troy/Oppenheimer, David B./Schultz, Marjorie M./Wellman, David (2003): Whitewashing Race: The Myth of a Color-Blind Society, Berkeley and Los Angeles: University of California Press.

Browne, Simone (2015): Dark Matters: On the Surveillance of Blackness, Durham: Duke University Press.

Consumer Reports (2010): "Consumer Reports Debunks the 'Racist' Kinect". In: Consumer Reports. Retrieved from https://www.consumerreports.org/cro/news/2010/11/consumer-reports-debunks-the-racist-kinect/index.htm (July 9, 2019).

Consumer Reports (2009): "Are HP Webcams Really Racist? Consumer Reports Weighs In". Retrieved from https://youtu.be/NphmOVolrBg (July 9, 2019).

Deng, Jia/Dong, Wei/Socher, Richard/Li, Li-Jia/Li, Kai/Li, Fei-Fei (2009): "ImageNet: A Large-Scale Hierarchical Image Database". In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition 1, pp. 248-255.

Desmond-Harris, Jenée (2014): "Don't Get What's Wrong With Blackface? Here's Why It's So Offensive". In: Vox. Retrieved from https://www.vox.com/2014/10/29/7089591/why-is-blackface-offensive-halloween-costume (July 9, 2019).

Dyer, Richard (1997): White: Essays on Race and Culture, New York: Routledge.

Dyson, Michael Eric (2018): White Fragility: Why It's So Hard for White People to Talk About Racism, Boston: Beacon Press.

Edwards, Philip (2015): "The Truth About Bathroom Sensors (And Why They Fail So Often)". In: Vox. Retrieved from https://www.vox.com/2015/9/3/9255805/bathroom-motion-sensors (accessed July 9, 2019).

Ellison, Ralph (1989): Invisible Man, New York: Vintage.

Fanon, Frantz (1967): Black Skin, White Masks (Charles Lam Markmann, trans.), New York: Grove Press.

Fair, Freda L. (2017): "Surveilling Social Difference: Black Women's 'Alley Work' in Industrializing Minneapolis". In: Surveillance & Society 15/5, 655-675.

Fitzpatrick, Thomas B. (1975): "Soleil et Peau". In Journal de Médecine Esthétique 2, 33-34.

Fitzpatrick, Thomas B. (1988): "The Validity and Practicality of Sun-Reactive Skin Types I through VI". In Archives of Dermatology 124/6, 869-871.

Fussell, Sidney (2017): "Why Can't This Soap Dispenser Identify Skin?". In: Gizmodo. Retrieved from https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773 (accessed July 9, 2019).

Google (2016): "Google Cloud and Autodesk Enable 10x Improvement in Media Rendering Efficiency". In: Google Cloud Platform Blog. Retrieved from https://cloudplatform.googleblog.com/2016/04/Google-Cloud-and-Autodesk-enable-10x-improvement-in-media-rendering-efficiency.html (accessed July 9, 2019).

Google (2015): "Picture This: A Fresh Approach to Photos". In: Google Blog. Retrieved from https://googleblog.blogspot.com/2015/05/picture-this-fresh-approach-to-photos.html (accessed July 9, 2019).

Hall, Stuart (1996): "The After-Life of Frantz Fanon: Why Fanon? Why Now? Why Black Skin, White Masks?". In: Read, Alan (ed.), The Fact of Blackness: Frantz

Fanon and Visual Representation, London: Institute of Contemporary Arts, pp. 12-37.

Hern, Alex (2015): "Flickr Faces Complaints Over 'Offensive' Auto-Tagging for Photos". In: The Guardian. Retrieved from https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos (accessed July 9, 2019).

hooks, bell (1992): Black Looks: Race and Representation, Boston: South End Press.

Ionescu, Daniel (2010): "Is Microsoft's Kinect Racist?". In: PCWorld. Retrieved from https://www.pcworld.com/article/209708/Is_Microsoft_Kinect_Racist.html (accessed July 9, 2019).

Irani, Farzin/Platek, Steven M./Bunce, Scott/Ruocco, Anthony C./Chute, Douglas (2007): "Functional Near Infrared Spectroscopy (fNIRS): An Emerging Neuroimaging Technology with Important Applications for the Study of Brain Disorders". In: The Clinical Neuropsychologist 21/1, pp. 9-37.

Kim, Meeri (2017): "Wearables Still Haven't Solved the Problems of Skin Science, But New Ideas are Coming". In: Wearable News. Retrieved from https://www.wareable.com/health-and-wellbeing/skin-science-complex-wearables-4441 (accessed July 9, 2019).

Krizhevsky, Alex/Sutskever, Ilya/Hinton, Geoffrey E. (2012): "ImageNet Classification with Deep Convolutional Neural Networks". In: Advances in Neural Information Processing Systems 25, pp. 1097-1105.

Kuznetsova, Alina/Rom, Hassan/Alldrin, Neil/Uijlings, Jasper/Krasin, Ivan/Pont-Tuset, Jordi/Kamali, Shahab/Popov, Stefan/Malloci, Matteo/Duerig, Tom/Ferrari, Vittorio (2018): "The Open Images Dataset V4: Unified Image Classificaiton, Object Detection, and Visual Relationship Detection at Scale". In: arXiv. Retrieved from https://arxiv.org/abs/1811.00982 (accessed July 9, 2019).

Levy, Steven (2015): "Bradley Horowitz Says That Google Photos is Gmail for Your Images. And That Google Plus Is Not Dead". In: Backchannel. Retrieved from https://medium.com/backchannel/bradley-horowitz-says-that-google-photos-is-gmail-for-your-images-and-that-google-plus-is-not-dead-54be1d641526 (accessed July 9, 2019).

Lyon, David (2008): "Biometrics, identification and surveillance". In: Bioethics 22/9, pp. 499-508.

Monea, Alexander (2019): "From Aristotle to Computational Topoi". In: Sundvall, Scott (ed.), Rhetorical Speculations: The Future of Rhetoric, Writing, and Technology, Logan: Utah State University Press, pp. 203-225.

Monea, Alexander (2016): The Graphing of Difference: Numerical Mediation and the Case of Google's Knowledge Graph". In: Cultural Studies ↔ Critical Methodologies 16/5, pp. 452-461.

Morgan, Eric (2006): "The World is Watching: Polaroid and South Africa". In: Enterprise & Society 7/3, pp. 520-549.

Nakamura, Lisa (2007): Digitizing Race: Visual Cultures of the Internet, Minneapolis: University of Minnesota Press.

Omi, Michael/Winant, Howard (2015): Racial Formation in the United States (Third Edition), New York: Routledge.

O'toole, Sean (2014): "Making, Refusing, Remaking: Adam Broomberg and Oliver Chanarin's Recent Photography." In: Safundi 15/ 2-3, pp. 1-14.

PhotoQ (2015): "Broomberg & Chanarin: Low Light". Retrieved from https://vimeo.com/123396189 (accessed July 9, 2019).

Plenke, Max (2015): "The Reason This 'Racist Soap Dispenser' Doesn't Work on Black Skin". In: Mic. Retrieved from https://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin#.1hOLrb9JR (accessed July 9, 2019).

Profis, Sharon (2014): "Do Wristband Heart Trackers Actually Work? A Checkup". In: CNET. Retrieved from https://www.cnet.com/news/how-accurate-are-wristband-heart-rate-monitors/ (accessed July 9, 2019).

Roth, Lorna (2009): "Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity." In: Canadian Journal of Communication 34/1, pp. 111-136.

Russakovsky, Olga/Deng, Jia/Su, Hao/Karuse, Jonathan/Satheesh, Sanjeev/Ma, Sean/ Huang, Zhigeng/Karpathy, Andrej/Khosla, Aditya/Bernstein, Michael/ Berg, Alexander C./Fei-Fei, Li (2015): "ImageNet Large Scale Visual Recognition Challenge". In: International Journal of Computer Vision 115/3, pp. 211-252.

Savage, Michael (1986): "The Imposition of Pass Laws on the African Population in South Africa, 1916-1984". In: African Affairs 85/339, pp. 181-205.

Sedaris, David (2001): "Don't They Know It's Christmas After All". In: This American Life 201. Retrieved from https://www.thisamericanlife.org/201/them (accessed July 9, 2019).

Sedaris, David (2004): "Six to Eight Black Men". In: Dress Your Family in Corduroy and Denim, New York: Little Brown and Co., pp. 157-165.

Simonite, Tom (2018): "When it Comes to Gorillas, Google Photos Remains Blind". In: WIRED. Retrieved from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ (accessed July 9, 2019).

Smith, David (2013): "'Racism' of Early Colour Photography Explored in Art Exhibition." In: The Guardian. Retrieved from https://www.theguardian.com/artanddesign/2013/jan/25/racism-colour-photography-exhibition (accessed July 9, 2019).

Spillers, Hortense (2003): "'Mama's Baby, Papa's Maybe': An American Grammar Book". In: Black, White, and in Color: Essays on American Literature and Culture, Chicago: University of Chicago Press, pp. 203-229.

Splinter, Hans (2010): "Sinterklaas 2010". In: Flickr. Retrieved from https://www.flickr.com/photos/archeon/5214550043/ (accessed July 9, 2019).

Szegedy, Christian/Liu, Wei/Jia, Yangqing/Sermanet, Pierre/Reed, Scott/Anguelov, Dragomir/Erhan, Dumitru/Vanhoucke, Vincent/Rabinovich, Andrew (2015): "Going Deeper with Convolutions." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9.

U.S. Patent No. 5,428,402 (1995) Retrieved from https://patents.google.com/patent/US5428402A/en (accessed July 9, 2019).

Weheliye, Alexander (2014): Habeas Viscus: Racializing Assemblages, Biopolotics, and Black Feminist Theories of the Human, Durham: Duke University Press.

Wevers, Rosa (2018): "Unmasking Biometrics' Biases: Facing Gender, Race, Class and Ability in Biometric Data". In: Journal for Media History 21/2, pp. 89-105.

Wilson, Benjamin/Hoffman, Judy/Morgenstern, Jamie (2019): "Predictive Inequity in Object Detection". In: arXiv. Retrieved from https://arxiv.org/abs/1902.11097 (accessed July 9, 2019).

Woodard, C. Vann (2001): The Strange Career of Jim Crow, New York: Oxford University Press.

Wynter, Sylvia (2001): "Towards the Sociogenic Principle: Fanon, Identity, the Puzzle of Conscious Experience, and What It Is Like to Be 'Black'". In: Durán-Cogan, Mercedes F./Gómez-Moriana, Antonio (eds.), National Identities and Sociopolitical Changes in Latin America, New York: Routledge, pp. 30-66.

Wzamen (2012): "HP computers are racist". Retrieved from https://youtube/t4DT3tQqgRM (accessed July 9, 2019).

Zhao, Jieyu/Wang, Tianlu/Yatskar, Mark/Ordóñez, Vincente/Chang, Kai-Wei (2017): "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints". In: arXiv. Retrieved from https://arxiv.org/abs/1707.09457 (accessed July 9, 2019).

Zunger, Yonatan (2015); Retrieved from https://twitter.com/yonatanzunger/status/615585375487045632 (accessed July 9, 2019).

Mapping the Democratization of AI on GitHub
A First Approach

Marcus Burkhardt

Over the course of the past 80 years the digital computer has radically changed the world we inhabit and the ways in which we relate to it and to each other. Likewise, computing has radically changed as well. At first, practical computing machines[1] carried individual names, but in the early 1950s proper names were quickly replaced by series designators which are emblematic for the era of mainframe computers and time-sharing systems. The 1970s and 1980s gave rise to micro, home and personal computers as well as graphical user interfaces that became prevalent in the 1990s. With the rise of the World Wide Web (WWW) during this decade networked computing and networking changed the face of computing again. Regardless of the burst of the dot-com bubble the Web flourished throughout the early 2000s by being reframed as Web 2.0 and social web. During this period computing devices became increasingly mobile and desktop computers were superseded by notebooks, smartphones and tablets, software gradually morphed into services and apps that rely on cloud infrastructures for distributed processing and storage.

In June 2017 Sundar Pichai, CEO of Google Inc., declared yet another paradigm shift in the history of computing. Innovation should neither be driven by approaching problems as first and foremost digital nor mobile, but instead by taking an AI first approach that is fueled by recent advances in the field of machine learning: "We believe smartphones should be smarter; they should learn from you and they should adapt to you. Technologies such as on-device machine learning can learn your usage patterns and automatically anticipate your next action saving you time" (Pichai 2018). This statement reflects a central promise of machine learning applications, namely the ability to adapt to unforeseen futures without prior programming of a particular event: visual recognition of specific objects or persons that the program did neither "see" nor was trained on before, self-driving cars that can deal with new situations safely or chatbots that conduct conversa-

1 For the concept of *practical computing machines* see Turing (1992).

tions with humans in an engaging manner. Conversely, the more such technologies are built into the fabric of everyday life the more concerns are raised about their potential risks, e.g. biases and inequalities inherent in training data sets. As a result, machine learning models often produce (social) structures instead of adapting to them. Drawing on debates in critical algorithm studies this paper asks how machine learning and artificial intelligence as fields of technological development and innovation are in themselves structured. By providing an initial mapping of the *coding cultures* of machine learning and artificial intelligence on GitHub the paper argues for the importance to attend more closely to the hitherto largely neglected infrastructural layers of code libraries and programming frameworks for the development of critical perspectives on the social and cultural implications of machine learning technologies to come.

Democratization of AI

In recent years the interest in artificial intelligence (AI) in general and in machine learning (ML) in particular skyrocketed once again. This ongoing development is to some extend driven by leading technology companies such as Google and its parent company Alphabet, Amazon Web Services (AWS), Facebook, IBM, Microsoft etc. It rests upon the massive accumulation of data by these companies on the one hand and the establishment of large-scale cloud infrastructures as well as infrastructural services on the other hand. However, these companies do not simply contribute to the rapid technological developments in AI and ML, they also take part in shaping the imaginaries of smart, intelligent and autonomous technologies as cornerstones of technological progress and enablers of social progress as well as economic prosperity for the years to come.

Central to this is the recent push towards the democratization of artificial intelligence. Google (IANS 2017), IBM (Moore 2018), Apple (Simonite 2017) and Microsoft (n.d.) alike mobilize the notion of democratic AI to promote the shift towards ML driven technological innovation. In this context democratization can be understood "as the action/development of making something accessible to everyone, to the 'common masses'" (Schmarzo 2017). For Microsoft this entails allowing "every person and every organization" (n.d.) to partake in the anticipated benefits of AI whose effects will supposedly be as far-reaching as that of the printing press:

> With the advent of the printing press in the 1400s we have an explosion of information—the first democratizing event around access that made it possible for humans everywhere to start learning. Access to information has only spread from there. […] The question is, how can we use all we have in terms of computational

> power to solve this fundamental constraint? To make better sense of the world? That's the essence of what AI is. It's not about having AI that beats humans in games, it's about helping everyone achieve more — humans and machines working together to make the world a better place. (Ibid.)

In view of the long history of exaggerated expectations of artificial intelligence, a certain skepticism regarding such a claim of a revolutionary caesura is warranted. What is more important, nevertheless, is the question how AI is actually made accessible, i.e. how democratization is enacted. Once again, the case of Microsoft can serve as a paradigmatic example. The company pursues a four-fold strategy: (1) utilize AI to develop new modes of interaction with ambient computing technologies; (2) build intelligence into every application; (3) allow developers to make use of these "intelligent capabilities"; (4) make computing infrastructure available as a service (ibid.).

The promise of democratization is directed towards both technology developers and their users. For developers this democratization entails the possibility to make use of AI in their own products and to partake in shaping the future of AI by having open or paid access to resources and services such as software libraries, pre-trained machine learning models, frameworks, platforms and infrastructures. Users on the other hand are enlisted in the democratization of AI as beneficiaries of technologies that are "infused" (ibid.) with artificial intelligence and machine learning. In such technologies, intelligence is typically enacted as a wide range of limited scope capabilities and features: software applications that play chess or the game of go, cameras that recognize faces and take pictures when people are smiling, speakers that are capable to recognize, interpret and execute voice-based commands, cars that drive autonomously, chatbots that engage in entertaining, helpful or informative conversations with human beings etc.

Practices of Machine Learning

The capacities of ML technologies are designed as well as staged to astonish its users. Among many researchers in the fields of science and technology studies and media studies Kate Crawford and Ryan Calo have argued for the "need to assess the impact of technologies on their social, cultural and political settings" (2016: 311). It is, thus, important to gain a critical understanding of machine learning technologies in general and the current drive towards democratic AI in particular. Such a critical understanding is all the more significant since modern day deep neural networks as well as other ML approaches supposedly evade human comprehension in principle.

Despite the secrecy imposed on algorithmic systems by corporations and the intrinsic opacity of machine learning systems (cp. Burrell 2016), there is much to know about ML technologies as Adrian Mackenzie eloquently argued:

> Machine learning is hardly obscure or arcane knowledge today. These techniques are heavily documented in textbooks […], in how-to books […], and numerous video and website tutorials, lectures and demonstrations […]. We can more or less read about and indeed play about with implementations in software […]. (2015: 431p.)

Drawing on such diverse resources Mackenzie himself became a critical student, practitioner and investigator of multiple situated, hybrid machine learning practices. In *Machine Learners* Mackenzie presents a hands-on inquiry of "machine learning as a form of knowledge production and a strategy of power" (ibid.: 9). Following Foucault's notion of archaeology Mackenzie unfolds an archaeology of six operational formations that are central to ML:

> vectorization, optimization, probabilization, pattern recognition, regularization, and propagation. These generic operations intersect in a diagram of machine learning spanning hardware and software architectures, organizations of data and datasets, practices of designing and testing models, intersections between scientific and engineering disciplines, and professional and popular pedagogies. (Ibid.: 18)

The approach to inquire machine learning not from afar, but by "learning to machine learn" (ibid.: 18) resonates well with Wendy Chuns claim "that software can only be understood *in media res*" (Chun 2008: 323). In the middle of things that constitute ML today Mackenzie engages not only with textbooks, tutorials, mathematical formulae, algorithms, and data sets, but also with numerous software libraries. While code libraries and programming frameworks are often referenced as crucial infrastructural elements of today's software culture they are hardly studied closely in critical research (cp. Berry 2011; Marino 2014). *Machine Learners*, too, acknowledges the importance of code libraries frequently, but does not research them in detail. In passing, however, Mackenzie offers some valuable insights into how code libraries and programming frameworks shape machine learning practices by crystallizing "a repertoire of standard operations, patterns, and functions for reshaping data and constructing models that classify and predict events and associations among things, people, processes, and so on" (Mackenzie 2017: 23). Their architecture "classifies and orders" (ibid.: 77) machine learning as a field of interrelated practices as much as a domain of knowledge production. They constitute the "accumulating sediment of coding and related data practices […] in which machine learners take root" (ibid.: 23). For Mackenzie the

implementation and widespread use of code libraries are, thus, articulations of contemporary *coding cultures*.

Code libraries provide resources that developers can draw from and build upon. By offering predefined functions and functionalities they relieve developers from building software from the ground up. At the same time code libraries impose their functional logics and practical affordances on developers. In this regard code libraries can be considered as media of "co-operative action" consisting of accumulative resources that can be *laminated* into operational software (Goodwin 2018: 129). However, code libraries are sites of cooperation as well. They are created, contributed to, maintained, updated and deprecated in the "recursive publics" (Kelty 2008: 28) of code-sharing platforms such as GitHub.[2]

As the "Facebook of coding" (Wulf 2017) GitHub today hosts more than 96 million repositories (GitHub 2019), most of them contain codes or coding related resources. For today's coding culture GitHub serves as a center of gravity for hosting open and closed source software projects as well as for finding, contributing and debating code resources including libraries and frameworks (Mackenzie 2018). This applies in particular to current developments in ML. Yet, how exactly does ML as a practice and field take shape on GitHub? How do algorithmic techniques for ML, data sets, machine learned models and other resources circulate on GitHub? Which actors are involved in shaping this space of cooperation, exchange and negotiation and which strategies of power are deployed? Or to put it differently: How is the democratization of AI enacted on the infrastructural level of code and coding sharing practices?

Mapping the Democratization of AI in Code

For mapping the numerous heterogenous articulations and manifestations of machine learning and artificial intelligence on GitHub researchers can make use of the application programming interface (API) provided by the platform which allows for the retrieval of repository metadata that match certain search criteria, like specific keywords contained in the repositories description or topics assigned to the repository by its creator.[3] However, the GitHub search API poses certain restrictions: it allows only for a limited number of keywords per query and returns only a maximum of 1.000 results for each query. While this might be sufficient to explore the most visible projects on GitHub for a specific keyword in regard to

2 For a discussion of the relevance of version control systems and GitHub for contemporary coding cultures see Burkhardt (2019) and Mackenzie (2017, 2018).

3 A similar mapping of cultures of coding and sharing has been undertaken by Kollanyi (2016) in respect to Bots.

the number of forks or stars a repository received, this limitation disregards the variety of keywords associated with ML and AI as well as the long tail of software development and code sharing practices in the field of ML. In order to map the enactments and materializations of machine learning more comprehensively a search tactic must be deployed that charts the space of ML and AI related repositories iteratively by identifying and working through an extensive list of search terms and by restricting the search parameters (programming languages, numbers of forks, received stars and size) step by step:

→ search terms
→ search terms + programming language
→ search terms + programming language + fork count
→ search terms + programming language + fork count + star count
→ search terms + programming language + fork count + star count + size

This search tactic provides a snapshot view of a dynamically changing environment. The results remain incomplete, but are more comprehensive than the default limit of 1.000 results per query.[4] What is left out and remains invisible in principle are all the private repositories hosted on GitHub.

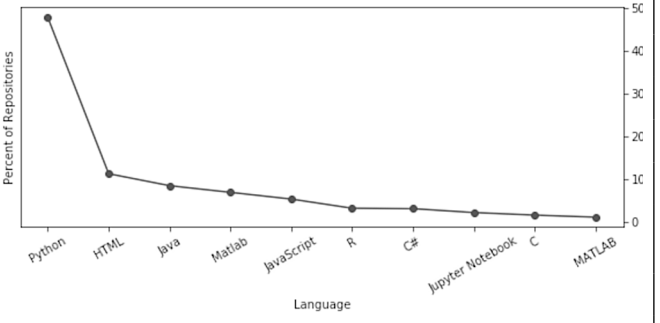
Assembling a dataset is all but the first step in mapping the enactments of machine learning in contemporary coding culture. The dataset contains 211.802 unique repositories.[5] Among those 41.818 contain artificial intelligence[6] as a keyword and 103.344 machine learning[7]. Remarkably only 3.064 repositories reference both conceptual fields although in public discourse the current summer of AI is largely attributed to developments in the area of machine learning.

4 Certain gaps in the result sets can be pinned down precisely: The result limit is exceeded for programming languages like Python that have neither been forked nor received a star and are relatively small (in the case of Python it is 0, 1, 2, 3 …, 11 KB). However, in other cases gaps emerge as inconsistencies between the supposed number of results returned by the GitHub API and the actual number of results retrieved.

5 Search terms: ai, dl, nn, ml, artificial-intelligence, artificialintelligence, artificial intelligence, machine learning, machinelearning, machine-learning, machine intelligence, deep learning, deep-learning, deeplearning, neural nets, neuralnets, neural-nets, neural net, neuralnet, neural-net, neural networks, neuralnetworks, neural-networks, neural network, neuralnetwork, neural-network, neural, bigdl, caffe, caffe2, cntk, coreml, deeplearning4j, keras, lasagne, mlib, mlpack, moa, mocha.jl, mxnet, neon, Paddle Paddle, pylearn, pylearn2, pytorch, scikit-learn, shogun, singa, tensorflow, tflearn, theano.

6 Variations considered: artificial intelligence, artificial-intelligence, artificialintelligence, ai.

7 Variations considered: machine learning, machine-learning, machinelearning, ml.



Language	%
Python	47.74
HTML	11.25
Java	8.46
Matlab	6.90
JavaScript	5.33
R	3.19
C#	3.09
Jupyter Notebook	2.18
C	1.59
MATLAB	1.10
Total	90,83

Fig. 1: Distribution of top 10 programming languages used in the retrieve repositories.

Machine learning and artificial intelligence materializes in a range of programming languages. *Python* (together with Python-based *Juypter Notebooks*) is by far the most often used programming language for ML. Remarkably this is followed by eleven percent HTML repositories. As will be discussed later this is because not all repositories on GitHub contain code as a resource. Many contain informational and educational resources such as tutorials, book manuscripts, research papers, course materials or collections of links.

The popularity or relevance of repositories can be inferred from the number of forks as well as stars they received. In the context of GitHub in particular and the Git version control system in general forks are copies of a repository. Forking constitutes a central practice in Git-based collaboration: "Most commonly, forks are used to either propose changes to someone else's project or to use someone else's project as a starting point for your own idea" (GitHub a). Stars on the other hand are platform specific indicators of some kind of interest of users in a repository: "You can star repositories and topics to keep track of projects you find interesting and discover related content in your news feed" (GitHub b).

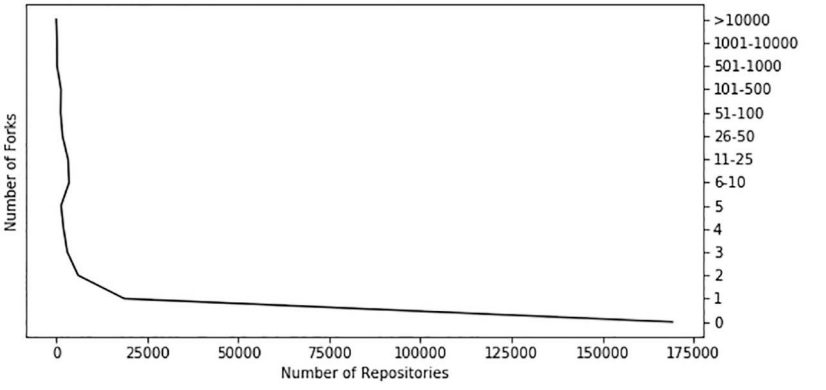


Fig. 2: Distribution of forks by repository

The distribution of stars and forks across repositories adheres to the power law. Only few repositories gain high visibility, while the largest number of repositories has no forks and received little to no stars. The long tail of machine learning is a space of testing out, experimenting with and learning by doing. This space, however, is also populated with course assignments, original research in progress, personal collections of resources on machine learning, and new, emerging, failed or abandoned code libraries. Among the more than 200.000 unique repositories that were retrieved for this article less than 0.8 percent have more than 100 or more forks and only 1.8 percent received 100 or more stars.

	Total number	≥100 forks	≥100 stars
Artificial intelligence repositories	41.818	284	593
Machine learning repositories	103.344	799	1.677
Deep learning repositories	33.749	857	1.842
Neural network repositories	46.681	558	1.365
All repositories	211.802	1.751	3.951

High level observations on the use of programming languages or the distribution of popularity offer some initial insights into how artificial intelligence and machine learning is articulated on GitHub. A more detailed analysis of the 200 most often forked and starred repositories, however, reveals the heterogeneity and diversity of resources developed, published, collaborated, maintained, debated, updated and downloaded. Among those top 200 repositories are 42 that can be categorized as code libraries or programming frameworks. However, more than

fifty percent contain informational and educational resources. About 15 percent of the repositories provide reference implementations for algorithms, machine learned models for specific application domains or software applications based on machine learning. And only few provide infrastructural resources, programming languages, experiments or datasets for ML.

