Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2019: 1–11

DOI: 10.1177/2053951718819569
journals.sagepub.com/home/bds
SAGE

Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media

Anja Bechmann[1] and Geoffrey C Bowker[2]

Abstract

Artificial Intelligence (AI) in the form of different machine learning models is applied to Big Data as a way to turn data into valuable knowledge. The rhetoric is that ensuing predictions work well—with a high degree of autonomy and automation. We argue that we need to analyze the process of applying machine learning in depth and highlight at what point human knowledge production takes place in seemingly autonomous work. This article reintroduces classification theory as an important framework for understanding such seemingly invisible knowledge production in the machine learning development and design processes. We suggest a framework for studying such classification closely tied to different steps in the work process and exemplify the framework on two experiments with machine learning applied to Facebook data from one of our labs. By doing so we demonstrate ways in which classification and potential discrimination take place in even seemingly unsupervised and autonomous models. Moving away from concepts of non-supervision and autonomy enable us to understand the underlying classificatory dispositifs in the work process and that this form of analysis constitutes a first step towards governance of artificial intelligence.

Keywords

Artificial intelligence, machine learning, classification, social media, Facebook, discrimination, bias

This article is a part of special theme on Knowledge Production. To see a full list of all articles in this special theme, please click here: http://journals.sagepub.com/page/bds/collections/knowledge-production.

There is a long tradition of equating knowledge with classification—in the physical sciences, the classification of subatomic particles is a core endeavor; in chemistry, Mendeleev's table was a fundamental breakthrough which gave us classes of elements; in botany and biology, the Linnean classification system is still at the root of scientific work and the central pursuit of cladistics is classification. The argument that Big Data can do without large scale classificatory work has been made at opposite ends of the spectrum by editor of Wired Chris Anderson, and social philosopher Bruno Latour (Anderson, 2008; Bowker, 2014 for critique; Latour, 2002; Shirky, 2005).

In their magisterial study about the use of "Big Data", Lehr and Ohm among others (Barocas and Selbst, 2016; Diakopoulos and Koliska, 2017; Lehr and Ohm, 2017) point to the profusion of layers at which social and political factors can enter into the deployment of Big Data to "automatically" and "dynamically" assign (in our reading) classes on the fly: for instance data collection; data cleaning; data partitioning; model selection; model training (including tuning and assessment); and model deployment. In this article, we shall consider the most relevant layers for understanding classifications as they arise in artificial

[1]Aarhus Institute of Advanced Studies, Aarhus University, Aarhus C, Denmark
[2]Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, CA, USA

Corresponding author:
Anja Bechmann, Aarhus Institute of Advanced Studies, Aarhus University, Hoegh Guldbergs Gade 6B, DK-8000 Aarhus C, Denmark.
Email: anjabechmann@aias.au.dk

intelligence (AI) and machine learning with the aim of making visible knowledge production.

Data collection is clearly one area in which data classification can occur. We may find the number of tweets per day (some 500 million in 2018—https://blog.hootsuite.com/twitter-statistics/, accessed 27 November 2018) or the number of Facebook users (2.3 billion active users in 2018—https://zephoria.com/top-15-valuable-facebook-statistics/, accessed 27 November 2018) staggering; however we always need to remember that these numbers do not of themselves provide representativity of the total population outside social media (Bechmann and Vahlstrup, 2015; Lomborg and Bechmann, 2014). What we get reflected back from large scale studies of these platforms is not society as it is, but a society that is classified immediately into users (of interest, accessible) and non-users (not of interest, inaccessible).

Another step at which classification work gets done is *data cleaning*. Walford (2014) has written beautifully about the work of data cleaning in the canopy of the Brazilian rain forest. Certain results from streaming sensors of the environment, which we may think of as objective representations of reality, get routinely rejected from the databases being built. If a temperature reading or a window reading is outside of the permitted range, it will be excluded in the scrubbing process. Thus anomalies are weeded out before they can be spotted—the world has a classified set of behaviors that can only exist within certain parameters.

Model training is a third step of classificatory work. Jaton (2017) has shown in the case of a new machine learning algorithm for detecting complex photographs (with more than one object in focus) that the innovators had to try to establish their training set in order to get their results accepted. The problem was that their model did not work as well as the finely tuned models when looking at a single focus image. Ultimately, their work was rejected because of this flaw. There is an argument, then that the scientific and organizational authority to create a training set was a core part of the process. And again, only recognition algorithms that worked optimally over a certain class of objects were considered valid.

We recently have witnessed how the largest communication platform in the world Facebook has weaponized political propaganda. This became especially clear in the case of Cambridge Analytica. The company collected data on millions of users through Facebook third party apps to understand the correlation between psychological profiles and platform behavior such as ''like'' patterns (Kosinski et al., 2013). This data inferred knowledge allowed the company to target more precisely voters with the specific message appealing to particular geo-located, profiles and potentially win over votes (Cadwalladr, 2017). According to John Dewey (1927), democracy can only be performed if different groups interact flexibly and fully in connection with other groups through ''free'' and ''open'' communication. Political micro-targeting brings into question whether such ''open'' communication is taking place. The increasing entrenchment of privately owned media ownership into an international oligopoly again questions ''free'' and ''open'' communication—particularly since we have limited access to the logics and structures on which social media are built.

One such logic is the use of AI in the algorithms of social media and how the reasoning of such machines on top of these structural problems can potentially create problems of visibility, redlining and other discrimination such as targeting, favoring and normalizing some people over others (Caliskan et al., 2017; Citron and Pasquale, 2014; Eubanks, 2017; Howard, 2005; Levin, 2016; Sweeney, 2013). AI and machine learning are concepts often used as synonyms to describe widely used yet controversial computational models employed to cluster and make sense of data to inform and predict actions in the Big Data era (see also Russel and Norvig, 2010).

