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.