Type	#
Informational/Educational resource	115
Library/Framework	42
Algorithms/Models/Applications	31
Infrastructure/Optimization	8
Experiments	2
Languages	1
Datasets	1

The informational and educational resources address a range of distinct audiences like machine learning beginners, learners of a framework, machine learning professionals as well as researchers. As a result, some repositories provide resources on the fundamentals of machine learning, materials for online courses, tutorials for different code libraries or introductory book publications. Others are framed as *collections*, *comprehensive lists* or *curated lists* on machine learning in general, state of the art research or more specific topics like infrastructures. In part these resources are created, assembled and provided by individual developers, researchers or authors. Yet, many informational and educational repositories are created and maintained by corporations to educate users *about* and recruite them *for* their products. The repository *amazon-sagemaker-examples*[8] released by *Amazon Web Services Labs* for example contains basic information and training materials on using the companies *SageMaker* platform for the training, optimization and deployment of machine learning models. Here the infrastructural complexities of machine learning practices become visible. Understood as a mode of databased programming machine learning relies on large scale processing capacities that today are provided as cloud computing services—a market dominated by global technology corporations like Amazon, Microsoft, Google, IBM etc. (Flexera 2019). What is more, engaging with the rich diversity of educational and informational resources reveals the central role of programming libraries for coding cultures. Learning to machine learn is deeply intertwined with learning to make use of code

8 https://aws.amazon.com/de/sagemaker/

libraries and frameworks. While there are repositories dedicated to do "Machine Learning From Scratch"[9], educational resources promising information on how to do machine learning with library x or framework y are more common.

The variety of repositories with informational resources is reflected in the heterogeneity of code repositories. Today, by far the most popular repository in machine learning and artificial intelligence is the TensorFlow framework.[10] Initially developed by Google Brain for internal use the framework has been released in 2015 as open source by Google and is since maintained by the company, but it has attracted a large number of external contributors as well. TensorFlow is a framework for deep learning that mainly supports the training and deployment of neural networks and is, thus, dedicated to a specific paradigm of machine learning that has become dominant throughout the past decade. Other popular libraries such as *scikit-learn*[11] focus on other areas of machine learning and in consequence support a broader range of approaches.[12] While it is possible to implement basic neural networks with scikit-learn they play a somewhat marginal role in this library. This is underlined by the missing support for deep and reinforcement learning as well as the use of GPU hardware (scikit-learn 2019, 5 and 7). TensorFlow and scikit-learn, thus, are different articulations and materializations of machine learning. Yet, they are similar in that they can both be considered as general-purpose frameworks, i.e. they are not explicitly focused on specific application domains.

A number of special-purpose machine learning libraries have gained a relatively high popularity as well. Among them are the *Unity ML-Agents Toolkit*[13], OpenCV[14], and the *ChatterBot*[15] and *Rasa*[16] frameworks. The ML-Agents Toolkit allows the implementation of so-called machine learning agents in the Unity game engine. As a plugin for a development environment for games the toolkit is of course aimed at this application area, but also allows for the development of algorithms for robotics or the training of self-driving cars in virtual environments (cp. Unity n.d.). Chatterbot and Rasa are frameworks for the implementation of AI-based chatbots. Machine learning here articulates itself as the use of pretrained models for natural language understanding and dialogue management as well as the ongoing

9 https://aws.amazon.com/eriklindernoren/ML-From-Scratch

10 https://github.com/tensorflow/tensorflow

11 https://github.com/scikit-learn/scikit-learn

12 For a comprehensive overview of machine learning paradigms and approaches see Domingos (2015).

13 https://github.com/Unity-Technologies/ml-agents

14 https://github.com/opencv/opencv

15 https://github.com/gunthercox/ChatterBot

16 https://github.com/RasaHQ/rasa

improvement of such models based on training data gathered in beta tests and during real world deployment. By providing code resources for the development of domain specific software, such frameworks are not for machine learning in general, but have some element of machine learning build in. This raises the question how special-purpose frameworks structure machine learning practices in specific ways as well as how they prescribe the operational logics and performances of applications created with them. In the case of chatbots that is to ask how *conversationality* and *communicability* is inscribed by such frameworks.

In addition to general-purpose and special-purpose libraries at least a third type can be distinguished which could be called meta-libraries. A popular example for this kind of library is *Keras*[17]. It is built on top of the deep learning frameworks TensorFlow, Theano[18] and CNTK[19] and provides a unified interface to a multiplicity of different code libraries. Keras, thus adds an abstraction layer which serves as frontend to the backend of multiple deep learning libraries. Two of those libraries are developed by global corporations—Google in the case of TensorFlow and Microsoft in the case of CNTK (Cognitive Toolkit). The third framework was mainly developed in academic contexts and is maintained by the Montreal Institute for Learning Algorithms. The end of its active development has been announced in 2017 citing among other reasons that "strong industrial players are backing different software stacks in a stimulating competition" (Bengio 2017). Indeed, large global digital technology companies compete in the space of open source machine learning libraries with respective frameworks. *PyTorch*[20] (Facebook), *Neo-AI-DLR*[21] (Amazon), *Aerosolve*[22] (AirBnB) as well as the already mentioned frameworks TensorFlow (Google) and CNTK (Microsoft) are just a few examples. The release of machine learning libraries in open source by Google, Facebook, Amazon, Microsoft etc. can be understood as an effort towards the democratization of AI. At the same time, the open source coding culture has become a space of corporate intervention and competition. How this affects the future of machine learning technologies as well as the ways in which machine learning will be built into the fabric of our technological world are questions that remain unanswered. Critical research, thus, needs to attend even more closely to the logics and politics of code libraries and programming frameworks.

17 https://github.com/keras-team/keras

18 https://github.com/Theano/Theano

19 https://github.com/microsoft/CNTK

20 https://github.com/pytorch/pytorch

21 https://github.com/neo-ai/neo-ai-dlr

22 https://github.com/airbnb/aerosolve

Literature

Bengio, Yoshua (2017): "MILA and the Future of Theano." 2017. https://groups.google.com/forum/#!topic/theano-users/7Poq8BZutbY.

Berry, David M (2011): The Philosophy of Software: Code and Mediation in the Digital Age, Basingstoke: Palgrave Macmillan.

Burkhardt, Marcus (2019): "Version Control. Zur Softwarebasierten Koordination von Ko-Laboration." In: Gießmann, Sebastian/ Röhl, Tobias/Trischler, Ronja (eds.), Materialität Der Kooperation, Wiesbaden: Springer VS, pp. 91–117.

Burrell, J. (2016): "How the Machine 'Thinks: Understanding Opacity in Machine Learning Algorithms." In: Big Data & Society 3/1. https://doi.org/10.1177/2053951715622512.

Chun, Wendy Hui Kyong (2008): "On 'Sourcery,' or Code as Fetish." In: Configurations 16/3, pp. 299-324. https://doi.org/10.1353/con.0.0064.

Crawford, Kate/ Calo, Ryan (2016): "There Is a Blind Spot in AI Research." In: Nature News 538/7625, pp. 311-13. https://doi.org/10.1038/538311a.

Domingos, Pedro (2015): The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World, New York: Basic Books.

Flexera (2019): "RightScale State of the Cloud Report 2019." Accessed June 9, 2019. https://info.flexerasoftware.com/SLO-WP-State-of-the-Cloud-2019-DE.

GitHub (2019): "The State of the Octoverse 2018." Accessed June 11, 2019. https://octoverse.github.com/.

— (a): "About Stars." Accessed June 11, 2019. https://help.github.com/en/articles/about-stars.

— (b): "Fork a Repo." Accessed June 6, 2019. https://help.github.com/en/articles/fork-a-repo.

Goodwin, Charles (2018): Co-Operative Action. Learning in Doing: Social, Cognitive and Computational Perspectives, New York: Cambridge University Press.

IANS (2017): "We Want to Democratise Artificial Intelligence: Google." YourStory.Com. Accessed June 9, 2019. https://yourstory.com/2017/08/democratise-artificial-intelligence-google/.

Kelty, Christopher M. (2008): Two Bits: The Cultural Significance of Free Software. Experimental Futures, Durham: Duke University Press.

Kollanyi, Bence (2016): "Where Do Bots Come From? An Analysis of Bot Codes Shared on GitHub." In: International Journal of Communication 10/0, pp. 4932-4951.

Mackenzie, Adrian (2015): "The Production of Prediction: What Does Machine Learning Want?" In: European Journal of Cultural Studies 18/4-5, pp. 429-45. https://doi.org/10.1177/1367549415577384.

— (2017): Machine Learners: Archaeology of a Data Practice, Cambridge: MIT Press.

— (2017): "Infrastructures in Name Only?: Identifying Effects of Depth and Scale." In: Harvey, Penny/Jensen, Casper Brunn/Morita, Atsuro (eds.), Infrastructures and Social Complexity: A Companion, London: Routledge.

— (2018): "48 Million Configurations and Counting: Platform Numbers and Their Capitalization." In: Journal of Cultural Economy 11/1, pp. 36-53. https://doi.org/10.1080/17530350.2017.1393443.

Marino, Mark C (2014): "Field Report for Critical Code Studies" In: Computational Culture: A Journal of Software Studies 4. http://computationalculture.net/field-report-for-critical-code-studies-2014%e2%80%a8/.

Microsoft (n.d.): "Democratizing AI." *Stories* (blog). Accessed September 5, 2018. https://news.microsoft.com/features/democratizing-ai/.

Moore, Todd (2018): "Think 2018: The Democratization of Artificial Intelligence." IBM Code (blog). March 20, 2018. https://developer.ibm.com/code/2018/03/20/think-2018-democratization-artificial-intelligence/.

Pichai, Sundar (2017): Google I/O'17 Keynote. https://www.youtube.com/watch?v=Y2VF8tmLFHw.

—(2018): Google I/O'18 Keynote. https://www.youtube.com/watch?v=QzbpXC0oxL0.

Schmarzo, William (2017): "Democratizing Artificial Intelligence, Deep Learning, Machine Learning with Dell EMC Ready Solutions." In: InFocus Blog | Dell EMC Services. Accessed June 7, 2019. https://infocus.dellemc.com/william_schmarzo/democratizing-artificial-intelligence-deep-learning-machine-learning-with-dell-emc-ready-solutions/.

scikit-learn (2019): "Scikit-Learn User Guide (Release 0.22.Dev0)." Accessed June 7, 2019. https://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf.

Simonite, Tom (2017): "Apple Just Joined Tech's Great Race to Democratize AI." WIRED. https://www.wired.com/2017/06/apple-siri-ai/.

Turing, Alan M. (1992): "Intelligent Machinery." In: Mechanical Intelligence, edited by Darrel C. Ince. Collected Works of A.M. Turing, Amsterdam: North-Holland, pp. 107-27.

Unity (n.d.): "Machine Learning." Unity. Accessed June 7, 2019. https://unity3d.com/machine-learning.

Wulf, Josh (2017): "9 Resources to Get Started Coding with JavaScript." Opensource.Com. Accessed June 7, 2019. https://opensource.com/article/17/6/get-started-coding-javascript.

On the Media-political Dimension of Artificial Intelligence

Deep Learning as a Black Box and OpenAI[1]

Andreas Sudmann

> 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—ANNs). First, this essay discusses whether ANN/DL has to be perceived as a fundamentally opaque or 'black box' technology, perhaps inaccessible or only partially accessible to human understanding. Second, 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 purpotedly 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 capital invested 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 ac-

1 This is a slightly revised and extended version of an essay that has been published under the same title in: Ramón Reichert, Mathias Fuchs (Ed.), *Rethinking AI. Neural Networks, Biopolitics and the New Artificial Intelligence, Digital Culture & Society* 4/1, 2018, pp.181-200.

cessibility 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?

1. 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). 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 by more than the half.[2] While this increase may not sound very impressive, 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 they were already harnessing big data, and now would have 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 optimised machine translations, etc. In general, DL algorithms are a universal instrument for pattern-recognition and prediction tasks, an 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 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

2 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.

(Mitchell 1997). This is the basic 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 natural language processing and computer vision.

Very broadly speaking, one of the 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 millions of iterations up until the very moment the system is 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 two or more intermediate layers that process the information through the network. On the lowest level of layers, the network analyses 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 (for example, parts of the object like a 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?).

2. 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 social sciences, whereas before 2016 many disciplines outside the hard sciences more or less ignored machine learning technologies like 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 discourses of philosophy of mind as well as cognitive psychology, which very early developed an interest in 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.[3] 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 does so 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, for instance in the humanities that tackle the topic of AI in general and DL technology in particular (Sudmann 2016, Pasquinelli 2017, Finn 2017, McKenzie 2017, Engemann/Sudmann 2018). 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 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 book *Superintelligence: Paths, Dangers, Strategies* (2014) attracted much public attention, partly because of its alarmist 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 humankind is able to handle an AI-driven society as long as we instantiate 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 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).

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

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).[4]

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').[5] 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.[6] For me, *every* entity involved in the process of generating and shaping AI technology can generally be perceived as a mediator (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.[7] For this very reason, it generally makes sense to differentiate between the concepts of mediator(s) and medium/media.

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

4 Accordingly, one possible approach would be to examine the politics of representation of different AI technologies in popular media like film and television.

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

6 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. The prefix "infra" highlights to look at media (operating) on a level that is generally invisible (for example on the level of code).

7 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).

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.[8] 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 or a medium in some ontologically 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 examines the *constitutive role of mediators* (and of 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 autonomous 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. To provide another example: When designing autonomous weapon systems, there is obviously a strong political interest involved in keeping hidden the mediators of the production of these technologies, as well as their functioning. On the other hand, with regard to the question of how AI affects the future of work, for example, it may be more important to make all mediators involved in this process as transparent and accessible as possible.

Such a plea does not mean that "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 to AI). This focus entails many relevant aspects, including the analysis of 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 or social 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

8 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).

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 at some point. For many scholars in the field, it is one of media studies' central tasks to focus on processes of knowledge translation 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 towards opening 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.

3. 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 the 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).[9] 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 seems to be true

9 Towards this definition, cf. W. Ross Ashby's 1956 book *An Introduction to Cybernetics*.

for DL systems. 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 in need of clarification. First, one should keep in mind that the math behind current DL technology is rather straight forward (cf. Goodfellow/Bengio/Courville 2016). Ultimately, it is a matter of statistics, albeit an advanced form of statistics. This aspect is important to highlight since we can observe a general tendency towards 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 trying 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.

4. 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 deduce 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 an '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 the prediction model. Without access to the latter, however, it is difficult for XAI models to fulfill the great promises of a democratic AI: To provide full accessibility, transparency and, control.

Another interesting tool that—in a way—helped to make visible how ANNs 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 differently than 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 those objects (for example, cats) in images 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 neural network 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.[10] 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). But does DeepDream also contribute to the democratization of AI? There are certainly

10 For a critical perspective on DeepDream from a media-theoretical and psychoanalytical perspective, cf. Apprich (2017). The following website shows some examples for the use of the DeepDream algorithm: https://deepdreamgenerator.com/#gallery

good reasons to be sceptical here, yet at least DeepDream offers an aesthetic approach to illustrate the functioning of artificial neural networks and in doing so enabling people to reflect, for instance, on the opacity of DL technologies. And DeepDream has perhaps also inspired a number of artists who have been using DL techniques in their work for some time. For example, the Parisian artist collective "Obvious" has explicitly declared that they want to democratize AI with their art. The group recently has made headlines by selling, for a considerable sum, a portrait at the auction house Christies which they had produced using a so-called GAN algorithm.[11]

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 this happens (if this is ever going to happen), DL technology needs 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. The latter would, however, be necessary in view of the claims to democratize AI, if one takes into account, for example, that laws and other forms of regulations and organization definitely require causal explanations. However, as long as DL models cannot adaeuqately address problems of causality, it is misleading or simply an exaggeration to speak of these approaches as forms of XAI.

11 At the same time Obvious was criticized for their project because they did nothing but use a foreign algorithm to generate this and other of their portraits.

5. The Politics of OpenAI

As I indicated earlier, providing models of an explainable and—more generally—a responsible AI has some obvious motivations. 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 that started as a "non-profit research company" (self-description)[12], 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.[13] This apocalyptic view has been around since the very beginning of AI research and even existed before. In fact, as media scholar Bernhard Dotzler already pointed at the end of the 1980s, you can find most well-established projections of the future of AI already in the work of Alan Turing (cf. Dotzler 1989). And yet, since very

12 It is worth noting and also telling that OpenAI is now a "capped-profit" company (cf. Coldewey 2019: online).

13 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).

recently, 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 historical point we might take the wrong path towards harmful applications of AI. 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 what could be called a basic form of common-sense knowledge of the world. In order to achieve this, the ANN has been trained, not with photos, but with hundreds of thousands of short video-clips (showing certain activities and hand gestures). Hence, the specificity of media, i.e. here the difference between still and moving images as a mediator of DL technology, is an essential factor for developing advanced forms of AI.

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 of parsing complexities that humans aren't able to cognitively process. 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. For example, 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) by 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 to 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 sceptical views on what AI systems are capable of (or not). In general, we should acknowledge that 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 will 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 democratizing 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 question, however, is: How can OpenAI legitimate its brand image and former status as a non-profit (and now cropped-profit) research company when it essentially does what all other big players in the AI game are doing—i.e. improving existing technology and/or finding the right path to develop an artificial general intelligence (AGI)? 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 are currently trying to develop highly efficient AGI systems. Instead, what we can learn by taking a closer look at those different research agendas is the simple fact that DL is not a homogeneous approach, but an umbrella term for very different approaches to AI research and design.

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 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 functions. 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 algorithms. 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 mediators and media: legal aspects, institutional procedures, environmental issues, existing political as well as legal regulations, and so on.[14]

These aspects are also important to consider if we 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 scandal has been very insightful, since it demonstrates how many of 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 or least not su much 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

14 In order to shed light on problems of opacity concernin the broad spectrum of mediators, Burrell's (2016) account of black boxing might be of use, with he breaks into three levels: "(1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully."

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 rather 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 may be more complicated than Bostrom's statement suggests. Neither is it 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: 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 over the processes of data collection, storage, processing, and transmission related to their personal lives and activities. The revelations of Edward Snowden have already 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 that 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.[15] These aspects have been the subject of cultural critique long before the current AI boom took off. We should therefore not simply discuss how

15 For example, as it has been revealed during the Cambridge Analytica scandal, the company used a back door in the Facebook API from c. 2012-2015 where a Facebook friend's choice to opt

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 protect us from the dangers of AI as well as the misuse of big data. Neither do we have any profound reason to believe that specifically companies like Open-AI, 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 that control both the data as well as the intelligent algorithms capable of making 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 political opposition or perhaps terrorists using AI for cyberattacks. This threat is not an instance of a big data or AI paranoia: As experts have recently demonstrated, by only having access to a so-called API, you are able to reverse-engineer machine learning algorithms with 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.

Acknowledgements

I like to thank Alexander Monea for providing very helpful feedback and comments on the original version of this text. I would also like to express my gratitude to the Center of Advanced Internet Studies in Bochum for the opportunity to do research on the topic of this paper as senior fellow in 2017.

in to the permissions for an app could give that app all of your personal information (including the contents of your private messages).

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.

Ashby, W. Ross (1956): *Introduction to Cybernetics*. London: Chapman & Hall.

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.

Burrell, Jenna (2016): "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society*, 3/1, pp. 1-12. https://doi.org/10.1177/2053951715622512.

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

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.

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].

Coldewey, Devin (2019): "OpenAI shifts from nonprofit to 'capped-profit' to attract capital." *Techcrunch.com*. (https://techcrunch.com/2019/03/11/openai-shifts-from-nonprofit-to-capped-profit-to-attract-capital/).

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.

— (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.

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

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/ac9e10ce-30b2-11e7-9555-23ef563ecf9a [Last access: 2018/03/06].

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

— (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].

— (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 Press.

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.

— (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/@jrod

thoughts/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].

— (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.

How to Safeguard AI

Ina Schieferdecker/Jürgen Großmann/Martin A. Schneider

1. Introduction

Artificial intelligence is a discipline within computer science that deals with the development of software-based systems that provide functions which require the execution of what is typically called (human) intelligence. However, since there is no widely accepted definition of human intelligence, there is also no widely accepted for artificial intelligence, sometimes also called machine intelligence (Legg, 2007). AI uses methods and tools from logic, probability theory, and continuous mathematics in order to provide perception, reasoning, learning, and action via software-based systems (Russell, 2016). And it provides already numerous practical applications in transportation, energy supply, health services, finance and banking as well as law and regulation: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy." (Stone, 2016)

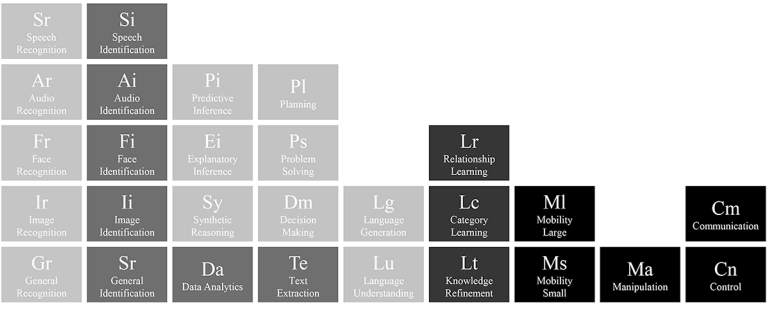


Fig. 1: Functional components in AI by Hammond (2016): Recognition of speech (Sr), audio (Ar), face (Fr) and image (Ir) and general recognition (Gr), Identification of speech (Si), audio (Ai), face (Fi) and image (Ii) and general identification (Gi); Data analytics (Da) and Text extraction (Te); Predictive inference (Pi), Planning (Pl), Explanatory inference (Ei), Problem solving (Ps), Synthetic reasoning (Sr), and Decision making

(Dm); Language generation (Lg) and understanding (Lu); Relationship learning (Rl), Category learning (Cl) and Knowledge refinement (Kr); Mobility at large (Ml) and at small (Ms); Manipulation (Ma), Communication (Cm) and Control (Cn), which can be used standalone or in combination e.g. to predict future events by recognizing sounds of technical systems and/or identifying images representing system states and/or correlating data and recognizing specific facts.

Technologies that are used to build AI by machine learning (in short ML), which is about improving problem solving accuracy or efficiency by learning to do something better, are numerous. Machine learning can e.g. be grouped along the learning type into methods for supervised, unsupervised or semi-supervised learning or along the knowledge extraction by symbolic computation or sub-symbolic processing. They can also be grouped along the principal approach, e.g. into regression, instance-based, regularization, decision tree, Bayesian, clustering, neural network, deep learning, and quite many other algorithms. Based on these, likewise numerous AI applications can be developed. Hammond (2016) presented a first taxonomy of AI functional components (Fig. 1). No matter which functional components are being used, AI-based systems are realized by use of software or also by use of sensors and actuators for the interconnection with the environment (Fig. 2). The software uses data which are interpreted by algorithms in order to provide automatisms for parts of or for entire processes in technical systems like in car engine control or in socio-technical systems like in autonomous driving.

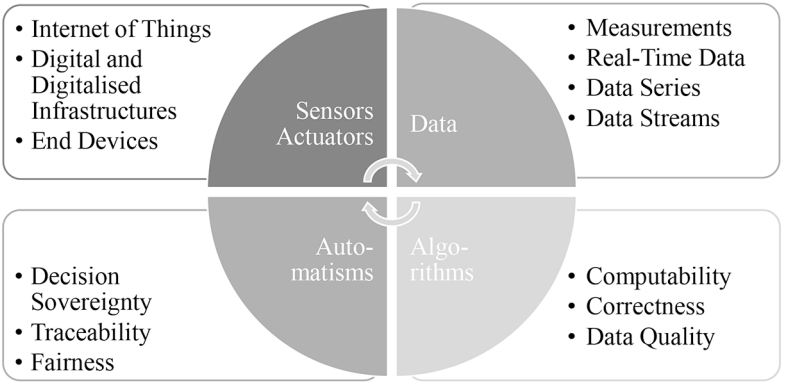


Fig. 2: Elements of software-based systems (WBGU, 2019). Sensors are part of the Internet of Things and generate different kinds of data such as measurements, series of measurements or data streams. Algorithms use these data in their computations or as training data. The algorithms are constrained by complexity, computability, and performance limits and possibly by the (in-)correctness of the implemented computation logic and by the (un-)biased (training) data. In result, software-based systems offer

automatisms for which it is essential to agree (and assure) decision sovereignty, traceability and fairness. Any decision in respect to the environment can finally be fed via software (into the cyberspace) and via actuators (into the environment).

2. Software Verification and Validation

Since any AI is also a software-based system, it is to be seen to which extent AI can be verified and validated with the established verification and validation (in short V&V) methods for software in general. V&V methods for software were revealed already with the software crisis back in 1968 (Wirth, 2008), when the term software engineering was coined. It pointed at the difficulties to design and develop useful and trustworthy software with the given resources and within the given time: "The major cause of the software crisis is that the machines have become several orders of magnitude more powerful! …(A)s long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now we have gigantic computers, programming has become an equally gigantic problem." (Dijkstra, 1972). And the newly coined term pointed at the necessity to develop practical and scalable engineering methods for software development. Since then, constructive and analytic methods for software quality engineering have been developed. They include methods for software engineering processes, software engineering tools and for software as such. A rough overview on these methods is given in Fig. 3.

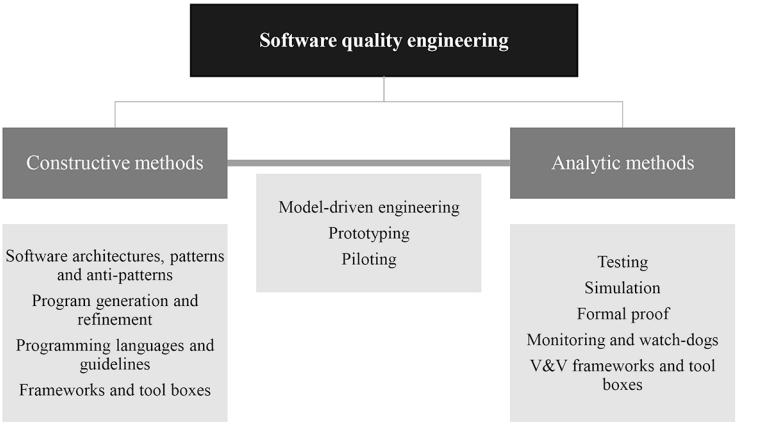


Fig. 3: Overview on software quality engineering methods. Software quality begins with the software design that is represented by software architectures which can make use of software patterns. Programs can be (partially) generated from these software designs and/or refined. The programs use typically high-level programming languages

which offer guidelines for best practice programming and which are supported by programming frameworks and tools. The achieved software quality is typically tested, checked by simulation or proven formally. The running software can be monitored and watch-dogs can check for constraint violations at run-time. All these analytic methods can also be automated by V&V frameworks and tools. Three specific (sets) of methods can be used both constructively and analytically: that is the use of model in software engineering, the early prototyping of software (or of V&V software) and the piloting of software (or of V&V solutions).

The software (program or code) tells the computer what to do, *"but that may be much different from what you had in mind".* (Joseph Weizenbaum, Computer Scientist, 1923-2008). However, by the systematic use of software quality engineering methods, software can be developed such that it is safe, secure, and trustworthy and that it can analyze and compute more data than any person and can do this more reliably.

Numerous international software engineering standards put the ground for software quality such as ISO/IEC 25010 (ISO, 2011) for software quality requirements and evaluation (SQuaRE) and software quality models. It argues about quality in use, external quality and internal quality of software and differentiates between functional suitability, reliability, usability, security, compatibility, portability, maintainability and performance/efficiency.

While these are all important software quality aspects that evolved over decades, interestingly, new aspects arise for AI in their use within socio-technical systems. Apparently,

- understandability, i.e. users and operators can get to know the features and services of the systems,
- interpretability, i.e. users and concerned people have access to clarifications of outcomes and their potential impacts,
- traceability, i.e. users and concerned people have access to more detailed analysis of outcomes in relation to a given situation/problem statement,
- explainability, i.e. users and concerned people receive descriptions, reasoning and justifications on the outcomes, as well as
- fairness, i.e. concerned people are treated the same wrt. commonly agreed rules for treatment, gain much more momentum.

3. AI Verification and Validation

Indeed, AI requires to quite some extent additional methods and tools for V&V (Van Wesel, 2017) since well-established testing technologies are short in V&V of AI. This is not only true because of the additional socio-technical quality aspects (see above), but also due to the different nature of logic-based software (most of the software in general so far and some of AI) and statistics-based software (most of AI, in particular in machine learning). Testing has limitations with respect to the dynamics of ML, the sheer size of the problem domain and the underlying oracle problem (Xie, 2011).

In addition, most of the AI is controlled by data. In this sense, a neural network is a generic function approximator whose structure reflects the actual functionality only to a very small extent. Hence, source code-oriented V&V techniques such as static analysis or white-box tests are only of limited use in this context. On the other hand, the trustworthiness and quality of the data becomes a central issue for the overall quality of the systems.

However, since systematic dynamic testing of software is the best-known and most effective V&V method, it will most probably also form the main basis for testing ML. In recent decades, research has developed industrial-grade techniques for increasing the quality, efficiency and reliability of testing. This includes in particular, automation strategies for dynamic testing such as automating test executions with test technologies like TTCN-3 (Testing and Test Control Notation [Grabowski, 2003]), for model-based testing to automate the generation of tests (MBT [Utting, 2012]), as well as the use of search and optimization algorithms for automated test selection and test suite reduction (Harman, 2015). Moreover, the combination of dynamic testing with verification approaches like source code analysis, model checking and symbolic execution allows for improvements in testing, that combines the rigor of verification processes with the scalability of dynamic testing (Godefroid, 2018). These techniques are applied to testing for functional as well as extra-functional properties like performance or security (Schieferdecker, 2012). Finally, the close integration of testing with system development processes and risk management (Felderer, 2014) improved the efficiency and transparency of testing so that testing has matured as one of the most important software quality measures in industry. Still, test automation as well as the use of models in testing are still underexplored: although a strong test automation is required, less than 14% of software testing professionals say that they use MBT (Binder, 2015). The potential of risk-based testing to steer test processes based on uncertainties has been shown especially in the area of critical system in terms of security and safety, which will likewise be applicable to AI (Erdogan, 2014).

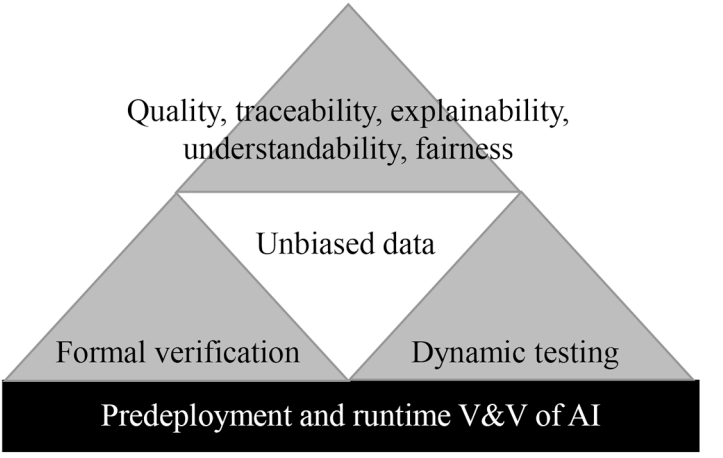


Fig. 4: The AI V&V pyramid. AI-based systems are to be verified and validated both in predeployment phases and at runtime. A combination of V&V methods from formal verification and dynamic testing is recommended, in particular for safety- and security critical AI-based systems. V&V will help to assure both quality and explainability requirements as well as enable the justification of bias in the (training) data used in AI.

Research on dedicated methods for verification and validation of ML is still at its beginning. Even so, testing is already part of the overall training set-up in ML, most testing is done to achieve more accurate models with respect to the initial training objectives. In supervised learning for example, test and validation data sets are used to provide evaluation of the ML model. Validation data sets are typically used during training to fine-tune the model parameters while test data sets are used on the final model to measure generalization errors. However, since individual test sets only provide a single evaluation of the model and have limited ability to characterize the uncertainty in the results, more advanced statistical testing approaches like cross-validation are used for model selection.

Ghosh et al (2016) combine ML and model checking in such a way that if the desired logical properties are not satisfied by a trained model, the model ('model repair') or the data from which the model is learned is modified systematically ('data repair'). Fulton and Platzer (2018) propose to combine formal verification with verified runtime monitoring so that safe learning can be guaranteed. The approach intervenes in the learning process whenever safety properties are violated and guides the learning process so that the result is compliant with the verification model. Approaches like DeepXplore (Pei, 2017), DLFuzz (Guo, 2018) and TensorFuzz (Odena, 2018) provide metrics for the quantification of neural coverage and simplify test automation. DeepTest (Tian, 2018) enables systematic testing of

neural networks under realistically changing environmental conditions especially for use in the automotive domain.

One of the socio-technical limitations of ML is the lack of transparency, i.e. its black box-approach. In order to address it, different approaches have been proposed such as

- model interpretation for image classifications, e.g. by understanding the activation maximization with saliency maps (Simonyan, 2013),
- model explanation by sensitivity analysis and local explanation vectors to provide reasons for the decisions of any classification method (Baehrens, 2010),
- model decomposition for interpreting generic multilayer neural networks by decomposing the network classification decision into contributions of its input elements (Montavon, 2017),
- extraction of decision trees from input data generated from trained neuronal networks (Krishnan, 1999),
- relevance propagation by pixel-wise decomposition of non-linear classifiers (Bach, 2015), and
- deconvolution methods to give insight into the function of intermediate feature layers and the operation of classifiers (Zeiler, 2014).

Another well-established way is to use test scenarios, i.e. test cases and their test data, for explaining ML decisions. The other socio-technical limitation of ML is the potential lack of fairness, i.e. the potential bias. Here, systematic generation of (training) data that cover well required categories and properties as known from test data generation is of help (Nguyen, 2016).

The ability to effectively test AI will be fundamental for the acceptance in broad scale and central for safety-critical areas like transportation and automotive, healthcare, or industrial automation. The provisioning of test technologies, tools, test scenarios with test cases and test data for AI will not only be a solid basis for V&V but also help in explaining AI and making them more transparent and unbiased. They can also be used to ensure safety and security of AI during runtime.

And last but not least, the tools for safeguarding AI contribute also to the democratization of AI: They are the basis for confirming or witnessing outcomes whenever AI-based systems are to be accounted. They can also become a digital common for the comparison and benchmarking of AI-based systems and by that contribute to a shared knowledge basis of AI.

Acknowledgment

This article resulted from several discussions with the research group on the criticality of AI-based systems lead by Diana Serbanescu, with the German Advisory Council on Global Change lead by Dirk Messner and Sabine Schlacke, with the participants at the conference "The Democratization of AI. Net Politics in the Era of Learning Algorithms" in Bochum, September 2018 led by Andreas Sudmann and with the members of the ITEA3 project "Industrial-grade Verification and Validation of Evolving Systems (IVVES) led by Jürgen Großmann.

References

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015): "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation", PloS one 10, no. 7, e0130140.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller (2010): "How to Explain Individual Classification Decisions", Journal of Machine Learning Research 11, no. Jun, 1803-31.

Binder, Robert V, Bruno Legeard, and Anne Kramer (2015): "Model-Based Testing: Where Does It Stand?", Communications of the ACM 58, no. 2, 52-56.

Dijkstra, Edsger W. (1972): "The Humble Programmer", Commun. ACM 15, no. 10, 859-66.

Erdogan, Gencer, Yan Li, Ragnhild Kobro Runde, Fredrik Seehusen, and Ketil Stølen (2014): "Approaches for the Combined Use of Risk Analysis and Testing: A Systematic Literature Review", International Journal on Software Tools for Technology Transfer 16, no. 5, 627-42.

Felderer, Michael, and Ina Schieferdecker (2014): "A Taxonomy of Risk-Based Testing", International Journal on Software Tools for Technology Transfer 16, no. 5, 559-68.

Fulton, Nathan, and André Platzer (2018): "Safe Reinforcement Learning Via Formal Methods: Toward Safe Control through Proof and Learning", Paper presented at the Thirty-Second AAAI Conference on Artificial Intelligence.

Ghosh, Shalini, Patrick Lincoln, Ashish Tiwari, Xiaojin Zhu, and Wisc Edu (2016): "Trusted Machine Learning for Probabilistic Models", Paper presented at the ICML Workshop on Reliable Machine Learning in the Wild.

Godefroid, Patrice, and Koushik Sen (2018): "Combining Model Checking and Testing", In Handbook of Model Checking, 613-49: Springer.

Grabowski, Jens, Dieter Hogrefe, György Réthy, Ina Schieferdecker, Anthony Wiles, and Colin Willcock (2003): "An Introduction to the Testing and Test Control Notation (Ttcn-3)", Computer Networks 42, no. 3, 375-403.

Guo, Jianmin, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun (2018): "Dlfuzz: Differential Fuzzing Testing of Deep Learning Systems", Paper presented at the Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering.

Hammond, Kris (2016): "The Periodic Table of AI", https://www.datasciencecentral.com/profiles/blogs/the-periodic-table-of-ai. Accessed 11 July 2019.

Harman, Mark, Yue Jia, and Yuanyuan Zhang (2015): "Achievements, Open Problems and Challenges for Search Based Software Testing", Paper presented at the IEEE 8th International Conference on Software Testing, Verification and Validation (ICST).

ISO/IEC 25010 (2011): Systems and Software Engineering-Systems and Software Quality Requirements and Evaluation (Square)-System and Software Quality Models. International Organization for Standardization.

Krishnan, R, G Sivakumar, and P Bhattacharya (1999): "Extracting Decision Trees from Trained Neural Networks", Pattern recognition 32, no. 12.

Legg, Shane, and Marcus Hutter (2007): "Universal Intelligence: A Definition of Machine Intelligence", Minds and machines 17, no. 4, 391-444.

Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller (2017): "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition", Pattern Recognition 65, 211-22.

Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune (2016): "Synthesizing the Preferred Inputs for Neurons in Neural Networks Via Deep Generator Networks", Paper presented at the Advances in Neural Information Processing Systems, 3387-3395.

Odena, Augustus, and Ian Goodfellow (2018): "TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing", arXiv preprint arXiv:1807.10875.

Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana (2017): "DeepXplore: Automated Whitebox Testing of Deep Learning Systems", Paper presented at the proceedings of the 26th Symposium on Operating Systems Principles, 1-18, ACM.

Russell, Stuart J, and Peter Norvig (2016): Artificial Intelligence: A Modern Approach. Malaysia; Pearson Education Limited.

Schieferdecker, Ina, Juergen Grossmann, and Martin Schneider (2012): "Model-Based Security Testing", arXiv preprint arXiv:1202.6118.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013): "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", arXiv preprint arXiv:1312.6034.

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller (2016): "Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence. Report of the 2015 Study Panel", 52. Stanford, CA: Stanford University.

Tian, Yuchi, Kexin Pei, Suman Jana, and Baishakhi Ray (2018): "DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars", Paper presented at the Proceedings of the 40th International Conference on Software Engineering, 303-314. ACM.

Utting, Mark, Alexander Pretschner, and Bruno Legeard. "A Taxonomy of Model-Based Testing Approaches", Software Testing, Verification and Reliability 22, no. 5 (2012): 297-312.

Van Wesel, Perry, and Alwyn E Goodloe (2017): "Challenges in the Verification of Reinforcement Learning Algorithms", NASA/TM–2017-219628.

WBGU (2019): Towards Our Common Digital Future. Flagship Report. Berlin: German Advisory Council on Global Change.

Wirth, Niklaus. "A Brief History of Software Engineering", IEEE Annals of the History of Computing 30, no. 3 (2008): 32-39.

Xie, Xiaoyuan, Joshua WK Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen (2011): "Testing and Validating Machine Learning Classifiers by Metamorphic Testing", Journal of Systems and Software 84, no. 4, 544-58.

Zeiler, Matthew D, and Rob Fergus (2014): "Visualizing and Understanding Convolutional Networks", Paper presented at the European conference on Computer Vision, 818-833, Springer.

AI, Democracy and the Law

Christian Djeffal

Digital technologies are in the process of reconfiguring our democracy. While we look for orientation and guidance in this process, the relationship between technology and democracy is unclear and seems to be in flux. Are technology and democracy mirroring each other?[1] The internet was first hailed as genuinely democratic technology and ultimate enabler of democracy. It is now often perceived as a major threat to democracy. The story of artificial intelligence (AI) might turn out to be quite the opposite. While there are many reflections on AI as a threat to or even as the end of democracy,[2] some voices highlight the democratic potentials of AI.[3] As is often the case, the research results depend on the premises underlying the research. This chapter is based on the assertion that technologies and media shape human affairs to a large extent, but that technology in turn is also shaped by human choices and decisions. There is a huge potential to endanger, game or even abolish democratic processes. On the contrary, there might also be opportunities to further democracy. Therefore, the extent to which AI impacts democracy is subject to the paths that are chosen in research, development and application of AI in society.

The main purpose of this chapter is to highlight the room for choice in the construction of AI and its impacts on the future of democracy. It will also inquire into how law and jurisprudence relate to these questions. From this perspective, current impacts of AI on democracy have an important indicative function. But in the face of further possibilities of inventions and regulative measures on different levels, they are only precursors to what will and should be possible. In that sense, this chapter is also an attempt to deal with developments and inventions we cannot yet grasp. The main argument is that it might be possible to influence them nevertheless. Therefore, the chapter will reflect on the possibility and necessity to democratize AI from a legal and jurisprudential perspective. It will then look at different ways to democratize AI.

1 On this question see Hofmann (2018).

2 Hofstetter (2016); O'Neil (2016).

3 Helbing (2019); Ennals (1987).

I. Democratizing AI: Possibility and Necessity

A. Understanding the Openness and the Power of Artificial Intelligence

In order to understand the relationship between artificial intelligence and democracy, it is necessary to clarify the concept of AI. The concept's crucial aspect lies not in a clear-cut definition of AI but in its openness. AI is a very broad concept indeed, and this might be the reason why this concept has outperformed other ideas, such as cybernetics, and is today the general term used in science, politics and the economy. Artificial intelligence is a term denoting a research question that inspires a hole sub-discipline of computer science today. This research question has been summarized as follows: Can systems solve complex problems independently?[4] The openness can already be seen in the initial definition of the term from 1955 in a grant proposal to the Rockefeller Foundation:

> We propose that a 2-month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.[5]

It is clever to frame a grant proposal in a way that encourages the imagination of those reading it. The way AI was used here does exactly that. The first aspect regarding the openness of AI that can be derived from this definition is that AI is a research question. It is not a theory offering explanations. It is not a general hypothesis or an idea framing certain aspects in a particular manner. The general research question of whether systems can solve complex problems independently is based on certain conjectures, but those are reduced to a minimum. The fact that AI is a question might also explain the longevity of the term. AI has seen so many ups and downs that commentators speak about "AI winters" and "AI summers."[6] As long as the general research question underlying AI is not solved in a manner that cannot be improved, it will continue to be interesting for AI researchers. Another aspect of the openness of AI relates to its basic assumptions. Comments by John McCarthy, one of the grant applicants and important figures in AI research, suggest that the term AI was coined to avoid the assumptions made in cybernet-

4 Mainzer (2019: 3).

5 McCarthy/Minsky/Shannon (1955).

6 Sudmann (57).

ics research and to get around the influence of Norbert Wiener.[7] While Wiener certainly made great contributions to the field of computer science and touched upon many important questions of AI that are still relevant today, he did so from another perspective. His idea of cybernetics, commonly held among many other important colleagues, is a general theory with strong assumptions. In contrast, the term AI has traditionally accommodated quite different views. One general disagreement has been termed as the strong and weak AI hypothesis.[8] The strong AI thesis departed from the idea that AI can either replicate or even surpass human intelligence. In contrast, the weak AI thesis only requires machines to act as if they were intelligent. It focusses generally on certain specific problems to be solved.

Another aspect of the openness of AI is that it does not relate to a single technology but to a whole set of technologies.[9] At the moment, technologies of machine learning[10] are considered to be either state of the art or even "real AI." Artificial neuronal networks, for example, fulfil certain tasks such as image recognition. They are trained on the basis of a great amount of training data, which is labelled so that the mathematical models underpinning the learning may continuously be adapted and improved. In contrast, generative adversarial networks improve themselves in an adversarial manner without the input of human training data. There are still many general ideas and architectures that might have been more popular in the past, such as decision trees, or that might become more popular in the future, such as evolutionary AI. Since AI is open for new approaches and breakthroughs, AI research continues to be a moving target. Systems that represented state-of-the-art-AI at one point in time do not qualify as being truly intelligent later. Different technologies require different resources. While AI is sometimes associated with big data applications that rely on training or analysis of huge amounts of data, big data is not a necessary requirement. There are also small data applications or applications that do not require significant training data at all. The resources vary accordingly. Artificial neural networks need large amounts of training data, sufficient memory space to store this data and enough power to compute it. It is especially important to note that the training data has to be annotated by human beings. Whether it is the interpretation of x-rays, skin cancer detection or crosswalk recognition in the context of automated driving, the data to train deep neural networks is dependent on human input. Large pools of human resources were even more crucial with the old expert systems popular

7 McCarthy (1989).

8 For a discussion see Russell/Norvig/Kirchner (2012: 1020).

9 Gasser/Almeida (2017: 59).

10 For overviews from different perspectives see Shalev-Shwartz/Ben-David (2014); Sudmann/Engemann Goodfellow/Bengio/Courville (2016).

in the 1990s. Experts had to design decision trees in many cases, which would then assist other people.

Furthermore, the general purposes of AI are also open. While it is often assumed that AI is synonymous with automation, there is indeed a disagreement about whether the goal of AI is rather augmentation than automation. While automation relates to the replacement of humans by machines, augmentation focusses on human-machine interaction in order to amplify human capabilities. This augmentation paradigm proved to be influential in different areas of computer science. Even the earliest research agenda by its most influential proponent Douglas Engelbart shows that there is a clear link to the research agenda of artificial intelligence:

> By "augmenting human intellect" we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems. Increased capability in this respect is taken to mean a mixture of the following: more-rapid comprehension, better comprehension, the possibility of gaining a useful degree of comprehension in a situation that previously was too complex, speedier solutions, better solutions, and the possibility of finding solutions to problems that before seemed insoluble. And by "complex situations" we include the professional problems of diplomats, executives, social scientists, life scientists, physical scientists, attorneys, designers—whether the problem situation exists for twenty minutes or twenty years. We do not speak of isolated clever tricks that help in particular situations. We refer to a way of life in an integrated domain where hunches, cut-and-try, intangibles, and the human "feel for a situation" usefully co-exist with powerful concepts, streamlined terminology and notation, sophisticated methods, and high-powered electronic aids.[11]

The systems capable of such an automation are to be "sophisticated" and able to deal with complexity. Those systems are to be combined with human intelligence. They are not intended to replace it. So, the general aim of artificial intelligence is also open regarding augmentation and automation. This openness in the general aim in the relationship of AI and humans is reflected in the variety of purposes for which AI systems can be used. The set of technologies described by the term AI are so called general-purpose technologies (GPTs). While the concept of GPTs has mainly been applied in economics,[12] it fits well as a category for analyzing social impacts of technology. The many purposes for the use of coal and steel have been captured in the phrase "swords to ploughshares." Maybe the same might

11 Engelbart (1963: 1).

12 Rousseau (2009).

be said about AI, which can fuel lethal autonomous weapon systems as well as assistive care robots. In order to understand AI, the comprehension of the general-purpose nature of the respective technologies is of the utmost importance. The technologies comprising AI are neither exclusively tied to certain risks and challenges nor to certain opportunities and advantages. There are many counter-intuitive examples for this proposition, but data protection and privacy are again very illustrative in that regard. AI systems can certainly be a threat to privacy and data protection as they allow the extraction of a lot of personal information. One interesting aspect is AI-powered shadow profiling. This means that people are profiled without any significant activity of their own. The information is provided by people around them. Circumstantial evidence, such as search queries from other persons in a social network, allow smart systems to reconstruct a profile of a person within that network and collect relevant personal information, without them having personally revealed anything. However, AI can also help to serve as a privacy enhancing technology. There is, for example, a general push for chatbots that learn the privacy preferences of a person in a short and simple conversation and then go on to adapt the privacy preferences in all networks and online-services the person uses. The purposes of AI systems can, therefore, both enhance and threaten privacy. As will be shown, the same is true for other principles and values such as democracy.

Aspect of Openness	**Alternatives**
Research question	Weak AI thesis, strong AI thesis
Technologies	Machine learning technologies (artificial neural networks, generative adversarial networks), good old-fashioned AI
Resources	Data, common sense, computation…
Aims of use	Automation vs. augmentation
Purposes	General purpose technologies: can go many ways regarding purposes like transparency and data protection

Table 1: Dimensions of Openness of AI

The importance of this openness can be appreciated to a fuller extent when recognizing the ways in which democracy can be shaped by technology. Firstly, there are different understandings and constructions of the meaning of democracy. While there is a common thread of self-government of a people, there are differing

views on how this self-government is to be exercised. Democracy is constituted in actual practices in society. Technology has always played a huge part in the actual practice of democracy. Democracy and technology are intertwined. "Democracy is not enacted and then mediated. It is preformed through acts of mediation. Technologies of mediation are and always have been inherent in the social enactment of democracy."[13] One, therefore, can go as far as tying practices of the use of certain technologies to specific ideas of democracy.[14] The use of technologies configures democracy. In the case of AI, being a set of general-purpose technologies, this configuration is generally open. Democracy is a process rather than a fixed and attainable state. It has to be constantly realized, using means like technological innovation, institutions, markets and competition, law and administration.[15] In the face of this openness, it is interesting to look at current and potential uses of AI in the context of democracy.

B. Empirical Insights

While general purpose technologies like the internet or AI can play out very differently, they are usually described in a particular way. The discourse on the internet and democracy began by hailing the potential beneficial effects of the internet on democracy.[16] Regarding AI, it seems to be the other way around: it is mainly regarded as a threat to democracy. AI is seen to have the potential to obstruct established democratic processes like elections and votes. There is also a fear that AI takes over decision making in many contexts. In order to paint a more nuanced picture, one has to appreciate the contingency of the technology and how it can be used in very different ways. The literature on the internet today recognizes its positive and negative effects on democracy.[17] The contingency of the internet means that "like every medium before it, from the alphabet to television, [it] is shaped by the ways that society chooses to use its available tools."[18]

The general-purpose nature of AI is also reflected in its relationship with the democratic process, especially in the context of elections. In this regard, AI is generally perceived as a threat. There have been several attempts to influence elections through automated systems that preformed different tasks. Fake news are spread in the context of elections to block and obstruct political discourse and to target voters on a granular level in order to engage or disengage them from

13 Coleman (2017: 27).

14 Bozdag/van den Hoven (2015).

15 Irrgang (2002: 173).

16 Pernice (2016).

17 For an overview see Ceron/Curini/Iacus (2017: 6).

18 Coleman (2017: vii).

voting.[19] One of the activities that has been fueled by AI and other digital technologies is micro-targeting. Micro-targeting denotes attempts to influence the behavior of people based on personal profiles and actions that are grounded in specific features of that profile. Those profiles provide specific information about certain persons; people can then be targeted individually through social media advertising instead of being addressed as part of a group with political posters or TV commercials. These actions can range from attempts to influence or obstruct democratic discourse to influencing or obstructing the actual decision-making of individuals. While the first micro-targeting efforts were used for canvassing campaigns, in which humans went door-to-door in order to influence the electorate, AI can also play a role in actions based on granular profiles of certain people. There have been several reports about the use of such technologies. Whereas the elections in the United States and Brazil and the Brexit vote have made the news, their use has also been debated in states like Switzerland and Austria.[20] AI systems can enhance the possibilities of micro-targeting on different levels. AI can help with the extraction of information by crawling the web and analyzing other sources of unstructured data. AI systems can also help to profile people. Furthermore, AI systems can automatically address persons based on their profiles through different channels like social media. Several aspects of these campaigns using micro-targeting are problematic.[21] First, the respective data has often been collected from public sources, in some instances illegally. This violates the respective persons right to data protection if the data was collected and used illegally. It also violates their right to personal autonomy, in that they are being influenced based on the collected data. Opting out of micro-targeting is not yet an option. What is more, micro-targeting can also be used for purposes of manipulation. Research on the topic also mentions the possible beneficial impacts—such as ensuring that voters receive the information that is relevant for them.[22] This could also make specific topics more relevant for elections and enhance the importance of certain groups, particularly when they are spread out and not organized.[23] Therefore, AI could help those conceived to be weak and less powerful to obtain more and better information.[24]

19 Bodó/Helberger/Vreese (2017: 3).

20 Eidgenössischer Datenschutzbeauftragter/Konferenz der schweizerischen Datenschutzbeauftragten (2018); Der Standard (2019).

21 A mapping of the threats can be found with Zuiderveen Borgesius/Möller/Kruikemeier/Ó Fathaigh/Irion/Dobber/Bodo/Vreese (2018: 87) On the same page, they collected reference on privacy and manipulation trends.

22 Zuiderveen Borgesius/Möller/Kruikemeier/Ó Fathaigh/Irion/Dobber/Bodo/Vreese (2018: 84ff).

23 Ibid.

24 Ennals (1987: 14).

This shows again the general-purpose nature of AI and the difficulty of putting it into one box. Micro-targeting can be detrimental, but it can also be beneficial to democracy. Yet, the applications existing today are only a preliminary view of what could be possible. Technological improvements, but more importantly also creative and innovative uses of the technology could lead to an even more profound impact of AI solutions on democracy. AI solutions can be something genuinely new or turn existing possibilities upside down. One example would be to empower voters through targeting and profiling candidates. A smart search engine could help to identify information concerning how parties or candidates think about certain issues. Empowering voters even further, one could come up with AI systems that predict future government behavior. One could try to compute the probability that parties or candidates act on certain promises. Indeed, it seems to be not entirely impossible to predict the likelihood of the question of whether certain promises will be kept in the future. This would be a use of profiling in a completely different way. While such a profiling of candidates and parties raises a series of problems and issues, it shows that the use of AI can vary greatly and also support voter empowerment. It could open their decision-making potential as opposed to narrowing it. Whereas there is currently great concern for using AI in the context of elections and votes, the future impact of AI is in fact open.

C. Law and Technology: Limitation, Motivation, Design

The law and technology have a multi-facetted relationship. This relationship can be broadly summarized in three functions: limitation, motivation and design. The impact of law on the relationship of technology and democracy will be explained along these lines. The law can add to the democratization of artificial intelligence in different respects. To include all these functions in the picture is particularly important since they highlight different perspectives that are best suited to create a full picture of the challenges and opportunities of AI in relation to democracy.

1. Legal Limits and Democracy

Human rights limit the use of AI, especially by public authorities. Human rights also trigger the need for democratic justification. Thereby, they further limit the possible uses of AI. The function of the law as limit to technology is possibly its best-known function. Legal obligations stemming from data protection, for example, limit the use of technology in several respects. Data protection law can ban the use of training data in machine learning, because there may be no legal grounds for such use or existing allowances do not cover the respective purpose. For instance, under the EU General Data Protection Regulation (GDPR) data pro-

cessing has to be justified according to Art. 6.[25] Systems taking automated decisions have to comply with Art. 22 GDPR. This provision allows such decisions only when the requirements in sections 2 and 3 are met.[26] Sections 2 and 3 refer to decisions based on contracts, statues or explicit consent.

2. Motivation

The law can also motivate the use of technology in different forms. This motivation can relate to "the development, advancement and application of technology by the administration or even make it compulsory."[27] There are different ways in which democracy as a legal principle can motivate the use of technology and AI specifically. Looking into international instruments about democracy, one can spot questions of technology in different contexts.[28] In human rights law, there are several rights that point to democratic governance. Some human rights instruments explicitly point to the crucial importance of technology in order to enhance democracy.[29] One area in which this is of particular importance is the inclusion of persons who are not yet able to effectively participate in democratic procedures and democratic discourses. The United Nations Convention on the Rights of Persons with Disabilities. Art. 4 para. 1 (g) obliges "to undertake or promote research and development of, and to promote the availability and use of new technologies, including information and communications technologies, mobility aids, devices and assistive technologies, suitable for persons with disabilities, giving priority to technologies at an affordable cost." This is an example of a progressive human rights clause that motivates states and other actors to employ technologies in order to further human rights. Many AI technologies help persons with disabilities, especially blind and death people. These technologies also empower their respective users to participate in democratic discourse. Therefore, Art. 4 para 1 (g) has an effect on people's democratic inclusion.

3. Design

Another function of the law is to structure and guide the design process. The law sets design goals, it shows how to balance different goals and even highlights possibilities to solve issues on the technical level. A good example for that is the privacy by design clause in Art. 25 sec. 1, which provides as follows:

25 Art. 6 provides that processing of data is only lawful if its requirements are met.

26 Abel (2018); Martini (2018).

27 Djeffal (2019: para 16).

28 This research is based upon the collection of documents by Ehm/Walter (2015).

29 See for example ga-Res. 68/164. Strengthening the Role of the United Nations in Enhancing Periodic and Genuine Elections and the Promotion of Democratization, United Nations A/RES/68/164, adopted by the General Assembly on 18 December 2013 (70th plenary meeting).

> Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

Art. 25 section 1 entails a direct obligation to include privacy considerations into the process of designing or adopting an application. It is, however, also possible to have rather indirect obligations. It has recently been claimed that constitutional principles such as human rights, the rule of law and democracy also should be included in the process of designing AI.[30] This would further the law's function to influence technologies at a very early stage. These obligations also have to be applied by those developing the systems directly. In order to meet those obligations, several methodologies have been invented in different domains. While there are different standardization processes regarding constitutional values, there is yet no specific standard in dealing with AI and democracy. To date, nothing specifies a general obligation to include the principle of democracy into the design of AI.[31]

D. Legal Reasons and Lessons for the Democratization of AI

This section sketches the main legal reasons for democratizing AI as well as some learnings from the relationship of law and democracy. Democracy as a principle is enshrined in the constitutions of many states, be it implicitly or explicitly; it is also a basic value for international organizations such as the Council of Europe.[32] Such a constitutional principle demands its realization in the public sphere. Apart from this very general democratic requirement, there are more specific lessons that can be drawn from the way that law functions. Three insights will be discussed in greater detail below.

30 Nemitz (2018).

31 Current value-sensitive design standards can be found with the respective ISO projects and in IEEE's P7000 series.

32 See for example the preamble of the statue of the Council of Europe from 5 May 1949, ETS No. 001.

1. Justification

As mentioned above, human rights add another layer to the limitation of technology. They set absolute limits on the behavior of public authorities and force them to realize human rights. Human rights are also tied to democratic decision-making. Whenever a measure touches upon human rights, it can only be lawful when there is a democratic justification underpinning it. The Covenant on Civil and Political Rights uses the phrase "rights shall not be subject to any restrictions except those which are provided by law." Meanwhile, the European Convention on Human Rights uses the phrase "in accordance with the law." This means that restrictions of human rights must be provided by law.[33] In order to qualify as a justification, the impact must be described by law in a manner that is understandable for the individual. The law here is a proxy for a democratic *ex ante* decision-making. Any impact upon human rights must be preceded by a democratic decision allowing for the precise impact and providing for safeguards for excessive and arbitrary uses. Another example is the Charter of Fundamental Rights of the European Union which provides that "[a]ny limitation on the exercise of the rights and freedoms recognized by this Charter must be provided for by law."[34] This provision makes the need for democratic justification explicit. In the absence of such a justification, a measure is necessarily unlawful. The law is a vehicle to enforce human rights. It is also a medium for democratic decisions. This strong link between human rights and democracy mediated by the law also affects the relationship between AI and democracy. Whenever AI systems have an impact on human rights, their use is to be justified.

This necessity for democratic justification does only depend on the fact that the human capacity to make decisions affected by an AI system. Democratic justification is not only triggered by specific human rights. The need for justification of AI systems certainly applies to so-called automated decision systems (ADMs) that are often in the focus of academic attention. This is only one among many ways in which human rights can be at issue. When AI is used as a watchdog for IT security or for maintenance of critical infrastructure, it is crucial for realizing human rights. While the right to privacy and self-determination might be the most obvious examples of such impacts, other subtler influences also need to be considered. For example, ADMs are frequently regulated, but this regulation never applies to recommendation systems. Yet, these recommendation systems can have substantial impacts on human rights. Highlighting the interdependence between artifi-

33 See for example the explanation by Greer (1997: 9).

34 Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391-407. The relationship between legal democratic justification and human rights is not as universal in every human rights instrument. The universal human rights covenants, for example, require legal democratic justification only in certain cases.

cial intelligence and human rights, there is a rather clear criterion for the need of democratic determination. This is not the "power" of a machine to decide, but the impact on human rights. The interrelation between human rights and democracy can limit the public use of AI systems. If there is such a relationship, democratic justification is necessary—irrespective even of how human rights are affected.

2. Supremacy

Tied to this necessity for democratic justification issues impacting human rights is the idea of the supremacy of legitimate democratic decisions. This concept has found its expression in the idea that the norms made by the organization with the highest democratic legitimacy take precedence over other norms. Therefore, many jurisdictions which regard parliament as the highest democratic authority rely on the "sovereignty of parliament" and regard parliamentary laws as taking precedence over all other legislative acts. Other jurisdictions describe a normative hierarchy in which the constitution is at the top and acts of parliament in the second place. While constitutional law derives its legitimacy from the *pouvoir constitutant*, statutory law relies on the legitimacy of parliament and yet other norms stem from actors with less legitimacy. Higher norms take precedence over lower norms, in cases of conflict, lower norms are either rendered invalid or inapplicable. The hierarchy of legal norms is generally grounded in different levels of democratic legitimacy.[35] In cases in which technology has normative force, this general idea would require that the law as a proxy of democratic decision takes precedence over functional requirements of technology and must actually guide democratic decisions.[36]

3. Democratic Rebalancing

From a legal point of view, the notion of democracy is open. While there are many ways to understand and construct what democracy ought to mean, constitutional law is generally open towards the multiple understandings and theories of democracy. This openness allows the law to adapt to different contexts and different situations, especially when changes and reforms are at issue. Such reforms can happen on different levels, but they always change prior democratic processes and sometimes even the notion of democracy itself. One pattern that can be discerned from the way in which courts deal with these issues could be described as the mode of rebalancing. Courts remain flexible and open towards changing existing

35 From a legal positivist standpoint, it would also be possible to arrive at the same conclusion arguing with validity. One would then have to argue that the basic reason for validity is democracy.

36 See for example Schulz/Dankert (2016) It is important to note, however, that such a hierarchy must be based on democratic legitimacy and not on a formal distinction of primary and secondary rules.

processes, but they require active steps that would rebalance the situation from a democratic standpoint. This rebalancing can mean that there are measures that effectively democratize the new institutional arrangements. Two examples from other contexts can illustrate this. In the process of European integration, there were many treaty revisions creating new competences or transferring competences from the national to the European level. The German Federal Constitutional Court had to deal with creation and transfer of competences on several occasions. In its famous Lisbon judgment, the court allowed for a transferal of competences, but it also required institutional arrangements in the German legal order, enabling the legislature to effectively play a role in European politics. So, while it agreed to supranational power transferals, it only did so on the condition that the national legislature could influence politics at the higher level.[37] In another case, the Constitutional Court of Baden-Württemberg had to deal with a transferal of powers from the collegiate of professors to the president of a university. The court allowed for this transferal of power, but only on the condition that the president become accountable to the collegiate of professors, which in practice meant that a democratic election process had to be created.[38] These cases show that changes and reforms with an impact on democratic processes are—from a legal standpoint—not to be evaluated in a binary fashion of "yes" or "no." Changes sometimes require democratic rebalancing. If there are disputes on how to rebalance those changes democratically, those disputes can ultimately be resolved in legal proceedings. These questions of rebalancing play an important role when actions and decisions are delegated to AI systems on a greater scale. Instead of arguing that this would be undemocratic, the question is rather whether this delegation to AI systems can be rebalanced. This flexible view present in different democracy cases also has the potential to shift the relationship between AI and democracy. Instead of asking whether AI should be democratized, the question is how it can be democratized and whether the respective measures are enough.

II. How to Democratize AI

If there is a need to democratize AI, how can it be put into practice? An instrumental approach to that question would first look at instances in which there are democratic choices and secondly at ways in which these decisions can be made. As

37 BVerfG, Judgment of the Second Senate of 30 June 2009—2 BvE 2/08—paras. (1-421), http://www.bverfg.de/e/es20090630_2bve000208en.html para 273ff.

38 Landesverfassungsgericht Baden-Württemberg, judgment 14.11.2016, 1 VB 16/15, obtaible at https://verfgh.baden-wuerttemberg.de/fileadmin/redaktion/m-verfgh/dateien/161114_1VB16-15_Urteil.pdf p. 43ff.

in every other democratic decision, there are different tools ranging from the ordinary processes of parliamentary deliberations and decisions to more direct versions of democratic participation. Each method might have specific advantages in a certain setting. Such a democratic toolbox could contain the following elements among others:

- ordinary parliamentary processes to debate and regulate artificial intelligence
- use of specialized parliamentary committees to determine certain issues
- empowerment of experts to make certain decisions according to preconfigured principles
- direct involvement of citizens regarding certain questions through
  - participatory methods
  - sortition: involving groups of randomly selected citizens in order to fulfil an office or make certain decisions
  - random sample voting: in order to vote on specific questions, a representative sample of the population is selected

For the sake of understanding the range of choices to be made about technologies and specific technical artefacts, it is helpful to distinguish between different layers analytically, despite the fact that the interrelations between the different layers are obvious. Focusing on specific choices regarding technical artefacts, there are choices that are rather technical and others that are rather social. Therefore, a distinction is to be made between a social and a technical layer. Furthermore, some decisions are not made with a view to a specific artefact but rather regarding a technology. These choices are situated in a layer of governance. On every layer, there are specific questions to be outlined.

A. Technological Layer

1. Design Choices

An important step in the democratic determination of technology is understanding the choices that are made in the course of inventing or applying a technology. Many design choices are made in the development. Some design choices are made intentionally, some have important consequences. From a democratic perspective, one must understand and highlight specific choices. These choices relate to architectures, applications and all other features of the technologies used. Whenever there is an alternative, there is a choice. Understanding choices also requires a democratic mindset that is open to several possibilities without automatically preferring certain outcomes. Computer scientists especially, who are trained to achieve specific goals such as efficiency, regularly do not see behind the choices that maximize their preferred value.

In order to appreciate choice in the case of machine learning, questions of optimization are very interesting.[39] Machine learning systems are optimized to attain certain goals, they receive feedback and adjust their model accordingly. In many cases, the goals towards which a model is optimized are not set in stone but rather contingent. An algorithm that distributes children to certain schools within an area can be optimized according to different goal functions: One could be the shortest way to school. Another would be the safest way to school. Yet, one could also define other goals such as a good mix of students in school from an ethnical or economic perspective. Such choices often result in trade-offs. They require an active choice. One trade-off that has become better known as of late is the choice between using data and being able to understand discrimination. Machine learning models are often trained on data that contains implicit biases—at the same time, training data may not contain explicit references to age, gender or other criteria. Thus, the decisive information is not present and it becomes impossible to understand whether there is bias in the data and consequently also in the algorithm and whether remedies are possible. Yet, including more data, e.g. age or gender, impacts the right to privacy and data protection. Especially in possible cases of discrimination, it would often be necessary to use special categories of personal data, such as data revealing racial or ethnic origins, that is heavily protected under many data protection regimes.[40] Therefore, it is necessary to weigh privacy and data protection against fairness in this regard. Another trade-off can happen when it comes to weighing transparency and accuracy. It is possible that some algorithms have higher scores than comparable alternatives but are based on models so complex that they are not intelligible for humans. There is an increasing awareness in the computer science community that choices are not only made in the process of using existing technologies but also in the process of research and development. In the same way that privacy enhancing technologies were invented, new communities have sprung up doing research to improve AI in specific directions. One example is the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT), that looks specifically at new research on fairness, accountability and transparency in socio-technical systems. Similar conferences or tracks on AI panels show how research and development can also be specifically directed towards certain aims. Again, there is an element of choice even when it comes to creating or improving technologies. In this case, these choices can be exercised by researchers, but also influenced by funding agencies. An element of choice is often present at different stages.

39 Haferkamp (2017).

40 See for example Art. 9 GDPR.

2. The Principle of Designability

Scholars and institutions have called for the inclusion of democracy by design in the context of AI.[41] In line with the idea of value-sensitive design, democratic values should be included in the design process. Not only are design choices to be made in a democratic manner, the very way in which the application operates is to be democratic. Yet, this general idea encounters several difficulties. One problem is that there are varying concepts of democracy and they can play out quite differently in the design of an application.[42] One way to structure the different forms of democratic legitimacy is to divide them into input, output and process legitimacy. Technical requirements can be quite sophisticated. Depending on the context in which the AI application is used, democracy can involve very different actors as well: in the smart city context, democratic decisions will often require decision making or participation by the municipal population. In national settings, it will be more about involving parliament in decisions. For these reasons, the assertion of democracy by design means a lot of uncertainty for developers. What would be needed from a technical perspective is a principle that developers can grasp and one that supports democratic values in design processes without prejudging certain understandings of democracy.

My suggestion to address this challenge would be to formulate a design principle of designability. The principle of designability is aimed at translating general democratic values into design in a general and workable manner. It ought to have at least two tiers that need to be addressed by developers: The first tier is the changeability of the system. The second tier is its intelligibility. Different ideas of democracy rely on the idea that they are open and flexible to different forms of change: changes in government, changes in opinion after an informed discourse and so on. This is particularly the case if there is uncertainty about how a decision plays out in practice. In such a situation, changeability is a requirement for democratic participation. Yet, such changeability has to be enhanced by design. This can be done by choosing a specific architecture or using specific methods. Considering that machine learning entails the possibility to adapt, it is changeable by definition. Another tier for designability is the intelligibility of the system. Intelligibility is not used in its general sense in computer science, that is the possibility to understand the logic behind a given system's actions. Intelligibility must be constructed democratically. A general target here could be that a system is intelligible for all people affected by the actions of the system. While not everybody will in effect decide upon whether and how to employ the respective AI system, the ideal would be that everybody should have the chance to. This standard of in-

41 See for example Nemitz (2018); Die Bundesregierung (2018: 33, 44); High Level Expert Group on Artificial Intelligence (19).

42 Bozdag/van den Hoven (2015).

telligibility can be rather narrow in the case of systems that are targeted only at a specific group of people. In contrast, generally applicable AI systems should meet general standards of intelligibility. Therefore, the tier of democratic intelligibility fits in with current discourses on transparency. Yet, in the context of designability, intelligibility is not limited to specific actions or decisions made by the system. The people affected by the system have to understand it and the choices underlying it. They have to know whether and how the system can be changed. Like any design principle, designability will hardly ever be achieved fully. But it can point developers in the right direction. While intelligibility points to the possibility of democratic deliberations, the tier of changeability indicates the possibility of change and opens up potentials to effectively govern the technical artifact.

B. Social Layer

AI is not only designed on the technical level, many social constructions surrounding AI systems play a crucial role.[43] These social constructions are not inevitable, they are the fabric of choices and assumptions that are shaping technology and society at the same time. The law is a mechanism that can make socio-technical choices subject to democratic determination.

1. Understanding Impacts

It is important to appreciate the social impact of technology, but also to understand that the recognition of such impacts are social constructs themselves. Recently, different methods to assess the impacts of AI have been proposed.[44] Impact assessment is a prerequisite for uncovering choices on the technological level. Sometimes, the respective choices only become apparent and understandable when the social impacts are known. The discussion about fairness in AI took off when several researchers criticized discriminatory effects of algorithmic systems. The same is true for transparency. To learn about the consequences of technologies before harm and damage occurs is far from easy. As the history of technology shows, the knowledge about the consequences of technologies often comes too late. The discovery of radiation is a telling and sad example, since many of the scholars discovering this technology did not know about its dangerous effects and later died from cancer. It took some time to understand the effects. In many other instances, the causal relationship between technology and impact was not as apparent or more contested. In these instances, the law has profound effects on the social construction of technology.

43 Stamper (1988).

44 Reisman/Schultz/Crawford/Whithattaker (2018); ECP (2018).

Firstly, human rights law can provide for a consensus that a certain consideration is worthy of protection. In order to know what constitutes an impact, one has to construct a value that is to be protected. The law can create a consensus of what that is. The right to privacy is a good example of a right that has been invented through deduction in an evolutionary manner from other legal institutions.[45] Once there is an agreement on what is to be protected as human right, a special protection is in place. As has already been shown, this protection entails the need for democratic justification of decisions affecting human rights. Another important feature of the law is its ability to recognize and balance impacts in a holistic manner. Impacts are not negative by definition. They can equally be beneficial. While it is important to be critical towards new developments and to understand new dangers and disadvantages, it is as important to appreciate the benefits and potential opportunities. In order to assess the impacts of technologies, it is crucial to have all of the future possibilities in mind. This is also true from a human rights perspective. As shown above, technologies also have the potential to further human rights. Therefore, the consequences have to be weighed against each other. In order to assess such situations in legal proceedings, several jurisdictions have developed a proportionality test.[46] It is a practical way to assess a measure holistically and to structure the argument in a way that allows for many considerations and to weigh them against each other. It also arrives at practical conclusions that are communicated to those affected by the decisions. The principle of proportionality actually allows for a socio-technical evaluation on different levels.

2. Designing AI through Social Construction

Yet, there is an even wider sense in which the impacts of AI are socially constructed. This applies to a large part of the influence of AI systems. Especially in the case of data analytics, there can be different goals and aims: to discover certain correlations, to discover probabilities of certain actions or to actually show probabilities of how certain alternative actions might play out.[47] While it is true that those systems can have profound normative effects, such effects often stem from the social construction of the system instead of being falsely pinpointed as inherent in the technology. Whereas big data analytics tools compute certain probabilities, for example, the meanings of those probabilities and the role they should play is actively constructed.[48] One illustrative example is the misuse of scores for cred-

45 See one early argument in Warren/Brandeis (1890).

46 Klatt/Meister (2012).

47 On this basis a distinction is made between descreptice, prescrpitive and descriptive analytics by Hoffmann-Riem (2017).

48 See for example Schlaudt (2018).

itworthiness as a reliability score for employees.[49] It is obvious that a system that is designed to compute the probability of a person repaying debts is not made to assess the respective persons reliability when it comes to the job. Yet, the choice to use the system in another context is by no means a choice that has anything to do with the design of the system. It is rather a social choice for a transfer to a different social context.

The same holds true for the use of certain probabilities. In many instances, the law shows how probabilities have completely different meanings in different contexts. In police and security law, there are also different probability requirements that are formulated from a social perspective. Measures that have low impacts on human rights have to meet a lower probability threshold, while measures with higher potential impacts have to meet higher probability standards. It is an active choice, and a democratic decision, to link a specific competence of the authorities to a certain probability.

There are numerous ways in which to construct the meaning of outputs of AI systems. The law not only makes this meaning explicit; it opens up the social construction of technology for democratic deliberation and democratic decision-making. The outputs of AI systems can be rendered illegal and irrelevant. They can be made subject to human oversight and human decision-making. Furthermore, they can be bestowed with the force of the law. In German law the assessment of civil servants, decisions must not be based on fully automated assessments of specific personality features.[50] The above-mentioned Art. 22 GDPR provides for a right of human oversight and makes fully automated decisions subject to human decisions. Yet, there are provisions clarifying that fully automated decisions do have the force of law. Take for example § 35a of the Federal Code of Administrative Procedure. The provision states: "An administrative act may be adopted in full by automatic systems, provided that it is authorised by a legal act and that there is no discretion or margin for assessment." This provision clarifies that there can be completely automated administrative acts, i.e. decisions with legal force for specific individuals or groups. This basically means that those systems can render decisions that have the force of law and can also be enforced. Two examples for such decisions are intelligent traffic systems that can automatically set speed limits or impose overtaking bans when there are dangers for the drivers due to wheather or traffic. Another example is fully automated speeding tickets issued from detection systems that automatically send the respective notices.

49 O'Neil (2016: 147-149).

50 See § 114 section 4 of the German Federal Civil Servants Law.

3. AI as Customary Law

AI systems can have real world impacts which depend to a large extent on a social construction that attributes these consequences to the system. This leads to the question of what the requirements of such acceptance should be. This question is currently addressed by the field of computational social choice.[51] The hidden moral choices in the process of designing AI is one of the main motivations to engage with the interlaces of social choice and computer science. So, the proponents of computational social choice try to find criteria to design AI systems in a legitimate way. One feature that is striking with machine learning is that it is actually based on data that is often produced by those to whom the system applies. Research projects have, for example, used inquiries and simulations in order to obtain user data on how automated cars should react in specific situations.[52] Yet, a democratic view on this ethical design focus reveals certain issues: The first problem is that different assumptions can lead to quite varied results, which might all have a claim to be ethical. Different ethical theories can even produce opposite results. Take for example utilitarianism and principled ethics. While certain actions detrimental to one person but beneficial for the majority could be regarded as ethical from a utilitarian perspective, they would be regarded as unethical from a principled point of view. In the end, it might be necessary to choose among many alternatives. To state that there is only one right and moral solution to be preferred over all other solutions is to discriminate against all other possible solutions. It neglects various approaches and different solutions to a single question. In such a setting, there is no room left for choice. Another question is whether artificial agents can genuinely make moral decisions or whether they are just simulating them. From a moral point of view, the question of actual judgement is paramount. This problem is tied to the question whether machines can actually think, which has attracted contentious reflection from Turing to Searle.[53]

The basic argument of this section is that computational choice theorists should think in legal instead of moral terms. Building upon Kant, one could attribute actions with external effects to the law, while questions that remain internal are in the realm of ethics. AI systems often have profound normative effects. While most ethical considerations focus on output legitimacy, one could merge computation and law in a way that democratic input legitimacy is achieved through legal means. Machine learning applications are generally trained with data that represents the behavior of certain actors. While there is no general formalized rule about what the significance of such practice is, I would like to make the argument

51 Brandt/Conitzer/Endriss/Lang/Procaccia (2016) A overview of the literature regarding AI is given by Prasad (2019).

52 Awad/Dsouza/Kim/Schulz/Henrich/Shariff/Bonnefon/Rahwan (2018).

53 Turing (1950); Searle (1980).

that machine learning could—under certain conditions—be regarded as customary law. This would highlight computational and social choices that allow for a democratic expression through an AI system. Building upon an analogy from certain law-creating practices, it could be possible to formulate requirements for AI as a medium for democratic decisions.

Customary law used to play a very important role for the governance of certain communities that regarded specific practices as binding. Spurred by the increasing complexity of modern societies and the possibilities of new printing technologies, customary law lost much of its importance. It mainly relied on unwritten practices of smaller communities that formed over time. While courts in the common law countries continued to rely on once formed principles and turned them into arguments the judiciary could build on, one legal system in which customary law has retained its importance is international law. In international law, there is still a manageable number of participants whose practice can be qualified as custom. Several trends of digitization assist a new knowledge dimension that might lead to a revival of customary law in different areas. First, datafication opens new avenues to store and understand the behavior of certain actors. Big data represents the idea that huge amounts of data can be stored and analyzed. Secondly, trends like the internet of things allow for the collection of data in a constant, automated and ubiquitous manner. The internet of things signifies a trend of networked devices in different human environments. AI technologies can help to analyze and understand the data in a way that makes the practice comprehensible and understandable. Together, those technologies make actual practice of people visible.

However, the question remains as to whether this custom is meant to be generalized in human exchange. Scholars of computational social choice have thought about this issue and come up with criteria that were to be considered in the process of building an AI that represents practice. Baum, for example has developed with three general criteria:

1. Standing: Who or what is included in the group to have its values factored into the AI?
2. Measurement: What procedure is used to obtain values from each member of the selected group?
3. Aggregation: How are the values of individual group members combined to form the aggregated group values?[54]

The requirements of customary law are in some sense complementary, in some sense different from the questions above. The formal criteria for the formation of

54 Baum (2017: 545).

customary law are a practice (*consuetudo*) and the belief that this practice is to be regarded as law (*opinio iuris sive necessitatis*). The practice must be consistent and general, even though this does not mean that the practice is uniform and universal.[55] The most important question regarding general practice in the context of customary law is whether there is sufficient representation. This is due to the fact that some actors remain tacit and do not engage in the practice. The second criterion is the so-called *opinio iuris*. That is the belief that the respective practice is based upon a legal obligation to act in that way. This criterion actually legitimizes the normative force of the practice. In order to fulfil the criterion of *opinio iuris*, data subjects must produce the data in the knowledge with the purpose of influencing a system that acts upon that data. This criterion makes the legitimacy of an AI system subject to a sovereign decision of users. The system simply learns what the practice of human beings is. It learns what the data subjects want the practice to be. In this setting, informational self-determination is not only the power of personal data; it is a conscious exercise of power through one's data. The data subject is not a resource from which personal data are extracted. In this setting, the production of data becomes a democratic act like voting.

C. Governance Layer

In order to analyze the impact of AI on democracy, it is not enough to look exclusively at specific systems. It requires an analysis from the macro level focusing on technologies or even AI as a whole. This is here denoted as the governance layer.

1. Framing

The democratic governance of AI is influenced by the way in which AI is framed. AI is regularly put in specific contexts or seen a certain way. Frequently, scholars talk about the ethics of AI,[56] another current is to talk about AI and human rights. While scholars discuss and analyze within one frame, there is relatively little discussion about the choice between frames. Yet, the frames do have significant effects. Take for example the choice between an ethical and a political frame.[57] The frames lead to completely different ways of thinking about technology. Compare stem cell engineering and the creation of a 5G network infrastructure. Stem cell engineering is predominantly construed as an ethical issue whereas the latter is commonly perceived as a political issue. Of course, there are many issues we would conceive of as being political in the context of stem cell research and there can be many ethical questions in building a 5G infrastructure. Constructivist

55 Crawford (2012: 23ff.).

56 Mittelstadt/Allo/Taddeo/Wachter/Floridi (2016).

57 For this reflection see Djeffal (2019).

scholars have highlighted that frames and theories influence the object of scientific inquiry. Therefore, it is an active choice to put AI in certain context and to inquire into the ethics or politics of AI or to look to the relationship of AI and human rights. This choice necessarily contains certain preferences that are inherent or follow from the frame that was adapted. Every frame also provokes some blind spots. Some aspects become invisible.

One attempt to generally describe the impact of AI on society is the concept of "algocracy." This term contrasts other forms of government such as democracy or monarchy with a system in which power is (increasingly) exercised by automated systems.[58] The term algocracy is mostly used in a critical manner.[59] It highlights that algorithms are becoming more and more important when it comes to issues of governance. Instead of adding to the growing corpus of literature on this issue, I would like to highlight the constructivist nature of algocracy. This leads to the question of what is highlighted by this term and what is left out of the picture. Building on the basic insights from actor network theory (ANT), I argue that the frame of algocracy tends to blur and hide human agency. Algocracy highlights machine power but overlooks how humans impact the perceived automated actions. One of the basic arguments of ANT is to ignore the distinction between subjects and objects and to appreciate technology as part of the social in a network with human actors using it.[60] This analysis allowed the proponents of ANT to uncover the agency of technical artefacts. My basic argument is that this theory might today be used upside down in order to uncover human agency instead of machine agency. The theory of algocracy represents a critical part of the AI discourse that frames AI specifically as automated decision systems and looks at their increased power. With the focus on increasing ability and power of those systems, it is sometimes forgotten to reflect on how these systems are used and interwoven with human agency. As outlined above, there are many ways in which the social surrounding determines the design of AI applications. In many cases, the law is part of constructive efforts to bestow AI with normative force. A frame that is complementary to algocracy would not exclusively look at the fact that more and more decisions are delegated but at *how* they are delegated and *who* controls and influences the automated systems. As many proponents of ANT have argued, the focus would not be on a single class of actors but rather on their interrelation.

58 Yeung (2018).

59 Danaher (2016).

60 Latour (2000: 180).

2. Organizational aspects

Another way to impact the development and deployment of artificial intelligence is through organizational measures. Many of the recent AI strategies contain such measures. On the one hand, organizational changes are aimed at enhancing technological progress in the field of AI. New institutions are founded, either to engage directly in research and development, to fund such activities or to enhance the network of already existing organizations. The United Arab Emirates made headlines with a minister for artificial intelligence[61] and the German government recently founded an agency for "innovation leaps" tasked with funding research and development for ground-breaking innovations and increasing implementation. On the other hand, newly founded organizations also exercise oversight over AI systems. In fact, there are indeed many organizations endowed with this task already. Organizations like the US Federal Drug Administration or its counterparts in Europe and elsewhere have engaged in the certification of AI systems that are considered to be medical products. There are also calls for more oversight institutions.[62] Following examples in Canada, some states have founded AI observatories that aim to find out about the social consequences of AI. The future of work is one of the issues often addressed in this context.[63]

Organizational change is not always expressed merely in new organizations. Sometimes, organizations change from within by adapting to new tasks. One important development in this regard is the question whether a new job profile is needed across organizations. Data scientists are one profile that is currently on the rise. Yet, some think that a completely new profile of algorithmists might be needed.[64] The idea behind this is to have people with specific technical skills so that an organization maintains agency when it must deal with AI systems. The interesting aspect of this idea is that expertise would also be available to organizations that have previously not been associated with technological expertise. The job profile of an algorithmist has the potential to democratize agency when it comes to questions of algorithms. Knowledge about AI systems would be generally available. A question separate from this specific profile would be the interdisciplinary mix of teams working on certain AI issues. If AI is used in specific contexts, there might be a minimum requirement of roles and perspectives that need to be present. Therefore, organizations developing, using or assessing AI systems should think about what the right mix of these teams would be. While computer scientists are a necessary component of such teams, they are never enough. All in all, organizational challenges and changes are a very good example of how algo-

61 Tendersinfo (2017).

62 Tutt (2017).

63 See for example Die Bundesregierung (2018: 26).

64 Mayer-Schönberger/Cukier (2013: 189-192); Hill (2015: 284).

rithms impact their social surroundings and how changes in the socio-technical context of AI systems can effectively contribute to the respective governance.

III. Conclusions

The 1947 constitution of Bremen, an entity of the German federal state, contains a very interesting provision about the relationship between man and machine. The constitution states in Art. 12 section 1: "The human being ranks higher than machines and technology." This provision addresses experiences from the process of industrialization, during which machines, technologies and the new possibilities of production gained importance. It is interesting that the founders of the constitution felt the need to remind the people and those in power of the fact that human beings should rank higher. During industrialization, this did not address the increasing capabilities of machines to act so intelligently that they may even be considered as persons. It was rather the fact that, as capacities of production, so much importance was conferred upon them. So, the basic idea was to argue for a human-centered view despite the huge social and economic importance of technical artefacts. This basic idea can also be translated to the process of digitization, in which machines engage in solving problems that require a degree of intelligence previously considered exclusively reserved for humans. One aspect of this normative centricity of human beings is their exclusive status as the bearers of human rights. Equally important is the aspect of effective self-determination of people in the face of technologies' increasing possibilities. To rank higher does not only mean that humans must not be harmed by new technological possibilities. It means that people need to be in the driver's seat. It can be understood as a call for effective self-determination on different levels.

If AI continues to fulfil the high expectations and has continued impacts on societal development, it will be even more important for an all-encompassing value-sensitive development. From the perspective of the constitution of Bremen, one necessary component would be to think about the democratization of AI. In order to do this, it will be crucially important to understand AI as a set of general-purpose technologies that can be used in very different circumstances and very different ways to achieve multiple tasks. While it is important to understand where AI currently threatens democracy, it is as crucial to appreciate its opportunities. To understand the openness of the use and potential of technology allows us to choose whether to develop the technology further and which path to take. When it comes to the democratization of AI, some general truths about democracy apply: Democracy is a process, not an achievable result. It can be lost very easily, and everyone must work for it continuously along the way. Once we stop striving for it, it is gone. From this perspective, AI is just another challenge that has the

potential to bring society closer to the ideal lying behind Art. 12 section of the Bremen Constitution, as well as many other democratic provisions: to meaningfully put people in the normative center of all public power.

References

Abel, Ralf B. (2018): »Automatisierte Entscheidungen im Einzelfall gem. Art. 22 DS-GVO. Anwendungsbereich und Grenzen im nicht-öffentlichen Bereich«, in: Zeitschrift für Datenschutz 8, pp. 304-307.

Awad, Edmond/Dsouza, Sohan/Kim, Richard/Schulz, Jonathan/Henrich, Joseph/Shariff, Azim/Bonnefon, Jean-François/Rahwan, Iyad (2018): "The Moral Machine experiment", in: Nature 563, pp. 59-64.

Baum, Seth D. (2017): "On the promotion of safe and socially beneficial artificial intelligence", in: AI & SOCIETY 32, pp. 543-551.

Bodó, Balázs/Helberger, Natali/Vreese, Claes H. de (2017): "Political micro-targeting: a Manchurian candidate or just a dark horse?", in: Internet Policy Review (IPR).

Bozdag, Engin/van den Hoven, Jeroen (2015): "Breaking the filter bubble: democracy and design", in: Ethics and Information Technology 17, pp. 249-265.

Brandt, Felix/Conitzer, Vincent/Endriss, Ulle et al. (Hg.) (2016): Handbook of computational social choice, Cambridge: Cambridge University Press.

Ceron, Andrea/Curini, Luigi/Iacus, Stefano M. (2017): Politics and big data. Nowcasting and forecasting elections with social media, London, New York: Routledge.

Coleman, Stephen (2017): Can the internet strengthen democracy?, Cambridge, UK, Malden, MA: Polity.

Crawford, James (2012): Brownlie's Principles of Public International Law, Oxford: Oxford Univ. Press.

Danaher, John (2016): "The Threat of Algocracy: Reality, Resistance and Accommodation", in: Philosophy & Technology 29, pp. 245-268.

Der Standard (2019): Post löscht alle Informationen zu Parteipräferenzen, https://derstandard.at/2000095874780/Post-loescht-alle-Informationen-zu-Parteiaffinitaet.

Die Bundesregierung (2018): Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz, https://www.bmbf.de/files/180718%20Eckpunkte_KI-Strategie%20final%20Layout.pdf.

Djeffal, Christian (2019): "Normative Guidelines for Artificial Intelligence", in: Thomas Wischmeyer/Timo Rademacher (Hg.), Regulating Artificial Intelligence, Wien, Berlin, New York: Springer, forthcoming.

ECP (2018): Artificial Intelligence Impact Assessment, https://airecht.nl/s/Artificial-Intelligence-Impact-Assessment-English.pdf.

Ehm, Frithjof/Walter, Christian (Hg.) (2015): International democracy documents. A compilation of treaties and other instruments, Leiden, Boston: Brill Nijhoff.

Eidgenössischer Datenschutzbeauftragter/Konferenz der schweizerischen Datenschutzbeauftragten (2018): Leitfaden. der Datenschutzbehörden von Bund und Kantonen zur Anwendung des Datenschutzrechts auf die digitale Bearbeitung von Personendaten im Zusammenhang mit Wahlen und Abstimmungen in der Schweiz, https://www.edoeb.admin.ch/dam/edoeb/de/dokumente/2018/Leitfaden%20Wahlen.pdf.download.pdf/Leitfaden%20Wahlen%20und%20Kampagnen%20final.pdf.

Engelbart, Doug (1963): "A Conceptual Framework for Augmentation of Mans Intellect", in: Vistas in Information Handling 1, pp. 1-29.

Ennals, Richard (1987): "Socially useful artificial intelligence", in: AI & SOCIETY 1, pp. 5-15.

Gasser, Urs/Almeida, Virgilio A.F. (2017): "A Layered Model for AI Governance", in: IEEE Internet Computing 21, pp. 58-62.

Goodfellow, Ian/Bengio, Yoshua/Courville, Aaron (2016): Deep Learning, Michigan: MIT Press.

Greer, Steven C. (1997): The exceptions to Articles 8 to 11 of the European Convention on Human Rights, https://www.echr.coe.int/LibraryDocs/DG2/HRFILES/DG2-EN-HRFILES-15(1997).pdf.

Haferkamp, Björn (2017): "Was ist optimal? Nutzen und Fallstricke der Optimierung", in: Björn Bergh (Hg.), Big Data und E-Health, Berlin: Erich Schmidt Verlag, pp. 59-68.

Helbing, Dirk (2019): "Machine Intelligence: Blessing or Curse? It Depends on Us!", in: Dirk Helbing (Hg.), Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution, Cham: Springer International Publishing, pp. 25-39.

High Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Hill, Hermann (Hg.) (2015): Auf dem Weg zum Digitalen Staat – auch ein besserer Staat? (= Verwaltungsressourcen und Verwaltungsstrukturen, Band 30), Baden-Baden: Nomos.

Hoffmann-Riem, Wolfgang (2017): "Verhaltenssteuerung durch Algorithmen – Eine Herausforderung für das Recht", in: AöR (Archiv des öffentlichen Rechts) 142, pp. 1-42.

Hofmann, Jeanette (2018): "Digitalisierung und demokratischer Wandel als Spiegelbilder?", in: Franziska Martinsen (Hg.), Wissen – Macht – Meinung. Demo-

kratie und Digitalisierung die 20. Hannah-Arendt-Tage 2017, Weilerswist: Velbrück Wissenschaft, pp. 14-21.

Hofstetter, Yvonne (2016): Das Ende der Demokratie. Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt, München: Bertelsmann.

Irrgang, Bernhard (2002): Technischer Fortschritt. Legitimitätsprobleme innovativer Technik (= Philosophie der Technik, Band 3), Paderborn: Schöningh.

Klatt, Matthias/Meister, Moritz (2012): The Constitutional Structure of Proportionality, Oxford: Oxford Univ. Press.

Latour, Bruno (2000): Pandora's hope. Essays on the reality of science studies, Cambridge, Mass.: Harvard University Press.

Mainzer, Klaus (2019): Künstliche Intelligenz – Wann übernehmen die Maschinen? (= Technik im Fokus), Berlin, Heidelberg: Springer.

Martini, Mario (2018): "Art. 22", in: Boris Paal/Daniel Pauly (Hg.), [Duplikat] Datenschutz Grundverordnung: DS-GVO, München: C.H. Beck.

Mayer-Schönberger, Viktor/Cukier, Kenneth (2013): Big Data. Die Revolution, die unser Leben verändern wird, München: Redline.

McCarthy, John (1989): "Review of The Question of Artificial Intelligence edited by Brian Bloomfield", in: Annals of the History of Computing.

McCarthy, John/Minsky, Marvin/Shannon, Claude (1955): A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html vom 31.03.2017.

Mittelstadt, Brent D./Allo, Patrick/Taddeo, Mariarosaria/Wachter, Sandra/Floridi, Luciano (2016): "The ethics of algorithms. Mapping the debate", in: Big Data & Society 3, 1-21.

Nemitz, Paul (2018): "Constitutional democracy and technology in the age of artificial intelligence", in: Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 376.

O'Neil, Cathy (2016): Weapons of math destruction. How big data increases inequality and threatens democracy, New York: Crown.

Pernice, Ingolf (2016): "E-Government and E-Democracy. Overcoming Legitimacy Deficits in a Digital Europe", in: HIIG Discussion Paper Series.

Prasad, Mahendra (2019): "Social Choice and the Value Alignment Problem", in: Roman V. Yampolskiy (Hg.), Artificial intelligence safety and security, Boca Raton: CRC Press, pp. 291-314.

Reisman, Dillon/Schultz, Jason/Crawford, Kate/Whithattaker, Meredith (2018): Algorithmic Impact Assessments. A practical framework for public agency and accountability. AI NOW, https://ainowinstitute.org/aiareport2018.pdf.

Rousseau, Peter L. (2009): "General Purpose Technologies", in: Steven Durlauf/L. Blume (Hg.), Economic Growth, London: Palgrave Macmillan UK, pp. 74-79.

Russell, Stuart/Norvig, Peter/Kirchner, Frank (2012): Künstliche Intelligenz. Ein moderner Ansatz, München: Pearson Higher Education.

Schlaudt, Oliver (2018): Die politischen Zahlen. Über Quantifizierung im Neoliberalismus (= Klostermann Rote Reihe, Band 102), Frankfurt am Main: Vittorio Klostermann.

Schulz, Wolfgang/Dankert, Kevin (2016): "'Governance by Things' as a challenge to regulation by law", in: Internet Policy Review 5, x.

Searle, John R. (1980): "Minds, brains, and programs", in: Behavioral and brain sciences 3, pp. 417-424.

Shalev-Shwartz, Shai/Ben-David, Shai (2014): Understanding machine learning. From theory to algorithms, Cambridge: Cambridge University Press.

Stamper, Ronald (1988): "Pathologies of AI: Responsible use of artificial intelligence in professional work", in: AI & SOCIETY 2, pp. 3-16.

Sudmann, Andreas: "Szenarien des Postdigitalen:. Deep Learning als Medien Revolution", in: Andreas Sudmann/Christoph Engemann (Hg.), Machine Learning – Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz, 55-74.

Sudmann, Andreas/Engemann, Christoph (Hg.): Machine Learning – Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz.

Tendersinfo (2017): United Arab Emirates. Minister of Artificial Intelligence Minister delivers talk on AI at DPC event, http://www.tendersinfo.com/ vom 06.01.2017.

Turing (1950): "Computing Machinery and Intelligence", in: Mind A Quarterly Review of Psychology and Philosophy 59, pp. 433-460.

Tutt, Andrew (2017): "An FDA for Algorithms", in: Administrative Law Review 69, pp. 83-123.

Warren, Samuel D./Brandeis, Louis D. (1890): "The Right to Privacy", in: HLR (Harvard Law Review) 4, pp. 193.

Yeung, Karen (2018): "Algorithmic regulation: A critical interrogation. A Critical Interrogation", in: Regulation & Governance 12, pp. 505-523.

Zuiderveen Borgesius, Frederik J./Möller, Judith/Kruikemeier, Sanne/Ó Fathaigh, Ronan/Irion, Kristina/Dobber, Tom/Bodo, Balazs/Vreese, Claes de (2018): "Online Political Microtargeting: Promises and Threats for Democracy", in: Utrecht Law Review 14, pp. 82.

Rethinking the Knowledge Problem in an Era of Corporate Gigantism[1]

Frank Pasquale

A preeminent theorist of laissez-faire, Friedrich von Hayek called the "knowledge problem" an insuperable barrier to central planning. Knowledge about the price of supplies and labor, and consumers' ability and willingness to pay, is so scattered and protean that even the most knowledgeable authorities cannot access all of it. No person knows everything about how goods and services in an economy should be priced. No central decision maker can grasp the idiosyncratic preferences, values, and purchasing power of millions of individuals. That kind of knowledge, Hayek said, is *distributed*.

However, in an era of artificial intelligence and mass surveillance, the allure of central planning has reemerged—this time, in the form of massive firms. Having logged and analyzed billions of transactions, Amazon knows intimate details about all its customers and suppliers. It can carefully calibrate screen displays to herd buyers toward certain products or shopping practices, or to copy sellers with its own, cheaper, in-house offerings. Mark Zuckerberg aspires to omniscience of consumer desires, profiling nearly everyone in Facebook, Instagram, and Whatsapp, and then leveraging that data trove to track users across the web and into the real world (via mobile usage and device fingerprinting). Indeed, you don't have to use any of those apps to end up in Facebook/Instagram/Whatsapp files—profiles can be assigned to you. Google's "database of intentions" is legendary, and antitrust authorities around the world have looked with increasing alarm at its ability to squeeze out rivals from search results once it gains an interest in their lines of business. Google knows not merely what consumers are searching for, but also what other businesses are searching, buying, emailing, planning—a truly unparalleled match of data processing capacity to raw communication flows.

Nor is this logic limited to the online context. Concentration is paying dividends for the largest banks (widely assumed to be too big to fail), and major

1 This essay originally appeared as "Tech Platforms and the Knowledge Problem" in *American Affairs*, Summer 2018. It is reprinted with kind permission of *American Affairs*.

health insurers (now squeezing and expanding the medical supply chain like an accordion). Like the digital giants, these finance and insurance firms not only act as middleman, taking a cut of transactions, but also aspire to capitalize on the knowledge they've gained from monitoring customers and providers in order to supplant them and directly provide services and investment. If it succeeds, the CVS-Aetna merger betokens intense corporate consolidations that will see more vertical integration of insurers, providers, and a baroque series of middlemen (from pharmaceutical benefit managers to group purchasing organizations) into gargantuan health providers. A CVS doctor may eventually refer a patient to a CVS hospital for a CVS surgery, followed up by home health care workers employed by CVS who bring CVS pharmaceuticals—all covered by a CVS/Aetna insurance plan, which might penalize the patient for using any providers outside the CVS network. While such a panoptic firm may sound dystopian, it is a logical outgrowth of health services researchers' enthusiasm for "integrated delivery systems," which are supposed to provide "care coordination" and "wraparound services" more efficiently than America's current, fragmented health care system.

The rise of powerful intermediaries like search engines and insurers may seem like the next logical step in the development of capitalism. But a growing chorus of critics questions the size and scope of leading firms in these fields. The Institute for Local Self-Reliance highlights Amazon's manipulation of both law and contracts to accumulate unfair advantages. International antitrust authorities have taken Google down a peg, questioning the company's aggressive use of its search engine and Android operating system to promote its own services (and demote rivals). They also question why Google and Facebook have been acquiring at least two companies a month, for years. Consumer advocates complain about manipulative advertising. Finance scholars lambaste megabanks for taking advantage of the implicit subsidies that too big to fail status confers.

Can these diverse strands of protest and critique coalesce into something more durable and consistent? This essay explores possible forms to channel social and economic discontent over the next few decades. I start by giving an accounting of where we are: a hierarchical, centralized regime, where corporate power is immense, and where large national apparatuses of regulation seem to be the only entities capable of reining it in. Against this economic reality, I can discern two vital lines of politico-economic critique at present.

Populist localizers want a new era of antitrust enforcement to rein in giant firms. These Jeffersonian critics of big tech firms, megabanks, and big health care combinations are decentralizers. They believe that power is and ought to be distributed in a just society. They promote strong local authorities, who are located closer to their own citizens.

Others have promoted gigantism as inevitable or desirable, and argue that we simply need better rules to cabin abuses of corporate power. Today's Hamiltoni-

ans argue that massive stores of data are critical to the future of artificial intelligence—and thus, to productive dynamism of the economy. They focus on better regulating, rather than breaking up, leading firms.

Jeffersonians and Hamiltonians have very different long-term views on what an optimal economy looks like. In the long-run, their visions are probably irreconcilable. However, in the short run, both sets of reformers offer important lessons for policymakers grappling with the power of massive tech, finance, and health care firms. This essay explores those lessons, specifying when a Jeffersonian approach is most appropriate, and when Hamilton's heirs have the better approach.

The Jeffersonian/Hamiltonian Divide

The tech policy landscape is often bleak. Corporate-funded think tanks strive to keep reform options in a relatively narrow window of tweaks and minor changes to existing law. The curse of over-specialization in the academy also keeps many law and policy professors on a short leash. Nevertheless, there are pockets of vitality in the field. Two camps that have arisen include a decentralizing camp, which I'd call Jeffersonian, and a more centralizing, Hamiltonian tendency that is comfortable with industrial "bigness."

The Jeffersonian school has coalesced around the problem of lax antitrust enforcement in the United States, and competition promotion more generally. The Open Markets Institute, kicked out of the New American Foundation for being too hostile to Google, has led the charge. Leaders at OMI, like Matt Stoller and Barry Lynn, argue that the Federal Trade Commission (FTC) should break up Facebook, establishing Instagram and WhatsApp as competing social networks. Lina Khan, also at OMI, has written an exhaustive critique of Amazon's gigantism that is already one of the *Yale Jaw Journal*'s most downloaded articles. The emphasis on subsidiarity in Catholic Social Thought is also a font of decentralist theory, often invoked by conservatives to protect the autonomy of local authorities and civil society institutions.

The Hamiltonians include traditional centrists (like Rob Atkinson, who recently co-authored *Big is Beautiful* with Michael Lind), as well as voices on both ends of the political spectrum. Recapitulating Schumpeter's praise of monopoly as a spur to growth, Peter Thiel's *Zero to One* is a paean to monopoly power, justifying its perquisites as the just and necessary reward for dramatic innovation. On the left, Evgeny Morozov does not want to see the data stores of the likes of Google and Facebook scattered to a dozen different versions of these services. Rather, he argues, they are most likely natural monopolies: they get better and better at each task they take on when they have access to more and more pooled data from *all* the tasks they perform. The ultimate left logic here is toward fully automated luxury

communism, where massive firms use machine learning and 3-D printing to solve hunger, save the environment, and end the problem of scarcity.[2] Left centralizers also argue that problems as massive as climate change can only be solved by a Hamiltonian approach.

The Jeffersonian and Hamiltonian visions lead to very different policy recommendations in the tech space. Jeffersonians want to end Google's acquisition spree, full stop. They believe the firm has simply gotten too powerful. But even some progressive regulators might wave through Google's purchase of Waze (the traffic monitoring app), however much it strengthens Google's power over the mapping space, in hopes that the driving data may accelerate its development of self-driving cars. The price of faster progress may be the further concentration of power in Silicon Valley. To Jeffersonians, though, it is that very concentration (of power, patents, and profits) in megafirms that deters small businesses from taking risks to develop breakthrough technologies.

Facebook's dominance in social networking raises similar concerns. Privacy regulators in the US and Europe are investigating whether Facebook did enough to protect user data from third-party apps, like the ones that Cambridge Analytica and its allies used to harvest data on tens of millions of unsuspecting Facebook users. Note that Facebook itself clamped down on third party access to data it gathered in 2013, in part thanks to its worries that other firms were able to construct lesser, but still powerful, versions of its famous "social graph"—the database of intentions and connections that makes the social network so valuable to advertisers.

For Jeffersonians, the Facebook crackdown on data flows to outside developers is suspicious. It looks like the social network is trying to monopolize a data hoard that could prove essential raw materials for future start-ups. However, from a Hamiltonian perspective, securing the data trove in one massive firm looks like the responsible thing to do (as long as the firm is well-regulated). Once the data is permanently transferred from Facebook to other companies, it may be practically very hard to assure that it is not misused. Competitors (or "frenemies," in Ariel Ezrachi and Maurice Stucke's terms) cannot access data that is secure in Facebook's servers—but neither can hackers, blackmailers, or shadowy data brokers specialized in military grade psy-ops. To stop "runaway data" from creating a full-disclosure dystopia for all of us, "security feudalism" seems necessary.

Policy conflict between Jeffersonians and Hamiltonians, "small is beautiful" democratizers and centralist bureaucratizers, will heat up in coming years. To understand the role of each tendency in the digital sphere, we should consider their approaches in more detail.

---

2 Authors in this vein include Leigh Phillips and Michal Rozworski, *People's Republic of Walmart (Verso, 2019); Aaron Bastani, Fully Automated Luxury Communism* (Verso, 2019), and Peter Frase, Four Futures (2016).

The Jeffersonian Critique of Absentee Ownership

The largest, most successful firms of digital capitalism tend to serve as platforms, ranking and rating other entities rather than directly providing goods and services. This strategy enables the platform to outsource risk to vendors and consumers, while it reliably collects a cut from each transaction. Just as a financial intermediary may profit from transaction fees, regardless of whether particular investments soar or sour, the platform pockets revenues on the front end, regardless of the quality of the relationships it brokers.

This intermediary role creates numerous opportunities for platforms. For example, they police transactions and adjudicate disputes that used to be the preserve of governments. I call this powerful new role of platforms "functional sovereignty," to denote the level of power a private firm reaches when it is no longer one of many market participants, but instead, the main supervisor and organizer of actual market participants. Platforms like Amazon and Google are functionally sovereign over more and more markets, playing a quasi-governmental role as they adjudicate conflicts between consumers, marketers, content providers, and an expanding array of third and fourth parties.

Personalization is a mantra for digital strategists, who tend to assume it is a "win-win" proposition. For example, tailored search results both guard Google's users against distraction and tend to connect them to products they want. However, online markets premised on ever greater knowledge of our desires and "pain points," income level and wealth, can easily tip toward exploitation. Platforms have an interest in intensively monitoring and shaping certain digital spheres in order to maximize their profits (and, secondarily, to maintain their own reputations). However, in their ceaseless quest to annex ever more sectors into their own ecosystems, they all too often bite off more than they can chew. They tend to overestimate the power of automation to process all the demands that modern marketplaces generate.

This has led to another problem, familiar from the history of monopolistic enterprise: absentee ownership. When a massive firm buys a store thousands of miles away from its headquarters, it owns the store, and will seek profit from it, but it may only assess its performance in crude terms, with little interest in the community in which the store is embedded. The store may neglect traditional functions it served, simply in order to maximize the revenues that its absentee owner demands. A present owner, resident in the community, is more likely to run the store in a way that comports with community interests and values, since the present owner will itself experience any improvement or deterioration the store causes in its community.

Similar dynamics emerge online. Google owns the largest collection of videos online, but its YouTube subsidiary's profitability depends on calculated neglect

of many aspects of the platform. Over the past two years, a litany of critics have flayed the firm for promoting disturbing, tasteless, shocking, and abusive content, even to children. The recent Google announcement that it would promote Wikipedia links to debunk the conspiracy theory videos YouTube does so much to promote, represents yet more layers of outsourcing—from a for-profit corporation to a non-profit that in turn delegates power over content to volunteers managed by a shadowy layer of administrators.

For Jeffersonians, the answer here is obvious: there should not be one, behemoth corporation with power over so many videos. YouTube says it needs the scale to keep its offerings free; Jeffersonians respond that the ad-driven business model is just a way to undercut subscription services which could better manage their offerings. Jeffersonians also point out that it is very difficult to know the extent to which services like YouTube are actually serving users and content producers, and to what extent they exist simply to maximize ad revenue.

A Hamiltonian Perspective on New Digital Utilities

The guiding spirit of Jeffersonians is the original intent of U.S. antitrust law—that immense corporations were so capable of dominating their customers, employees, and communities, that they needed to be broken apart. Dividing a large corporation into smaller part is a "structural remedy," because it addresses fundamental ownership stakes and control in society. This populist demand to break up the largest corporations has inspired antitrust attacks on firms ranging from Standard Oil to Brown Shoe to Microsoft.

More recently, though, antitrust authorities have been more cautious about breaking up large firms. Both the Department of Justice and the Federal Trade Commission have narrowed their interest to focus almost entirely on large firms' present, price effects on consumers. So a massive firm that undercuts competitors by reducing quality is of little concern to them. Nor is the possibility that the same firm will, eventually, once it has monopolized a space, raise prices dramatically for customers (or reduce wages for workers). Instead, there is a single-minded devotion to efficiency—more, for less, faster. Free or low prices in the short run trump other considerations.

To see the practical effects of this obsession with the short-term, imagine searching for "weather" in Google, and instantly seeing its own weather forecast filling your mobile screen. Had it linked to three forecasting sites in that precious screen space, it might have directed more exposure and advertising revenue to sites with diverse interfaces, more or less information, or other variations. For example, the site WeatherSpark used to give a beautifully precise image of storms' movement over time—the perfect visual analogue to Accuweather's min-

ute-by-minute forecasts of rain or clear skies. But WeatherSpark no longer offers that service, and who knows how many other startups gave up on occupying this space. To establishment antitrust authorities, there is no ground to intervene—consumers get the basics of weather from Google's interface, and it is free. It's a short-termist outlook that omits long-run considerations in the name of a presentist scientism. In their worldview, there is no room for argument about whether better or worse alternatives do or should exist. Antirust is supposed to protect "competition, not competitors"—and a singular lack of concern for quality translates into profound lack of interest in whether current or future competitors could do a better job than a digital behemoth. But how can we know if there is competition, if there are no competitors to provide it?

In the wake of this narrowing of antitrust law, more Hamiltonian voices have called for a revival of public utility law to cabin the power of massive online firms. The utility regulators of the early 20th century did not want to see 10 different phone companies digging up the streets to provide competition in calling services. Nor did they envision localized power generation (however tempting that prospect may now be for those pursuing a distributed, renewable grid based on solar power). Instead, these regulators accepted the massiveness of telecom, power, and other firms as an inevitable aspect of modern economic rationalization. They just wanted a state (and unions) massive enough to offer countervailing forces.

For the Hamiltonians, an agency like the Federal Communications Commission provides a behavioral alternative to structural remedies. A Federal Search Commission, for example, could monitor how Google treats competing firms in search results, and force it to provide alternatives to its own services in such results.[3] European competition authorities may effectively create such an agency, if they are serious about policing Google's treatment of vertical search competitors (that is, narrow gage searching for certain types of goods or services).

Hamiltonians identify with technocratic left-liberalism. They want to deploy tools like cost-benefit analysis and advanced data analysis to calculate just when it might make sense for a service to be folded into a conglomerate, and when it makes sense to create rules that presume the independence of firms. However, there are more ideologically ambitious endorsements of industrial scale and scope. For example, Evgeny Morozov warns against efforts to split up Google or Facebook, since advances in AI may only be possible when truly massive amounts of data are consolidated. In a recent podcast, the socialists of Chapo Trap House joked that they were happy to see Amazon consolidate power. Once it takes over every business in the country, it will be easy to "cut off the head" and simply impose government control over the economy. "Free Whole Foods hot bar for every-

3 O. Bracha & F. Pasquale, *Federal Search Commission: Access, Fairness, and Accountability in the Law of Search*, 93 Cornell L. Rev. 1149 (2008).

one!" was the imagined denouement. Similarly, if all the private health insurers in the US merged, the stage would finally be set for "single payer:" the government need only take over the one insurer left standing.

Authors at *Jacobin* (including Alyssa Battistoni, Peter Frase, Christian Parenti) are also articulating a neo-Hamiltonian approach of advanced corporate capacity tempered by countervailing power of government and labor unions. Allowing centralization into large peak organizations like Germany's general trades union council and mega-manufacturers, would enable corporatist negotiations over the division of the spoils from the types of investment made possible by massive concentration of resources. Germany's largest trade union recently negotiated to reduce its members' workweek to 28 hours, while also getting a 4.3% pay raise—exactly the type of deal U.S. workers could have gotten had productivity gains since the late 1970s been widely shared, and had business and labor been similarly organized.

At its most ambitious, the Hamiltonian vision tends toward a dream of a robust universal basic income guaranteed under fully automated luxury communism. Artificial intelligence and robots mimic workers, who still are paid for the data they (or their forbears) contributed to advance AI's development. Hamiltonianism can be the economic equivalent to geoengineering—an embrace of the radically new and large-scale, arising out of the sense that inequalities and climate change are such massive problems that only rapid technological advance can solve them. Jeffersonians adhere to something like a precautionary principle, questioning whether any entity should accumulate the power necessary to, say, compare everyone's genomes, convert millions of workers' movements into patterns of behavior programmable into robotics, or maintain social credit scores on all citizens.

Reconciling Jeffersonian and Hamiltonian Perspectives

All these trends suggest new fault lines in economic thought for the 21st century. To alleviate these tensions, we should return to some seminal tensions in the neoliberal project. In the 1930s and 40s, the University of Chicago economist Henry C. Simons warned that monopolies posed a mortal threat to classical liberal ideals of free and open markets. In his *A Positive Program for Laissez Faire*, written in 1934, Simons argued that "the great enemy of democracy is monopoly, in all its forms: gigantic corporations, trade associations and other agencies for price control, trade-unions—or, in general, organization and concentration of power within functional classes." However, by the 1950s, George Stigler and Aaron Director supplanted Simons at Chicago, and offered a far more hands-off approach to antitrust law. They viewed concentrated *state* and *union* power as a far greater threat to society than concentrated corporate power. And since the former was

needed to combat the latter, they downplayed the harm that massive corporations could pose (outside a narrowly delimited category of conduct that was to become ever smaller as Chicago scholars like Robert Bork shrank the domain and force of antitrust law).

What if Chicago had followed Simons's path instead of Director's? Neoliberals might have embraced a more even-handed approach to confronting excessive power in society. Antitrust authorities would have better resisted behemoth firms' aspirations to centralize data collection and control of workers. Policymakers could have better balanced efforts to reduce state power with parallel efforts to decrease corporations' ability to work their will upon communities and workers. A 1950s era policy agenda to reduce union power looks risible in the 2010s, when union density has declined so precipitously, while corporate concentration has risen.

Jeffersonians have their own blind spot when it comes to labor. Too much of the Jeffersonian literature idealizes small-holders, advancing an idea of every-man-as-entrepreneur. But most of us are, and will be, working for someone else for most of our life. Thus Atkinson and Lind are right to argue, in *Big is Beautiful*, that small businesses should be held to many of the same labor and consumer protection laws that now only govern larger corporations. Otherwise, the wizards of franchising and platform capitalism will simply find new ways to disaggregate existing concerns into smaller units, to get around regulation. Undercapitalized and judgment-proof small businesses are the perfect business law-breakers, since they have little to lose if caught.

However, a core insight of the Jeffersonians must be respected: there really is no "one best way" to handle many products and services. The question then becomes, how to identify optimal scale and scope of enterprise in different industries. When a firm has a bona fide need for data to solve a problem (such as calculating optimal routes for a fleet of self-driving cars), that is a much better rationale for "bigness" than simply using data to rearrange commercial transactions to its own advantage. Stacy Mitchell of the Institute for Local Self-Reliance has observed that, "when third-party sellers post new products, Amazon tracks the transactions and then starts selling many of their most popular products." However much that practice increases economic productivity, it does so at an unacceptable cost of concentrating power in one firm while discouraging entrepreneurship outside it. Policymakers should protect vulnerable sellers against it.

The structural concerns of the Jeffersonians are a first line of defense against over-concentration in the economy. Competition authorities should take them seriously, particularly when there is no substantive *productive* rationale for bigness. If Amazon needs to buy equipment manufacturers to pursue vertical integration to make a better Kindle, fine—but if it is acquiring other firms simply (or mainly) to enhance its bargaining power relative to consumers or suppliers, that is not a le-

gitimate rationale for mergers. Similarly, authorities need to recognize that mergers in the name of "better service" or "cheaper inferences" about users can lead to overwhelming bargaining power for a platform vis a vis advertisers it serves—and its ability to intrude upon the privacy of its users. Those are the key reasons why the FTC should have blocked Google's purchase of DoubleClick, and Facebook's acquisition of Instagram and WhatsApp.[4]

It will be politically difficult to "unscramble the omelet" of currently dominant firms. Authorities are wary of reversing mergers and acquisitions, even when they are obviously problematic in hindsight. While Jeffersonians may keep our digital giants from getting bigger, Hamiltonians will need to monitor their current practices, and intervene when they transgress social norms. Thanks to the movement for algorithmic accountability, we know that algorithmic corporate decisionmaking is frequently deployed to arbitrage around extant anti-discrimination, due process, and media law. Agencies like the Consumer Financial Protection Bureau, the Federal Communications Commission, the Federal Trade Commission, and state attorneys general, should closely monitor platforms in order to ensure that they are actually giving their users a fair shot at access to customers, advertising, and growth. These firms no longer are mere market participants. They make markets, and need to be treated as such. Even Mark Zuckerberg recently conceded that the question is not *whether* to regulate Facebook, but *how*. Other tech CEOs should adopt a similar openness to the societal values they have shunned for so long.

Context Matters

There is ongoing struggle over what responsibilities the domination of an online space should entail. Investors demand a fantasy of monopolization: their firm not merely occupying a field, but developing "moats" against entrants, to guarantee both present returns and future growth. However, the day-to-day reality of operational budget constraints pushes the same firms toward the pathologies of absentee ownership.

Law can help resolve these tensions. Competition laws take aim at the functional sovereignty of large tech platforms, reducing the stakes of a firm's domination of a field. At the very least, antitrust authorities should have blocked Facebook's purchases of Instagram and Whatsapp, instead of letting its juggernaut of domination over communication roll up some of the few entities capable of providing alternative modes of association online. Ten, twenty, or one hundred social net-

4 For more on the advantages and disadvantages of antitrust policy here, see Frank Pasquale, Dominant Search Engines: An Essential Cultural & Political Facility, in *The Next Digital Decade* (2011).

works could eventually emerge, if competition law were properly enforced, and interoperability standards could assure smooth connections among confederations of social networks, just as AT&T, T-Mobile, and Verizon customers can all talk to one another seamlessly. If that diversity emerged, we could worry less about a few persons in Silicon Valley essentially serving as a world Supreme Court deciding which expression was appropriate for a so-called "global community," and what should be banned or obscured (in oft-secretive algorithmic manipulation).[5]

When industrial giants can't be broken up, there are still many ways to neutralize their power. Utility-style regulation mitigates the worst failures of absentee owners, as well as the caprices of the powerful. The state can require Google to carry certain content on YouTube, just as it has required cable networks to include local news. Moreover, whenever policymakers are afraid that firms like Google, Amazon, or Uber are taking too large a cut of transactions, they can take a page out of the playbook of insurance regulators, who often limit insurers to taking 15% to 20% of premiums (the rest must be spent on health care). That kind of limit recognizes the infrastructural quality of these firms' services. We would not want to live in a world where the electric company can endless jack up charges in order to take advantage of our dependence on it. Digital monopolists should face similar constraints.

Though Jeffersonian trust-busters and Hamiltonian utility regulators have very different views of political economy, each counters the untrammeled aspirations (and disappointing quotidian reality) of stalwarts of digital capitalism. They also help us understand when giant firms can help us solve the "knowledge problem" Hayek identified, and when they exacerbate it via obscurity and obfuscation.[6] If conglomeration and vertical mergers actually help solve real-world problems—of faster transport, better food, higher-quality health care, and more—authorities should let them proceed. Such industrial bigness helps us understand and control the natural world better. But states should block the mere accumulation of bargaining power and leverage. Such moves are exercises in controlling persons—a much less salubrious aim of industrial organization. Economic policy focused on productivity and inclusive prosperity will balance and do justice to important insights from both Jeffersonian and Hamiltonian critics of our increasingly sclerotic economy.

5 Kate Klonick and Thomas Kadri, "How to Make Facebook's 'Supreme Court' Work," *N.Y. Times*, Nov. 17, 2018.

6 Walter Adams and James W. Brock, *The Bigness Complex* (Stanford University Press, 2004).

Artificial Intelligence and the Democratization of Art

Jens Schröter

I want to be a machine
Andy Warhol

I. Introduction

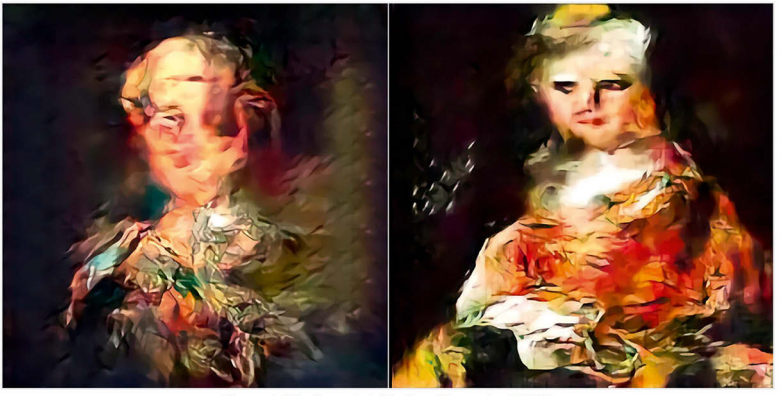
The current hype on Artificial Intelligence is exaggerated insofar, as the much-discussed procedures of machine learning adress relatively specialized problems, e.g. of image recognition or more generally of pattern discrimination. A universal, general artificial intelligence is not the goal and perhaps not even possible. Nevertheless, there is one related phenomenon that emerges in sometimes nervous discussions nowadays: It is the question whether the systems of machine learning can be 'creative'. This discussion heated up as Alpha Go was able to beat Lee Sedol in Go—especially the famous move 37 in match two on the 10th of march 2016, which seemed radically new even for experienced Go-Players was discussed: No one foresaw this move—so could it be considered as creative or not?[1] For obvious reasons, this essay cannot go into the depths of the theory of creativity or 'creative subjectivity'[2], but it will adress a recent phenomenon in the discourse on the 'new' AI, a phenomenon, in which the question of creativity is especially urgent: that is the field of art, art created by AI.

1 See the fundamental critique of the idea of machine creativity by Mersch (2019).

2 Cp. amongst many others Sternberg (1999) and Reckwitz (2012).

The AI-Art Gold Rush Is Here

An artificial-intelligence "artist" got a solo show at a Chelsea gallery. Will it reinvent art, or destroy it?



AI-generated "faceless portraits" by Ahmed Elgammal and AICAN.

Fig. 1: AI Art Goldrush is here

The AI Art Gold Rush is here is the title of a critical essay by Ian Bogost (2019). The images depicted somehow look like deformed Renaissace portraits, perhaps with a little Francis Bacon in it (Fig. 1).

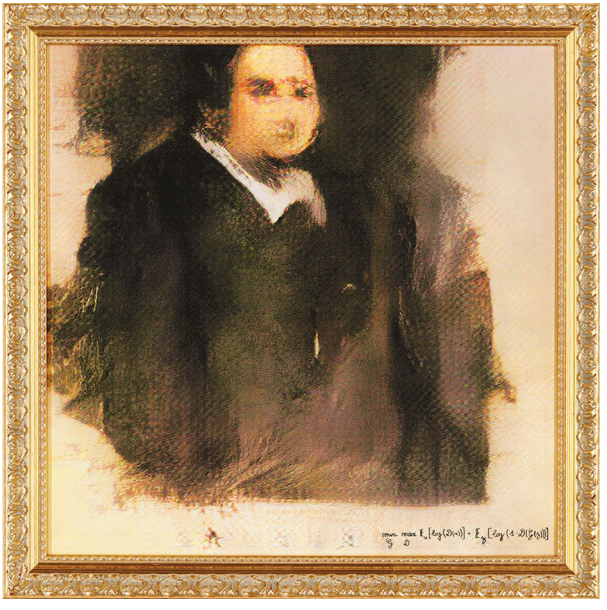


Fig. 2: Portrait of Edmond de Belamy

Especially the slightly fuzzy character of the portraits resembles the famous *Portrait of Edmond de Belamy*, auctioned for 432.500 $ at Christie's in october 2018. That this computergenerated image was sold for a comparatively high price was surprising, the joke with an algorithm as the signature spurred again a discussion on the 'creativity' of AI and the question emerged if an AI system can be an artist or an author. In addition, a discussion began whether the art collective *Obvious* who used the AI system is the 'real' author or even the programmer who developed some of the algorithms.[3] But the artwork was surprising also in another respect: It took recourse to a quite conservative genre and a quite conservative style of depiction: the blurriness seemed not only to evoke a certain 'technicity', but a conservative notion of artiness[4]—see the somewhat weird discussions on the 'blurred' style of the impressionists (cf. Payne 2007). Given that the development of painting in the twentieth century developed new forms like abstraction raises the question why one needs a conservative style of painting in order to demonstrate the creativity of AI?

And even the joke with the signature shows a certain traditional understanding of art, insofar as many artists (see the famous quote of Andy Warhol above) problematized the traditional myth of the artist: just think of surrealistic automatic writing, Cage's aleatoric processes or the doubling gestures of Elaine Sturtevant to name just a few. Perhaps *Obvious* understood the painting as purely ironical—and intended to ridicule the the 'newness' of AI by the very act of foregrounding its conservative 'taste'. Often the 'democratization' of new technologies need the adaption of conservative and established forms to be adaptable to mass markets. Therefore, an abstract work of art might be regarded as insufficient to prove AI's creativity, for the simple reason that many people still have problems in accepting abstract art forms as art at all—or as 'too easy' for real art.

Hence and this is the central argument of my short essay it is crucial to historicize this discussion on 'AI art', the implications of 'machine creativity', and therefore the (possible) *automatization of artistic work*: Sometimes, we forget that the idea of computational machines, or more specifically AI, making art adds to a rather nervous discourse on the automation of work through smart machines.[5] If machines can produce art, the assumption goes, they could mass produce art for everyone, serially, industrially. Then there would no longer be auratic works of art made by rare geniuses. And nobody would have to pay millions for artworks.

---

3 Cp. the insightful article by Sudmann (2019).

4 AI systems do not necessarily have to produce blurry images—but 'modern art' in a way has to transgress notions of realism in a way.

5 Cp. my article "Digitale Technologien und das Verschwinden der Arbeit" (2019). See also Bogost (2019): "Given the general fears about robots taking human jobs, it's understandable that some viewers would see an artificial intelligence taking over for visual artists, of all people, as a sacrificial canary."

In part II, I want to look specifically on 'information aesthetics', a discourse and practice from the 1960s, already driven by the idea to produce art *by* computers (not mainly *with* computers as tools).[6] In part III, I will discuss somehow speculative reasons why the idea to automatize artistic work (and hence simply mass produce 'art' with machines) did not seem to work back then.[7] In part IV, I'll come back to the recent 'gold rush' in 'AI art' and re-read it in the light of the discourse on information aesthetics.

II. Short Remarks on Information Aesthetics

The origin of information aesthetics is the attempt to formally determine the 'measure' of aesthetics. In 1933, David Birkhoff formulated an equation (Fig. 3) in which the variable O denotes the measure of the order of a given work and the variable C the 'complexity'. M is the degree of how aesthetic an artwork is.

$$M = O/C$$

Fig. 3: Birkhoff-Equation

According to it, the more ordered and the less complex a work of art is, the more aesthetic it would be. Apart from the fact that it is difficult to understand exactly how to determine the degree of order and complexity in a specific case, this attempt to express the 'aesthetic quality' of art in an equation seems strange to us today. Nonetheless, especially in the 1960s there were several attempts to formally understand art and its aesthetic criteria and consequently to produce it synthetically with computers (although the computers were slow, the output possibillities limited and computer technologies were only available in research institutions and large companies).

---

6 There are even more precursors, for example the computer cluster Iamus, who is composing music and even released an album, (see: http://melomics.uma.es/). The slogan is, not surprising, "music for everybody, everything", promising a democratization of art. Another important example would be *Aaron*, which is a software system that in collaboration with its inventor Harold Cohen produces paintings (see: https://www.computerhistory.org/atchm/harold-cohen-and-aaron-a-40-year-collaboration/).

7 It's quite problematic that an otherwise excellent collection of essays on ‚Computers and Creativity' (McCormack/d'Iverno 2012) does not include a single contribution that tries ot relate the central question of the volume to sociological questions concerning the art system. 'Art History' is only mentioned a few times in the outstanding contribution by Frieder Nake—except of Nake's paper most contributions in this volume de-socialize and de-historicize the question of the possible creativity of machines.

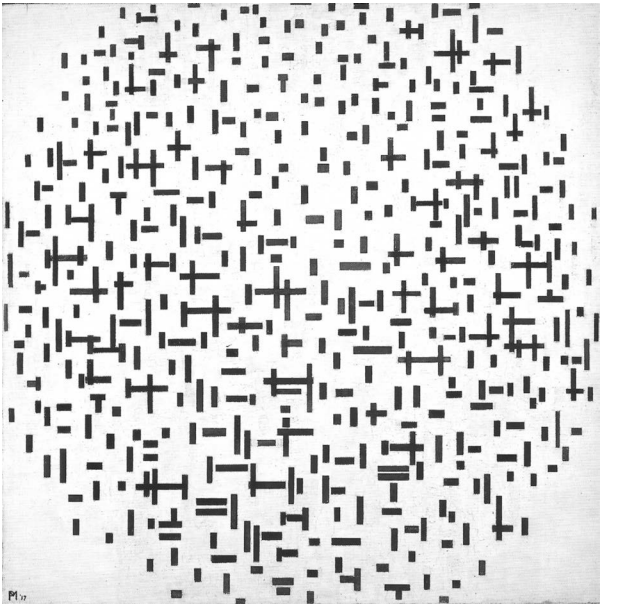


Fig. 4: Piet Mondrian, Composition With Lines (1917)

In 1967, Michael Noll describes in his essay "The Computer as a Creative Medium" an information aesthetic experiment with a painting by Mondrian (Fig. 4).[8] Noll writes:

> [An] experiment was performed using Piet Mondrian's "Composition With Lines" (1917) and a computer-generated picture composed of pseudorandom elements but similar in overall composition to the Mondrian painting. Although Mondrian apparently placed the vertical and horizontal bars in his painting in a careful and orderly manner, the bars in the computer-generated picture were placed according to a pseudorandom number generator with statistics chosen to approximate the bar density, lengths, and widths in the Mondrian painting. Xerographic copies of the two pictures were presented, side by side, to 100 subjects with educations ranging from high school to postdoctoral; the subjects represented a reasonably good sampling of the population at a large scientific research laboratory. They were asked which picture they preferred and also which picture of the pair they thought was produced by Mondrian. Fifty-nine percent of the subjects preferred the computer-generated picture; only 28 percent were able to identify correctly the picture produced by Mondrian. In general, these people seemed to associate

8 It is by the way one of the earliest texts in which the computer is described as a medium—the computer's becoming a medium thus goes back to questions of art and aesthetics.

> the randomness of the computer-generated picture with human creativity whereas the orderly bar placement of the Mondrian painting seemed to them machine-like. This finding does not, of course, detract from Mondrian's artistic abilities. His painting was, after all, the inspiration for the algorithms used to produce the computer-generated picture, and since computers were nonexistent 50 years ago, Mondrian could not have had a computer at his disposal. (1967: 92)

Fig. 5: Simulated Mondrian by Michael Noll

Noll thus simulates a Mondrian (Fig. 5) on the basis of a statistical distribution which is supposed to describe the arrangement of the lines in Mondrian's work. And since this distribution looks more 'disordered', a group of observers, presumably not to be regarded as representative, identifies the simulated Mondrian as the real one, while the real one appears too regular and mechanical.[9] Does this preference show that the disorderly picture is understood to have a higher aesthetic quality? That would contradict Birkhoff, but it does remind us in an uncanny manner of the dis-

9 A similar discourse can be found in recent AI art: "According to Elgammal, ordinary observers can't tell the difference between an AI-generated image and a 'normal' one in the context of a gallery or an art fair" (Bogost 2019).

torted, blurred, and 'arty' qualities of recent AI-generated portraits. Or is it because the disorder is simply understood as a reference to human authorship? It could also reveal that the artiness of a given picture does not depend on perceptual features alone… It should be noted that Noll tries to find formal, algorithmic rules, allowing the production of a work of art that can be *identified* as a work of art. Noll is not the only one following this approach. Frieder Nake, one of the most important theorists and practitioners of information aesthetics, also tried to trace the pattern of image production in Paul Klee's work and thus produce a computer-generated Klee (Fig. 6).

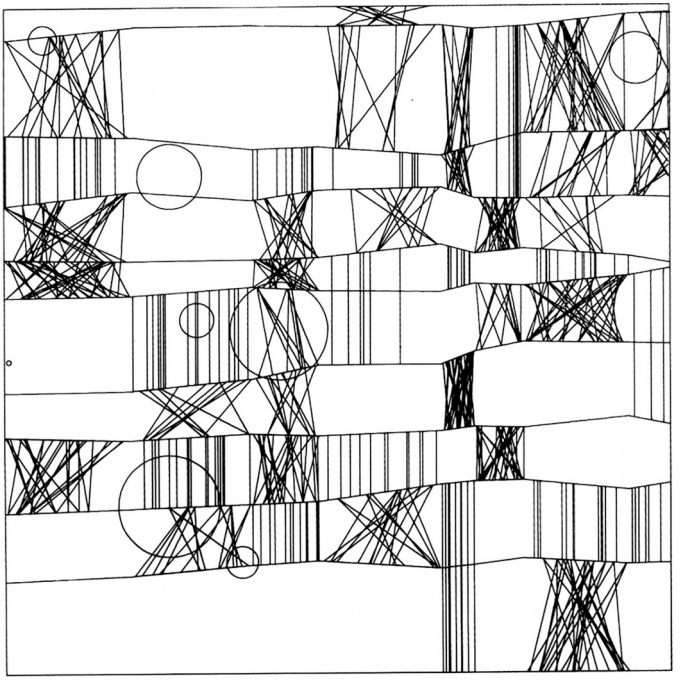


Fig. 6: Simulated Klee by Frieder Nake

Admittedly, these examples are taken out of context and an appropriate discussion of information aesthetics would have to consider at least the positions of Max Bense, Rul Günzenhäuser, and Abraham Moles, which can't be addressed in this essay. In any case, from today's point of view, these attempts seem rather strange, because they remove the works of art from their historical context and reduce them to abstract structures that can be formalized—similar to recent AI-produced art. It must be therefore emphasized that the aesthetic strategies of Mondrian and Klee have reacted to certain historical questions, not the least of them the history of art itself. Information aesthetics and AI art seem to understand artworks as ahistorical, formal structures—although they could also be seen as responses to

a certain historical, in this case computational, context of art as such (cf. to the vexed relation between form and history amongst others Buchloh 2015). This also seems to reappear in recent AI art: "That might be an inevitability of AI art: Wide swaths of art-historical context are abstracted into general, visual patterns." (Bogost 2019)[10]

However, these attempts are relatively characteristic of a larger development, namely the attempts to formalize cognitive labor and, if possible, to transfer it partially or entirely to machines, for which computers as symbol-processing machines are suitable. [11] Early texts from the domain of Computer Science, such as Douglas Engelbart's "Program on Human Effectiveness" or J. C. R. Licklider's "Man-Machine Symbiosis" from the 1960s, are programmatic for this. (cf. Licklider 1960: 4-11; Engelbart 1991: 235-244). Noll, Nake and others try to make aesthetic work formalizable and, in principle, executable by machines. Their attempts can be understood, whether intentionally or not, as contributions to the automation of aesthetic work.

In other forms of commodity production, these processes normally contributed to the cheapening of commodities—and therefore to 'democratization' in the sense of goods becoming more affordable for more people. That was obviously not the goal of information aesthetics, since—for instance—Nake operated as the author of the work (and not Klee); and if he would have called his work a Klee, this could have led to serious legal problems. But even if this democratization of art would have been the goal of information aesthetics—to start an industrial, cheap mass production of, say, Mondrians and Klees—this would not have worked in the art system: Art is not only a question of formal structures and strategies, it is also a question of historical and especially social places and roles. We will now turn to that.

III. Art, knowledge and work

Apparently, art it is not directly threatened by computerization. Art does not appear in the highly discussed Oxford research report, which started a nervous discussion on the disappearance of work: the only activity that resembles artistic practice is that of the 'art director', who gets off quite lightly with a 95th place on the computerizability probability list. (cf. Frey/Osborne 2013: 59) Artists can-

10 The same is true for Schmidhubers contribution to McCormack/d'Iverno (2012). It's again an approach to formalize aesthetic value of an object—without really posing the question if this value is not derived from the relative historical position of the artefact and not (only) of its formal, internal structure.

11 Cp. on the history of Automation Noble (1984).

not become the object of rationalization and their works will continue to be called 'works'. Artistic work seems to be a type of work that cannot or should not be formalized, algorithmized, and consequently neither reproduced mechanically without further ado. This suggests the suspicion that artistic work is not really work, but another form of activity or—at least—another form of work. But why?

At least at first glance, the art market looks exactly like any other market: artists are well aware that they have to earn money with their work. Some succeed to an almost inconceivable degree, but most have serious difficulties competing in the art market. Artists are also subject to the fact that work is "a necessity of refinancing their expenses" (Luhmann 1994: 191). Work, money for one's own work, market, competition, rich versus poor—this first glimpse suggests that the art business does not differ in any way from other forms of production. It is not a realm of freedom, but only a kind of service or consumer goods industry that serves a special market.

Therefore, we can find cases where work is technologically made superfluos in the art business. If one looks at a large studio, such as Studio Olafur Eliasson, one could observe that the introduction of new computer-assisted technologies directly leads to the disappearance of work. A series of jobs, e.g. website maintenance, management, public relations, logistics, up to the people that clean the studio, may be replaced. Inasmuch as Studio Olafur Eliasson also operates under capitalist conditions, it is likely to save costs by rationalizing the way work is done. However, this aspect is external; it does adress the problem that information and generative aesthetics have pointed to, mainly if the work of making art *itself* can be rationalized. Although Olafur Eliasson could not realize any of his elaborate projects without his team, the disappearance of the teamwork in the black box of the author's name does not seem to affect art 'in itself'. And insofar the teamwork is blackboxed it may change its composition without changing anything in the 'artiness'of Eliasson's art.

This finally brings us to the core of the problem raised by information aesthetics. Obviously, the idea of art without the intervention of a human author or causer—and even if its role consists precisely in demonstratively withdrawing—does not seem plausible to us. We do not see art appearing in nature.[12] The question implied in information aesthetics and AI art is different from the observations in Walter Benjamin's famous essay *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit*, where everything revolves around the question of the technical reproduction of the *product, the work of art*. But in information aesthetics and AI art the question of the *reproducibility of the work that produces the work of art* is central. The fact that the technical reproduction of this work does not seem possible

12 Although art was historically sometimes understood as being close to nature ('Kunstschönes' vs. 'Naturschönes'), cp. on this Kant (1914 [1790]).

does not have to be attributed—this would be a very traditional answer—to the 'genius', which is ultimately of divine origin and thus per se untechnical. Luhmann remarks: "The artist's genius is primarily his body" (2000: 38). One could therefore simply say that the separation of knowledge from the working body, characteristic of the progression of capitalism and perhaps first discovered by Marx, does not or cannot take place in art. But why?

One possible reason would be that the work of art—despite the attempts of the information aesthetics (at least according to Birkhoff) to formalize precisely the complexity—is *too complex* and thus the work that produces it cannot be sufficiently understood. Thus, it is noticeable that Noll and Nake focus on a certain type of painting, which is determined by the extensive recourse to basic geometric forms. Such forms seem simple enough to suggest their formalization, while other more complex ones would elude them. Obviously, technology progressed: Nowadays geometric patterns do not suffice to demonstrate the creativity of computing systems, it has to be (albeit blurry) portraits. Portraiture is historically connected to a history of 'genius'. As was remarked in the beginning, such portraits are hardly the status quo of art today[13]—using machines to create figurative art is less about AIs critically reflecting on contemporary art (at least when Nake and Noll made their works, abstract painting was quite central), it is more about the new powers of computer graphics.

But coming back to the vexed question of auctoriality: Even with geometric forms, it is true that we cannot quite imagine their existence as art without an artist. For even if an artist—like Nake, for example—were to define himself precisely by delegating all work to machines, we would still call the result 'a work by Nake', which is similar to the finding already mentioned that the production within a studio with a division of labour is 'black boxed' under an author's name. A similar process seems to occur with the debate regarding the *Portrait of Edmond de Belamy* on who is the author in the last instance.

The work of art is the result of a work in which body and knowledge, i.e. the knowledge of *how this specific work is to be produced*, cannot be separated—and that means that the function of the 'author' is central. (cf. Graw 2012: 43-45) At least, this is the ideological figure that has historically emerged as characteristic of the art system. Therefore, works of art must not be the connection between knowledge and *a false body*—this would be what we call forgery. Even if, for example, one were to take a work by Donald Judd that according to Sebastian Egenhofer (2008: 214) is "dissolved in the anonymity of the industrial dispositive" , it would still be pointless if another person or simply a company, based on the knowledge of how it is made, were to produce the same object again, as it is also done in principle in the industrial production of reproductions—it would not be possible

13 Perhaps with the exception of some types of 'post-modernist' neo-figurative painting.

to recognise this reproduction as a work of art (of 'Donald Judd'). Or let us take another example: Elaine Sturtevant borrowed the screen printing matrices from Warhol for the *Flowers* and printed the *Flowers* again, in 1991 even made an entire exhibition with *Warhol Flowers*—and Warhol, referring to the production process of the *Flowers, is frequently quoted for saying*: "I don't know. Ask Elaine." (Quoted in: Arning 1989: 44) Nevertheless, Sturtevant's appropriation of Warhol's knowledge is not a rationalization of Warhol's work in the sense that Sturtevant now simply makes 'cheaper Warhols', but she rather makes 'Sturtevants'. Diedrich Diederichsen points out that the work of assigning relevance by curators, critics, audiences, etc. also belongs to the work that creates the artwork and its market value. (cf. Diederichsen 2012: 99) The 'distributed' character of this work makes it impossible to rationalize—since it continuously accompanies the work, i.e. never ends, and can also take unpredictable turns in the future. 'Sturtevants' can become more important and more expensive than 'Warhols' in the future. It could even be argued, that Sturtevant by her appropriation makes the author-function of Warhol visible in the first place.

Of course, you can try to save production costs, but it does not make sense to offer Warhol's *Flowers* cheaper, because only *Flowers* from Warhol's Factory are accepted as originals, which of course does not exclude the possibility of producing inexpensive reproductions of *Flowers* as posters (which do not count as work of art, but its reproduction). Warhol's life ended in 1987 and that stopped the production of original 'Warhols'—and that's a necessity: In the long run, the mortality of artists makes artworks scarce and that's why they have market-value (cf. critically on the notion of scarcity Panayotakis 2012). In capitalism, no one can have an interest in a 'democratic' production of artworks.

Obviously, the crucial difference is that in art itself, reproductions, such as Sherrie Levine's re-photographs or Sturtevant's repetitions, are always originals (which, through repetition, exhibit the discourse of the original). In contrast, in 'conventional' commodity production there are no originals,[14] but only series of reproductions, e.g. of *Flowers* posters, which all refer to the original pieces of art but are all serially of equal rank. And once again: to be original means to be connected with the body of the artist. *This work cannot be detached from the body, which is why it cannot be formalized and rationalized.* The attempts of information aesthetics may seem pointless because here a rationalization dispositive from industrial production is transferred to an area that blocks it. Or as Adorno puts it:

14 A difficult case seems to be product counterfeiting, insofar as the term already implies an 'original product'—but even original products exist as a series of identical products. Product counterfeiting means that a product wrongly claims to belong to that series. This differs from the relationship between original and copy in art.

> On the other hand, however, whenever autonomous art has seriously set out to absorb industrial processes, they have remained external to it. […] The radical industrialization of art, its undiminished adaptation to the achieved technical standards, collides with what in art resists integration. (1997 [1970]): 217).

But to be clear about this: That does not mean that art in itself is an utopian realm of freedom, freed from the restrictions and rules of commodity production. It just means that it is a field of commodity production with different, historical, contingent rules and that's why art is in a way economically exceptional (cf. Beech 2015)

IV. Conclusion

It seems that the question whether machines can be creative or not can not be answered ontologically. What 'creative' means seems to be too historically contingent and malleable. Think of the statements why machine-learning is not creative: Bogost (2019) insists that "any machine-learning technique has to base its work on a specific training set." That's true—but doesn't that also apply to human artists who have to train their perceptual and e.g. painterly skills? Or, again Bogost (ibid.): "A neural net couldn't infer anything about the particular symbolic trappings of the Renaissance or antiquity—unless it was taught to, and that wouldn't happen just by showing it lots of portraits." Is that not true for humans too? And moreover: Is 'creativity' not always distributed between human and non-human actors? Don't human artists often say that their artwork 'answers' while they are in the painting process? Can one be an artist without any kind of non-human materials and mediators that are not only transparent tools for a pre-given 'vision' of a work (cp. Hensel/Schröter 2012)? And couldn't we even imagine advanced AIs, perhaps in the shape of humanoid robots that could be artists (cp. Kjøsen 2012 for a similar argument relating to the labour theory of value)? Could not the works they produce then be tied in the same way to their 'bodily' presence as is the case with human artists? Think of virtual popstars like Miku Hatsune[15], which can operate as a kind of enunciator, having a kind of 'signature' (here: her voice)—but of course the mortality of the artist as a kind of natural scarcity that limits the work is not given in such a case (on the notion of a virtual star, see Schröter 2000).

Although at the moment the idea of artworks produced by machines seems ahistorical and absurd, because artworks have a place in history and have to be tied to the body of the artist and therefore be scarce and so on, it may nevertheless be that in a far future things could change: Perhaps a democratization of art needs a different social context—as does the democratization of AI, ripping it out

15 https://en.wikipedia.org/wiki/Hatsune_Miku.

of the hands of big monopolies. That's why it seems so appropriate that the much discussed AI-painting is a portrait of Edward Bellamy—since Bellamy wrote in 1888 the famous novel *Looking Backward 2000-1887*. The AI looked back, so to speak, to a historical mode of painting in a historical style and with somewhat dated gestures (the signature), but this might also be a metaphor of looking back from a possible future, in which AI can be an artist. Bellamys novel centrally is about a very different economy of the future, a kind of post-capitalism, that might indeed be the precondition for the democratization of AI as of art. And in the novel several futuristic media are mentioned, which nowadays seem absurd of course—but which can also be read as metaphors of a future mediality, in which even 'creativity' might at least partially be automatized (for an opposing view, see Kelly 2019).

Bibliography

Adorno, Theodor W. (1997 [1970]): Aesthetic theory. Newly translated, edited, and with a translator's introduction by Robert Hullot-Kentor, London: Athlone Press.

Arning, Bill (1989): "Sturtevant." In: Journal of Contemporary Art Vol. 2, No. 2, pp. 39-50.

Beech, Dave (2015): Art and Value. Arts Economic Exceptionalism in Classical, Neoclassical and Marxist Economics, Leiden and Boston: Brill.

Bogost, Ian (2019): "The Al-Art Gold Rush Is Here. An artificial intelligence 'artist' got a solo show at a Chelsea gallery. Will it reinvent art or destroy it?" In: The Atlantic March 3 (https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/).

Buchloh, Benjamin H.D. (2015): "Formalism and Historicity" In: Ibid., Formalism and Historicity. Models and Methods in Twentieth Century Art, MA/London: Cambridge, pp. 1-87.

Diederichsen, Diedrich (2012): "Zeit, Objekt, Ware." In: Texte zur Kunst No. 88, pp. 95-101.

Egenhofer, Sebastian (2008): Abstraktion Kapitalismus Subjektivität. Die Wahrheitsfunktion des Werks in der Moderne, München: Fink.

Engelbart, Douglas C. (1991): "Letter to Vannevar Bush and Program On Human Effectiveness" In: James M. Nyce/Paul Kahn/et al. (eds.), From Memex to Hypertext. Vannevar Bush and the Mind's Machine, Boston/MA: Academic Press, pp. 235-244.

Frey, C. B./Osborne, M. A. (2013): "The Future of Employment. How Susceptible are Jobs to Computerisation. Resource Document", Oxford Martin School, University of Oxford. (http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf).

Graw, Isabelle (2012): "Der Wert der Ware Kunst. Zwölf Thesen zu menschlicher Arbeit, mimetischem Begehren und Lebendigkeit." In: Texte zur Kunst No. 88, pp. 43-45.

Hensel, T./, & Schröter, Jens. (2012): "Die Akteur-Netzwerk-Theorie als Herausforderung der Kunstwissenschaft." In: Ibid. (eds.), Die Akteur-Netzwerk-Theorie als Herausforderung der Kunstwissettschaft. Schwerpunktherausgeberschaft der Zeitschrift fur Ästhetik und Allgemeine Kunstwissenschaft No. 57, pp. 5-18.

Kant, Immanuel (1914 [1790]): Critique of Judgement. Translated with introduction and notes by J. H. Bernard, second edition, London: MacMillan.

Kelly, Sean Dorrance (2019): "A philosopher argues that an AI can't be an artist. Creativity is, and always will be, a human endeavor.", MIT Technology Review, 21.2.2019 (https://www.technologyreview.com/s/612913/a-philosopher-argues-that-an-ai-can-never-be-an-artist/).

Kjøsen, Atle (2012): "Do Androids Dream of Surplus Value?" (https://www.academia.edu/2455476/Do_Androids_Dream_of_Surplus_Value).

Licklider, J.C.R. (1960): "Man-Computer Symbiosis" In: IRE Transactions on Human Factors in Electronics Vol. HFE-1, No. 1, pp. 4-11.

Luhmann, Niklas (1994): "Kapitalismus und Utopie." In: Merkur Vol. 48, Issue 540, pp. 189-198.

Luhmann, Niklas (2000): Art as a Social System, Stanford: Stanford University Press.

McCormack, Jon/d'Iverno, Mark (2012) (Ed.): Computers and Creativity, Heidelberg: Springer.

Mersch, Dieter (2019): "Kreativität und Künstliche Intelligenz. Einige Bemerkungen zu einer Kritik algorithmischer Rationalität." In: Zeitschrift für Medienwissenschaft 21, forthcoming.

Noble, David F. (1984): Forces of Production. A Social History of Industrial Automation, New York: Knopf.

Noll, Michael (1967): "The Computer as a Creative Medium." In: IEEE Spectrum Vol. 4, No. 10, pp. 89-95.

Panayotakis, Costas (2012): "Theorizing Scarcity. Neoclassical Economics and its Critics" In: Review of Radical Political Economics Vol. 45, No. 2, pp. 183-200.

Payne, Stewart (2007): "Cataracts the key to Monet's blurry style." In: The Telegraph May 16 (https://www.telegraph.co.uk/news/uknews/1551703/Cataracts-the-key-to-Monets-blurry-style.html).

Reckwitz, Andreas (2012): Die Erfindung der Kreativität. Zum Prozess gesellschaftlicher Ästhetisierung, Berlin: Suhrkamp.

Schröter, Jens (2000): "Lara Croft. Funktionen eines virtuellen Stars." In: Ulrike Bergermann, Hartmut Winkler (eds.), TV-Trash. The TV Show I Love to Hate, Marburg: Schüren, pp. 123-144.

Schröter, Jens (2019): "Digitale Technologien und das Verschwinden der Arbeit." In: Thomas Bächle, Caja Thimm (eds.), Die Maschine: Freund oder Feind?, Wiesbaden: Springer VS, pp. 183-210.

Sternberg (1999) (Ed.): Handbook of Creativity, Cambridge: Cambridge University Press.

Sudmann, Andreas (2019): "KI-Phantasien: Kommt jetzt der Terminator mit Pinsel?" In: CCB Magazine February 8 (https://www.creative-city-berlin.de/de/ccb-magazin/2019/2/8/andreas-sudmann-ki-forschung/).

"That is a 1984 Orwellian future at our doorstep, right?" Natural Language Processing, Artificial Neural Networks and the Politics of (Democratizing) AI

Andreas Sudmann in conversation with Alexander Waibel, professor for Computer Science at the Karlsruhe Institute of Technology and also professor at the School of Computer Science at Carnegie Mellon University.

Andreas Sudmann: Alex, you are one of the pioneers in the area of Artificial Neural Networks (ANN) and Natural Language Processing (NLP). What was your initial motivation to enter this field of research?

Alexander Waibel: I have been working in academia for around forty years. Born in Germany and having German parents, I went to study at MIT and later to Carnegie Mellon, the leading universities in computer science and AI. And it was at those institutions that I developed the main thrust and inspiration for all my work, which is the question of how we learn and communicate as human beings and how we can build technology to help improve human communication. Back then, in the 1970s, when I went to university, people had already been thinking about AI, given that the first definitions of the field were proposed in the 1950s. In fact, Nobel laureate Herbert Simon, who was part of my thesis committee, had participated in the famous Dartmouth conference on building intelligent machines in 1956, where he and other researchers defined this early vision of AI.

In those days, everybody was attempting to build intelligent machines by search algorithms, rules, and logic formulas. But for me as a student, this seemed a bit like "des Kaisers neue Kleider" (the emperor's new clothes). I was listening to these famous people talking about a problem that to me would never be solvable with a rule-based approach to AI. It was intuitively clear to me that the amount of knowledge and facts that we learn in a lifetime is just so enormous that programming it all into rules would be impossible. And worse, they would have to be changed all the time, because the world around us is changing all the time. In fact, this *is* totally impossible, and so it was an early concern for me to say from the start: We will never achieve such goals unless we develop learning machines that can acquire such knowledge by themselves.

Aside from this fundamental scientific quest, though, another dimension always mattered to me as a scientist: Even though this might now sound like a cliché – it was a continuing desire to make contributions to society and to make the world a better place, as opposed to just following my own personal pursuits or interests. As a researcher, I am not so much simply curiosity-driven, but driven by practical goals. Practical goals provide a way to evaluate progress and can impact society in a positive way, once we are successful. And among them perhaps one of the most consistent goals for my work as a foreigner who grew up with 5 languages, was to build machines that can help us humans to translate between languages, by text or by spoken language. Throughout history, there have been many attempts to use machines for translating texts, which is hard enough in its own right. But when you try to connect people across language barriers, you also have to translate *spoken* language. In the 70's this seemed like a preposterous goal, and indeed, it seemed unsolvable, as speech added a whole new dimension of complexity and complication to the problem due to the fact that turning speech into text (before translating it) was an unsolved AI problem in itself. We did not know how to recognize speech and worse people never speak clean text… they make mistakes when they speak, they stutter, hesitate and correct themselves during speaking. And how would you then combine it with the other hard AI problem of translation? And all of that combined was so hard that it was unthinkable to realize in the early days of AI. But for me the dream was born and with youthful naivety and optimism we went for it. Needless to say, it was and remains a hard problem, a problem that we are still working to this very day. But despite the obstacles, challenges and delays along the way, we were able to see the fruits of our efforts. In retrospect, it is quite a privilege, actually, to be living in the *one* generation of humankind that sees language barriers disappear and to have had the opportunity to be working on the technologies made it possible.

The key to success scientifically was due to progress in machine learning methods combined with the explosive growth in available computing power and data that supports them. But for the vision to become reality also meant that academic progress had to be transferred to societal deployment. To do so, we started several companies that specialized on building wearable speech and language technology and eventually mobile speech translators. One of them, Jibbigo, built and sold the first ever Dialog translator on a phone. It was sold via the App Store and helped Tourists and Healthcare workers to communicate. The company was later acquired by Facebook, and we continued working on even more advanced deployments. For example, we are now developing new interpreting tools that help migrants in Germany to communicate with doctors if they cannot speak the language. In a University setting, we have installed automatic simultaneous interpretations services at KIT, so that foreign students can study in Germany and follow a German lecture by way of simultaneous interpretation during the lecture.

My team in Karlsruhe and I have also performed early experiments at the European Parliament to see if such a technology can be of assistance in this most challenging language environment. So it's really ultimately not only about translation alone but about how we can build technology that can bridge across barriers, that can bring the world together and make people understand each other better. And in order to master this hard problem, you really have to build machines that can learn effectively at multiple different levels.

And you have already worked out the necessary fundamentals in the 1980s and 1990s, especially with your research on so-called TDNN models. Perhaps you can tell us a little about how you came up with this particular approach and explain how it works?
Back while I was writing my PhD thesis at Carnegie Mellon I became fascinated with the idea of building learning algorithms that would mimic more closely the massively parallel, holistic learning that we perform as humans in the brain. I discovered that researchers in the 1950's had already proposed so-called "perceptrons", which did learn, but could only solve very simple classification tasks. Still, the fact that one could actually learn those functions was not only exciting, but seemed to be directly applicable to the fuzzy and ambiguous language and perception problems that I was working on. Again, this was a time when people believed they could solve speech recognition and language translation primarily by rules, a belief that seemed preposterous to me, given the enormity of facts and details that would have to be assembled. Nevertheless, simple perceptrons and similar methods also had severe limitations for speech recognition, because one could only train a single neuron at a time. The whole magic of the brain, by contrast is that it does not train single neurons but it trains entire *networks* of neurons, and that an *ensemble* of neurons can do much more powerful tasks. But how would we train an entire network?

It was just during that time that fortuitously a young assistant professor by the name of Geoffrey Hinton came to Carnegie Mellon, and started working on something called Boltzmann machines. While he was there, we had many wonderful discussions, and he introduced me to something they had been tinkering with at USC San Diego, an algorithm called backpropagation. It was much closer to what I was looking for and I immediately jumped on it. Backpropagation was a simple algorithm, an extension of the simple perceptron – except that this algorithm would now optimize the whole *network* of perceptrons and make sure that it was functioning in an optimal way. If you tell the entire network what it's supposed to do, it can in fact adjust each internal neurons in such a way that each of them will try to contribute to what is best for the whole ensemble of neurons. This seemed like a big step in the right direction, a big improvement toward classifying patterns, but for speech and language this was not enough. Because in most real world problems, recognizing patterns is not the only problem, but finding the

pattern that is to be classified in the first place. This always meant that one would have to segment a signal first to find the interesting patterns (sounds, images) before they could be classified. In speech, one would have to cut speech in such a way that you identify the beginning and the end of a particular phoneme, and then once you have that, you can try to apply a neural network for classifying these sounds and assemble them into a speech sequence. However, this meant compounding multiple separate hard problems. And the hard learn lesson in speech was that this is deadly as each of them makes mistakes. Therefore, it became clear to me that we needed a neural network that was not only a wonderful classifier but that would also recognize patterns independent of position, a property we would call "shift invariance". So what does that mean? It means that you are building a neural network that you do not just apply to a particular pattern, but that you move all neurons over a range of input, and let them essentially scan that input until it lights up whenever it finds a useful, helpful pattern. Networks of such units could thus learn to assemble all useful evident independent of small shifts in the signal. Such shift-invariance is, of course, necessary for speech, because speech flows by and changes all the time, but as it turns out it is also necessary for many other problems in AI, including images, music, games, language, and many more.

In all of these situations, your first challenge is to know *where* the useful patterns are before you can classify them correctly. Hence, classifying things by detecting them in a shift-invariant fashion was the key problem that we needed to solve. With that goal in mind, I then went to Japan as a post-doc, where I had access to some of the most powerful super-computers at that time. And with this computing power, I had the chance to develop a new model which then became known as the time-delay neural network (TDNN). It was still a multilayered ("deep") neural network, but it was now trained specifically for shift invariant classification. And as it turned out it worked fantastically well; it worked better than all other methods that existed back then.

So did this new TDNN model then replace other methods?

Sadly, we still did not have the necessary computing power to build networks that were large enough. Back then in Japan, we used the biggest supercomputers available, and compared them with other statistical or rule-based methods over benchmark data – and we found them to be much, much better. But when we tried to build larger networks and practical speech recognition systems, we still ran up against computational limits and had to make many compromises that hurt performance, and so other researchers could use simpler methods to gradually catch up, and get similar or even slightly better performances then we did. As a consequence – and this was in the late 1990s, early 2000s – people lost interest in neural networks and simply used other statistical methods.

Ten years passed and few people continued to work on neural networks, until around 2008, when – rather by coincidence – various people in the US actually tried these old neural networks again, but with the help of much more computing power and with much larger amounts of learning data that is now available over the internet. And, as it turned out, these neural network methods that had already been developed in the 1980s suddenly worked amazingly better than any other approach in the field of AI. And they did not just work a *little* better, but in fact they worked like 30 percent better. In our area, you know, entire PhD theses are written when progress of half a percent is made; so doing something that is 30 percent better is simply revolutionary. As a result, the entire community switched to neural approaches within two years. The other thing that came as a surprise is what happens when you add more layers in the network. In the 1980s, we had one or two so-called hidden layers and that was all we could compute. But now, with all this new computing power, we can do three, four, five, and more layers. No one expected that this would continue to improve performance, but it did. Today, we have networks used for speech recognition in our laboratory with 40, 80 or even hundreds of layers. And the exciting neural models that worked so well 20 years ago work even better today, too. TDNN's went on and got applied to image processing, games, speech, and other problems and became known by the more generic name: "Convolutional Neural Nets". They can now be found at the heart of most modern AI engines.

In terms of having access to powerful hardware and large amounts of data, it was certainly helpful that you worked for Facebook for some time.

Right, I was with Facebook for two years as a director, but of course I also have many friends working at Google, Amazon and Microsoft. Many of our students are now with Google, Amazon, Microsoft and so on. And many of them graduated from our labs. The massive amounts of data that these companies control is of course a treasure trove for learning programs. In those large Internet companies, they train huge neural networks over huge amounts of data using huge amounts of computing power, and the performance gains still grow. And that's surprising and impressive. But if you ask me what's the new breakthrough in AI today as opposed to 20 years ago, I would have to tell you: not that much. They are again very much the same network techniques and training algorithms as we were working on in the 1980s, except that we now use orders of magnitude more data and more computing power, and they actually work much, much better than we ever imagined.

So far, we have mainly talked about the technological aspects of speech recognition, machine translation, and ANN. Perhaps we can now talk about the political dimension of these models and applications of AI. What would you consider to be the most relevant political aspects concerning the field of natural language processing in general and speech recognition and machine translation in particular?

There is of course much to say about this. One important aspect would be the politics of research funding that affects us directly in terms of how we are doing science. Scientific support depends on political factors, and sometimes they work, sometimes they don't. And there are really fascinating differences between the US and Germany, or between countries in Europe and Asia, because each of these countries or cultures approaches scientific support differently and therefore has specific strengths and weaknesses. So I am politically active at that level to introduce and improve better mechanisms for research support in Europe. The other political dimension, though, is what we do with our research. As I have mentioned before, in my view of the world, I like to do projects with which I try to improve some aspect of society. And as I said before, research always meant for me to make people understand each other better. If with our research we can build machines that allow us to communicate better, then this means having fewer misunderstandings.

Throughout my career, I have founded several companies, and one of them was for building a handheld speech translator on a phone. It was the first mobile speech translation system on a phone ever. We launched that in 2009. The start-up company was called "Mobile Technologies" and the product was called "Jibbigo". You could speak into the phone and then the system translated the input into another language. It was a huge success. Apple, for example, ran commercials with it. It was used everywhere and people came back to us, saying: "I can finally understand my in-laws!", and "I can really understand other people!" And we also started doing humanitarian projects, for example, we built systems, say, in Thai and Khmer, so that American, European, or Japanese doctors could help rural people get healthcare, and we deployed similar things in South America.

Due to broad interest in this type of technology, the company was then acquired by Facebook (making the world "open and connected") in 2013 and for two years I led a team of scientists to build translation technology there. At Facebook, the use of the technology for translation of posts and other company use cases, however, turned out to be of higher priority than the interactive communication aspect of our speech translators I was keen on advancing, and so I returned to the University to continue our work on the educational and humanitarian aspects of this technology.

Due to Cambridge Analytica and other scandals, Facebook has increasingly been confronted with massive criticism, which is why the tech giant is all the more under pressure to meet their idealistic agenda. At the same time, companies like Microsoft or OpenAI are demanding the democratization of AI. What is your general opinion on this concept?
Again, you know, the world is much more complex than a simple slogan suggests. Facebook is a good case in point. I am sure their initial goal was to democratize news. If anyone can post news, how wonderful that can be! If anyone can provide facts on Wikipedia, how wonderful would that be! No more experts dictating their opinions, right? But if you really think this through – if anybody can publish trash about anybody and reach a worldwide audience – the benefit is not necessarily that you are making people heard that were unheard before, but you also open up a worldwide potential for abuse and manipulation. And that is exactly what we are realizing now. So democratization is fine. But the potential for massive manipulation and abuse is equally there in the same process, and therefore one has to be really careful.

In other words: You are more or less skeptical about this concept?
Once again, we should be careful and keep on thinking about what we are doing because tools like the Internet are so powerful. Sometimes you create things that have unintended consequences. And one must reevaluate the technology and strive to move it in a good direction. While the internet lead to democratization of information, we now see again massive concentration of information and power as well. Would you rather have a world in which only Google, Microsoft, Facebook, Amazon or Apple can have intelligent systems and everybody else is at their mercy with regard to using this technology? Would you like to have a world in which only one of the big tech giants can recognize anybody's face by a machine and nobody else is able to? These technologies effectively encourage monopolies, that are hording incredible amount of data and generate a lot of knowledge, but – despite the best intentions – at the same time also provide a lot of potential for manipulation. We have recently seen that this is the case with the Cambridge Analytica scandal. So, again, the question is: Would you like to see all data and AI to be concentrated with only three or four companies in the world?

These concerns also play out on a geo-political stage. While the internet was designed to be a great global unifying force, it now also threatens to break into major regional spheres with different moral and societal attitudes that compete for supremacy. In China, where there are fewer laws or restrictions to data collection and handling, we see that AI feeds the emergence of an automated mass surveillance state that is overseen by the government. Will this – by way of competition – undermine Western values of privacy, freedom, and independence? That is a 1984 Orwellian future at our doorstep, right? Indeed, democratizing it at least distributes the technology to a broader set of players and that is why antitrust

efforts are so important, domestically. But, if we talk about global balances – for example – Europe versus America: Europe does not have a large internet company and this creates asymmetries, where one continent is critically dependent on AI systems from another for its information and data management. This still works, because relations between America and Europe are amicable and supportive because both are Western democracies. But what about China, Russia, India? China is making dramatic progress in AI right know, and is spending billions of dollars on AI. So it is only a matter of time until China and others will be on par, if not more advanced then the US. And, without its own clear technology base and vision, Europe, could be at the mercy of what other players are doing and be much more vulnerable to external meddling and manipulation. For me, these are worrisome developments.

Nevertheless, it still makes a difference whether we talk about American or Chinese tech monopolies.

Right, the so-called GAFA [Google, Apple, Facebook, Amazon] have to be considered differently in some sense. In China, companies like Baidu, Alibaba or Tencent have very strong government connections. In the US, the playing field looks different because there is a public that watches these companies, and whenever one of them abuses their data power, it becomes a scandal and is immediately all over the news, which is very bad for the company. While I was with Facebook, I do have to say that – despite all the scandals – I was impressed by how much the company actually attempted to deal with their data in a responsible way. And the fact that they still produce scandals simply shows how hard it is to do that and how sensitive an issue it is. But I think that companies in the US also embrace the idea of democratizing AI because it is part of their business model. Of course, companies ultimately are very selfish and try to do what is best for them. But in the American context, protecting data is good for business, since scandals are terrible. Hence, Microsoft and Google are into democratizing AI because it supports their business strategies. Take Amazon, for example: One of its largest businesses is Amazon Web Services (AWS) that, among other things, includes renting nodes, so that a small company can do its computation on Amazon's servers. Microsoft or Google provide similar computational resources, and if they can provide AI services on top of that, then it is obviously also good for their business.

So which would you consider to be the most important political challenges of AI in the near future, from your perspective as a computer scientist?
I think that, first of all, we as scientists have a responsibility to be vigilant. But it gives rise to optimism that there are actually a lot of idealistic people working inside those big companies. Thus, the fact that there are scandals is good news, because it means that you cannot keep such things secret, that it forces society to keep thinking these things through, which is good. And regarding how politics should respond: Well, if you look at some of these senatorial debates, you realize that politicians cannot be deeply involved in every aspect of every technology, and hence may lack intuition about where it may go and how to respond. For this reason, I think it is very important for AI scientists to be vocal and active in a public dialog, so we (science, public, and governments) can ensure that we build these technologies to serve humanity, as opposed to serving our own political or financial interest.

What worries me in this context, however, is the fact that large companies are voraciously hiring scientists, and that universities have difficulties retaining talented people. And we should remember that universities are (or should be) spaces for open discussion and debate so that we are not manipulated by economic or political interests. So my point is that we should maintain a strong academic environment in all major areas in which AI is used. And this is a particular challenge in Europe: Without major internet companies, it naturally suffers from a continuing loss of talent. With a reference to my own AI laboratory in Germany, I can tell you that many of the best scientists, as soon as they are done with their pitches or degrees, get offers that are like seven times higher (or even more) than the ones we can afford at a university. And when young people are being offered those amounts of money plus a chance at building something with a major company like Amazon, Apple, or Facebook, they jump at the opportunity. In other words, the brain drain is enormous, not just from academia to industry, but between countries. Therefore, in an age of AI, Europe must move much more aggressively to provide for its future.

What can Europe do to change that? And how do politicians, people, or the public know which experts they can trust?
Well, as to your last question, I think by being less risk averse and doing more to encourage technology disruption: Europe has outstanding scientists and engineers. There is also outstanding support and freedom in Europe to carry out innovative and fundamental science. Europe has very bright, well-educated, and idealistic scientists. I could even argue that many of the top scientists in America were trained or started their career in Europe. That is not just an empty phrase, I could name many famous examples. But the area in which we are doing badly is letting the scientific advances challenge the status-quo in society. What is needed

is fast, practical, and disruptive moonshot projects. DARPA in the US, for example, has done that very successfully for the government. And so do companies like Google, Tesla, Amazon, all of which did not exist 30 years ago.

The other thing that should be improved is the technology transfer into industrial exploitation. In Europe, we actually have many entrepreneurs who start companies. The risk takers are there, the young people are there, the bright ideas are there, and the excitement and the eagerness to do this are there. What's missing is capital to support such activities and also more willingness and speed of mergers and acquisitions. For example, in the US, small companies are bought up very quickly. Some of the companies exist only twelve or eighteen months before they are being absorbed by a larger corporation. This is a healthy process as it speeds the transition from idea to concept to product to industry. But in Europe, that is very rare. Here, it is very difficult for companies to be bought. It takes a long time to go public, to enter the stock market and so on. The transition from a small successful start-up to a large business has so much friction in Europe that it misses many opportunities; speed is of the essence in this kind of game. And this ultimately drives many small companies and their young entrepreneurs to the US and China.

Another political-ethical concern that many people talk about these days is the problem of algorithmic biases. How are these problems related to your research in natural language processing and the translation of spoken languages?

I am glad that you are bringing us back to this topic. So far, we talked a lot about how AI affects society. Another important political dimension is to discuss how we pick projects that contribute to a society that we want to live in. And for me that means speech translation, because I think this is one of the big problems in Europe. Europeans speak 23 different languages, and these are only the official ones. In fact, there are many more languages in Europe. And this situation generates separation, misunderstandings, and also a considerable loss in business opportunities. One big reason why e-commerce is more challenging in Europe is because it is so fragmented. Each country in Europe has a different legal system, a different delivery system, and much of that is of course fossilized in language, because if everything has to be done in multiple languages, it complicates transnational business exchanges. But saying that everyone should learn English, Esperanto, or something like that would be ridiculous and also not desirable. In fact, having the variety and diversity of languages is something Europeans are rightfully proud of.

Against this background, technology must not be regarded as an obstacle, but as a tremendous problem solver if we want to develop a technology capable of text and speech translation that bridges these language barriers on all fronts, so that we actually have a language-transparent world. Imagine you are going to China, Russia, or Spain, and like to operate in these countries as if you are at home, with-

out any language disadvantages. But if you think that through, what such a scenario would mean if you are in all these countries without knowing the respective language, what all kind of assistance it would require so that you do not notice the language barrier anymore. And indeed to achieve this is the very vision we are working on.

Biographies

Alexander, V. N. is a philosopher of science, author of The Biologist's Mistress: Rethinking Self-Organization in Art, Literature and Nature. She is a Rockefeller Foundation Bellagio Center alum, former Public Scholar for the NY Council for the Humanities, member of the Third Way of Evolution group, and a director at the Dactyl Foundation. She is on the editorial boards of Biosemiotics journal (Springer Publishing) and Meaning Systems book series (Fordham University Press). Her work in saltational evolutionary theory appears in Fine Lines: Vladimir Nabokov's Scientific Art, published by Yale University Press. Alexander will work as a Digital Humanities Fulbright Scholar in Saint Petersburg, Russia in 2020.

Beverungen, Armin is (since October 2019) Junior Professor for Organisation in Digital Cultures at Leuphana University of Lüneburg, and has held previous research and teaching positions at the University of the West of England, Leuphana University and the University of Siegen. He is a founding co-editor of the journal *spheres: Journal for Digital Cultures* (www.spheres-journal.org) and the book series Digital Cultures (meson press). During the summer of 2019 he was a fellow at the Center for Advanced Internet Studies in Bochum. His research takes place at the interstices of media and organization studies, and is currently focused on the phenomenon of algorithmic management. His most recent publications include *Markets* (with Jens Schröter/Phil Mirowski/Edward Nik-Khak, meson press and University of Minnesota Press) and an edited issue of *Organization* on the theme of "the organizational powers of digital media" (with Lisa Conrad/Timon Beyes).

Burkhardt, Marcus is postdoctoral research associate at the chair for Digital Media and Methods at the University of Siegen. His research focuses on the history and theory of digital media, especially the kogi(sti)cs of database technologies, big data, and algorithmic environments as well as on media of knowledge production and dissemination, media philosophy and media theory.

Dippel, Anne is Scientific Researcher and Lecturer at the Department for Cultural Anthropology and Cultural History of Friedrich-Schiller-University Jena. She worked at the Custer of Excellence Image-Knowledge-Gestalt of Humboldt-University Berlin, has been research fellow at the Institute for the Advanced Studies

of Media Cultures and Computer Simulations (mecs), Leuphana University Lüneburg, and has been visiting assistant professor at the Science, Technology and Society Programme at MIT, and for the time of her field work an associated member of CERN collaboration. Dippel, A. (2017). Das Big Data Game. Zur spielerischen Konstitution kollaborativer Wissensproduktion in der Hochenergiephysik am CERN. In: *NTM 4/2017*, 485 -517. Dippel, A. & Warnke M. (Ed.) (2017). *Interferences and Events. Epistemic Shifts in Physics through Computer Simulations*. Lüneburg: Meson Press. Dippel, A. Arbeit (2018). In: Markus Rautzenberg, Daniel Martin Feige, Michael Ostritsch (Hrsg.): *Philosophie des Computerspiels*. Stuttgart: Metzler Verlag, 123-148.

Djeffal, Christian is Professor for Law, Science, and Technology at Technical University Munich. He focuses on the relationship between law, technology and society. He is specifically interested in new technologies like artificial intelligence and the internet of things. His interdisciplinary research addresses constitutional law, regulation and standards. Christian received his PhD from Humboldt University of Berlin for his thesis "Static and Evolutive Treaty Interpretation: A Functional Reconstruction" which was published by Cambridge University Press. As a Postdoc, he coordinated the research group "Global Constitutionalism and the Internet" at Alexander-von-Humboldt-Institute for Internet and Society.

Förster, Yvonne works as a Research Professor at Shanxi University, China and teaches Philosophy at Leuphana University Lüneburg, Germany. She received her PhD at the Friedrich-Schiller-University Jena on *Experience and Ontology of Time* (*Zeiterfahrung und Ontologie,* München: Fink 2012). As a visiting professor, she has taught aesthetics at Bauhaus University Weimar and been recently awarded Senior Research Fellowships at two Institutes for Advanced Studies (*Media Cultures of Computer Simulation* at Leuphana and *Cultural Sciences* at University of Konstanz). Her research focuses on philosophy of technology, aesthetics, phenomenology, and fashion as art. For more information see: www.yvonnefoerster.com

Großmann, Jürgen, Dr.-Ing. is an expert on model-based development, model driven testing as well as in security engineering and security testing. Furthermore, Jürgen Großmann has experiences in testing and modeling automotive software systems and applications, especially ITS systems. He has extensive experience in developing test benches and test laboratories for conformance and interoperability testing in this area. Jürgen Großmann has experiences in numerous standardization activities for various standardization bodies, including OMG, ETSI, ASAM and AUTOSAR.

Matzner, Tobias is professor for "Media, Algorithms and Society" at Paderborn University in Germany. His research focuses on the intersection of digital technology, culture, politics and ethics. Recent publications include: The Human is Dead—Long Live the Algorithm! Human-algorithmic ensembles and liberal subjectivity. In: *Theory, Culture & Society* online first 2019. "Grasping the ethics and politics of algorithms". In *The Politics and Policies of Big Data*. Ed. by Ann Rudinow Sætnan, Ingrid Schneider und Nicola Green. London: Routledge 2018.

McQuillan, Dan is Lecturer in Creative & Social Computing at Goldsmiths, University of London. After his Ph.D in Experimental Particle Physics he worked in learning disabilities and mental health, and later in human rights. His research interests include the social impact of AI and the practice of radical citizen science. Recent publications include "People's Councils for Ethical Machine Learning", "Data Science as Machinic Neoplatonism" and "Algorithmic States of Exception". A selection of his public writings can be found athttps://www.opendemocracy.net/en/author/dan-mcquillan/

Monea, Alexander is Assistant Professor at George Mason University, serving jointly in the English Department and Cultural Studies PhD Program. He researches the history and cultural impacts of computation, algorithms, and big data. His current book project examines how heteronormative biases get embedded in the machine vision algorithms and content filters that control the flow of internet communications. He has previously published papers in *Computational Culture* and *Cultural Studies ↔ Critical Methodologies* that examine the power of graph databases and that examine the implementation of Google's Knowledge Graph for the mapping of web semantics. He has also co-authored with Jeremy Packer an article for *The International Journal of Communication* on what they term 'media genealogy', a methodology for examining the emergence of technologies with a focus on their political and social impacts. Recent Publications: Monea, A. (2016). Graph Force: Rhetorical Machines and the *N*-Arization of Knowledge. *Computational Culture, 5*. Monea, A. (2016). The Graphing of Difference: Numerical Mediation and the Case of Google's Knowledge Graph. *Cultural Studies ↔ Critical Methodologies, 16*(5), 452-461. Monea, A. & Packer, J. (2016). Media Genealogy and the Politics of Archaeology. *The International Journal of Communication, 10*, 3141–3159.

Morin, Kevin is PhD Candidate at the National Institute for Scientific Research in Canada. His work is about innovation ecosystem in digital domain such as machine learnings and artificial intelligence. Mixing approach from STS and Critical Theory, his thesis describes the regimes of justification, actor networks, discourses and the digital culture of technosystem such as AI development.

Pasquale, Frank researches the law of big data, artificial intelligence, and algorithms. He has testified before or advised groups ranging from the US Department of Health and Human Services, House Judiciary Committee, and Federal Trade Commission, as well as directorates-general of the European Commission. He is the author of The Black Box Society (Harvard University Press, 2015), which has been translated into Chinese, Korean, French, and Serbian. The book developed a social theory of reputation, search, and finance. He has served on the NSF-sponsored Council on Big Data, Ethics, & Society. He has spoken on data policy at many universities and public lectures. Frank has co-authored a casebook on administrative law and co-authored or authored over 50 scholarly articles. He co-convened the conference "Unlocking the Black Box: The Promise and Limits of Algorithmic Accountability in the Professions" at Yale University. He is now at work on a book tentatively titled Laws of Robotics: The Future of Professionalism in an Era of Automation (under contract to Harvard University Press).

Reutter, Lisa is a Ph.D. student at the Department of Sociology and Political Science at NTNU, Trondheim, currently researching the application and use of machine learning in the Norwegian public sector. She is in addition part of the "Digital Infrastructures and Citizen Empowerment (DICE)" research project.

Rieger, Stefan studied German language and literature and philosophy. He wrote his dissertation on baroque data processing and mnemonics and his habilitation thesis on the relationship between media and anthropology (Die Individualität der Medien. Eine Geschichte der Wissenschaften vom Menschen, Frankfurt/M. 2001). Current research areas are: History of science, media theory, and cultural techniques. Since 2007, he is Professor of Media History at the Ruhr-Universität Bochum. Recent book publications*: Bunte Steine. Ein Lapidarium des Wissens* (Berlin: Suhrkamp 2014, mit Benjamin Bühler) and *Die Enden des Körpers. Versuch einer negativen Prothetik*, Wiesbaden: Springer.

Roberge, Jonathan is Associate Professor of cultural and urban sociology at the Institut National de la Recherche Scientifique, where he also holds the Canada Research Chair in Digital Culture. He is among the first scholars in North America to have critically focused on the production of algorithms, a research agenda which culminated into a book entitled Algorithmic Cultures (Routledge, 2016, translated into German at transcript Verlag, 2017). He currently works on a manuscript entitled The Cultural Life of Machine Learning to be out early in 2020 at Palgrave MacMillan (together with Micheal Castelle). Recent Publications: SEYFERT, R. et ROBERGE, J. (eds), Algorithmic Cultures. Essays on Meaning, Performance and New Technologies, London, Routledge, 2016.; ROBERGE, J. et MELANÇON, L., "Being the King Kong of Algorithmic Culture is a Tough Job After All. Google's Re-

gimes of Justification and the Meanings of Glass", Convergence. The International Journal of Research into New Media Technologies, vol. 23, no 3, 2017, pp. 306-324.

Schieferdecker, Ina, Prof. Dr.-Ing., is Co-Director of Fraunhofer FOKUS, Berlin and Professor for Quality Engineering of Open Distributed Systems at the Technische Universität Berlin. She is also Director of the Weizenbaum Institute for the Networked Society, the German Internet-Institute in Berlin. Her research interests include (urban) data platforms, critical infrastructures as well as conformity, interoperability, security, reliability and certification of software-based systems. She is President of the Association for Software Quality and Education (ASQF), member of the German National Academy of Science and Engineering (acatech), of the German Advisory Council on Global Change (WBGU), and member of the "Hightech Strategy 2025" of the Federal Ministry of Education and Research.

Schneider, Martin A. is engaged in advanced testing methods and techniques and works on model-based testing techniques for security aspects based on different fuzzing techniques, security test patterns and security testing metrics and on the integration of other methods like the risk analysis in order to increase the efficiency of fuzz testing. By that, he develops and evaluates innovative security test methods in the frame of various industrial domains with special focus on complex systems, cyber-physical systems and service-oriented architectures.

Schröter, Jens, Prof. Dr., is chair for media studies at the University of Bonn since 2015. He was Professor for Multimedial Systems at the University of Siegen 2008-2015. He was director of the graduate school "Locating Media" at the University of Siegen from 2008-2012. He is member of the DFG-graduate research center "Locating Media" at the University of Siegen since 2012. He was (together with Prof. Dr. Lorenz Engell, Weimar) director of the DFG-research project "TV Series as Reflection and Projection of Change" from 2010-2014. He was speaker of the research project (VW foundation; together with Dr. Stefan Meretz; Dr. Hanno Pahl and Dr. Manuel Scholz-Wäckerle) "Society after Money—A Dialogue", 2016-2018. Since 4/2018 director (together with Anja Stöffler, Mainz) of the DFG-research project "Van Gogh TV. Critical Edition, Multimedia-documentation and analysis of their Estate" (3 years). Since 10/2018 speaker of the research project (VW foundation; together with Prof. Dr. Gabriele Gramelsberger; Dr. Stefan Meretz; Dr. Hanno Pahl and Dr. Manuel Scholz-Wäckerle) "Society after Money—A Simulation" (4 years). April/May 2014: „John von Neumann"-fellowship at the University of Szeged, Hungary. September 2014: Guest Professor, Guangdong University of Foreign Studies, Guangzhou, People's Republic of China. Winter 2014/15: Senior-fellowship at the research group „Media Cultures of Computer Simulation", Summer 2017: Senior-fellowship IFK Vienna, Austria. Winter 2018: Senior-fellowship IKKM

Weimar. Recent publications: (together with "Project Society after Money") Postmonetär denken, Wiesbaden: Springer 2018; (together with "Project Society after Money"): Society after Money. A Dialogue, London/New York: Bloomsbury 2019; (together with Armin Beverungen, Philip Mirowski, Edward Nik-Khah): Markets, Minneapolis/London: University of Minnesota Press and Lüneburg: Meson (Series: In Search of Media). Visit www.medienkulturwissenschaft-bonn.de / www.theorie-der-medien.de / www.fanhsiu-kadesch.de

Senneville, Marius is an Urban studies M.A. student at INRS, where he works on the AI ecosystems of Montreal and Toronto. With an approach informed by both STS and the study of corporate governance, he focuses on the various configurations of collaborative work between industrial and academic scientists and the way they impact the research being produced.

Spilker, Hendrik Storstein is professor in the sociology of media and technology at the Department of Sociology and Political Science at NTNU, Trondheim. He has written several books and articles about the cultural politics of Internet and digitalization, including "Digital Music Distribution: The Sociology of Online Music Streams" (Routledge, 2018) and is currently heading a large research project on "Digital Infrastructures and Citizen Empowerment (DICE)."

Sudmann, Andreas worked in Göttingen, Regensburg, Vienna, and Berlin, and currently teaches and researches as an adjunct professor (*Privatdozent*) of media studies at the Ruhr University Bochum, where he also completed his habilitation in 2016. Most recently, he held a guest professorship in media studies at the Philipps University in Marburg followed by a research fellowship at the IFK in Vienna. Current research foci include media-theoretical and historical problems of artificial intelligence (AI), specifically machine-learning methods, aesthetics and politics of popular, and technical/digital visual media, forms and processes of seriality and documentary, media theory, and media critique. Sudmann is the author of several books and edited collections in the field of media studies and digital culture studies, among them *Computer Games as a Sociocultural Phenomenon* (published with Palgrave Macmillan in 2008) and *Digital Seriality* (special-issue of *Eludamos. Journal for Computer Game Culture*, 2014). His new book *Serielle Überbietung. Zur televisuellen Ästhetik und Philosophie exponierter Steigerungen* (*Serial Outbidding: The Televisual Aesthetics and Philosophy of Ostentatious Escalation*) has been published by Metzler (Stuttgart) in October 2017. Together wie Christoph Engemann, he has co-edited the anthology *Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz* (Machine Learning. Media, Infrastructures, and Technologies of AI).

Volmar, Axel is a postdoctoral fellow at the Collaborative Research Centre "Media of Cooperation" (SFB 1187 Medien der Kooperation) at the University of Siegen. He has been a Mellon Postdoctoral Fellow in the Department of Art History and Communication Studies at McGill University (2014-2016). His current research interests reside at the intersections of media studies, history of science, and sound studies and focus on the history of audiovisual telecommunications. He is the author of *Klang-Experimente. Die auditive Kultur der Naturwissenschaften 1761-1961* (Frankfurt/M: Campus, 2015) and has co-edited several special issues and collected volumes in media and sound studies. His recent publications include the collected volume *Format Matters. Theories, Histories, Practices*, co-edited with Marek Jancovic and Alexandra Schneider (Lüneburg: meson press, 2019).

Acknowledgments

This volume is the first publication of a newly founded book series entitled "AI Critique". It is based in part on contributions from an international conference that took place in Bochum in 2018 and that was generously funded by the Center for Advanced Internet Studies (CAIS).

A number of wonderful colleagues took part in this conference, including Friedrich Balke, Michael Baurmann, Armin Beverungen, Gioele Barabucci, Marcus Burkhardt, Anne Dippel, Christian Djeffal, Till Heilmann, Timo Honkela, Lisa Gotto, Tobias Matzner, Christoph von der Malsburg, Roland Memisevic, Daniel Neyland, Ina Schieferdecker, and Claus Pias.

Many of these scholars have also written an essay for this volume. Other colleagues joined the project later, including Dan Mcquillan, Alexander Monea, Frank Pasquale, Stefan Rieger, Jonathan Roberge, VN Alexander, and Jens Schröter.

This project was completed not least because of two fellowships, one by the CAIS in Bochum, the other by the IFK in Vienna. I am very grateful to both institutions for allowing me to do research under excellent conditions.

Sincere thanks also to Anna Tuschling und Bernhard Dotzler for their interest and commitment in starting this book series as well as to the many colleagues and friends who have helped me over the years in either researching or reflecting on AI in different ways: Friedrich Balke, Michael Baurmann, Regina Bendix, Ulrike Bergermann, Armin Beverungen, Shane Denson, Julia Eckel, Lorenz Engell, Christoph Ernst, Jens Eder, Oliver Fahle, Yvonne Förster, Lisa Gotto, Irina Kaldrack, Frank Kelleter, Peter Klimczak, Petra Löffler, Thomas Macho, Tobias Matzner, Alexander Monea, Roland Memisevic, Moritz Müller-Freitag, Till Heilmann, Kathrin Peters, Ben Peters, John Durham Peters, Claus Pias, Stefan Rieger, Simon Rothöhler, Jens Ruchatz, Gabriele Schabacher, Jens Schröter, Philipp Schweighauser, Florian Sprenger, Daniela Wentz, and Sabine Wirth.

My deepest gratitude goes to my family, especially to my wife Anne, for their support and love. This book is dedicated to them with all my love.