Despite the as yet imperfect state of these models for interpreting and predicting action from data, they have an increasingly significant influence on decisions made in an increasingly data-driven society. In line with critical algorithmic scholars (Ananny and Crawford, 2018; Boyd and Crawford, 2012; Cheney-Lippold, 2017; Citron and Pasquale, 2014; Elish and Boyd, 2018; O'Neil, 2016; Sandvig et al., 2016), we argue theoretically and show through empirical case studies that such models and associated classification *dispositifs* are central objects of study in order to provide a critical yet informed discussion on the knowledge production of AI. This article aims to provide a framework for analyzing social, cultural, and political classification *dispositifs* of supervised and unsupervised machine learning models as models that are seemingly autonomous.

We will theoretically discuss classification as a central knowledge producing concept and how it has been applied to AI in general. This will be followed by examples on the use of two different models with different degrees of classification—topic modelling with text2vec (unsupervised) and deep neural network picture pattern recognition with inception v.3 (supervised) applied to Facebook data in one of our labs. The purpose is to detail at what point in the work process of applying AI classification, as defined theoretically in the previous sections, takes place. The article contributes to the existing literature on knowledge production in AI and potential problems with accountability (Ananny and Crawford, 2018; Barocas and Selbst, 2016; Burrell, 2016; Dwork, 2006; Hardt, 2011; Kroll et al., 2017;

Lehr and Ohm, 2017) by focusing even more on the human nexus in the work process and at what point classification is carried out that can result in counter-productive outputs for democratic societies and their shared human values.

AI, machine learning, and the role of classification

Theories of AI have historically distinguished between strong/general and weak/narrow AI (Searle, 1980; Slezak, 1989). Searle (1980) describes weak AI as "a powerful tool" that, for instance, allows for us to examine larger amounts of data in a more rigorous and precise way than what we as human brains would be able to. An example of such processing is using Baysian methods to create data inferred classifiers (Rieder, 2017). With the current fascination of Big Data, such uses of AI are widespread in all areas of the digital layer of everyday life through predictions that lead to actions, be it within robotics, communication optimization and manipulation, or behavior adjustments.

However, weak AI—as the most widespread use of AI—is best understood in the historical context in which strong AI as an "imitation game" (Turing, 1950) has played a major role as driver for AI research and developments, over the past 70 years and still present today within the development of humanoids (Kanda et al., 2004). In comparison to weak AI, strong AI has the goal of imitating human behavior and communication. Strong AI is defined by Searle as a computer mind with intentionality acting as a human mind that is able to "understand and have other cognitive states" (p. 417). Similar visions can be traced back in time (Buchanan, 2005), but newer vision is often connected to the Turing test (1950), where Turing asks "can a machine think?". For a program to pass the test, the human must not be able to tell in a dialogue that the person is speaking to a computer (pretending to be a woman). In this vision, humans are used as a benchmark to measure computational success in a simulation optic. Many examples demonstrate that it is difficult to establish a structural classification scheme or an artificial understanding of the implicit rules and tacit knowledge (Polanyi, 1966) socially inferred from a specific society. With a growing global media arena that complexity only increases. Inspired by the Turing test, many developers have tried building software bots that chat with human/bot peers within communities such as Twitter but have failed to encode classifiers of such tacit knowledge. Using AI on social media will, on the surface, seem like strong AI, because the machine would have to decode logics of human communication and behavior and adapt to changes in this.

An example of such an attempt was the release of the Twitter chatbot Tay powered by Microsoft in 2016 that turned into a female hating, Nazi sympathizer and had to be shut down only 16 hours after release. Why did the experiment fail? Alba (2016) suggests that the feedback loop is problematic because the AI acted on top of the input it was provided with. So, if the input data intentionally or unintentionally display unacceptable classes of behavior or social values, the AI will mirror these unless there is an intervention in the programming phase, e.g. hardcoded adjusted thresholds or black listing outcome variables. As digital social scientists, we would add that it could also play a role that the logics changed without the social chat bot noticing the social cues. From trying to mirror a traditional Twitter conversation, the chat turned into a game where the teenagers tried to make the social chat bot deliberately display "unacceptable" social values, triggering classifiers that were not acceptable in the wider community. The debate about simulating human thought, and of originality and intentionality (strong AI) versus human enhancing and computing power to process enormous amounts of data (weak AI) is important to the discussion on classification. We will argue that it is precisely the quality of the classifiers that result in the perceived success or failure for a given social context. As the discussion on weak and strong AI suggests, classifiers are present in both cases. In weak AI applications, classifications are generated from a lot of training data and learning iterations. In the strong AI vision they are meant to imitate humans' way of approaching the world, training and learning being a significant part of this, but also being able to recognize social cues in shifting contexts.

Classification theory, social categories, and claimed lack thereof

Classification is a natural part of human reasoning and is important in order to understand the world around us (Foucault, 1971); as the boxes that we structure the world around and in. However, despite such boxes being dynamic, these widely accepted boxes only allow us to see the world from certain cultural and historical standards, often exclude those at the margins and the "residual categories" that do not fit into our system or negations of the standardized classifiers: "deciding what will be visible and invisible in the systems" (Bowker and Star, 1999: 44).

Legally protected classes of people may well be covered in classification systems, however computer systems are often blind to other kinds of potential discrimination: for instance, towards left-handed people, people with allergies (Star, 1990), people who are predicted to be depressed or pregnant in the near

future, or receptive to political campaign messages (Tufekci, 2014). In the contemporary media landscape where platforms create increasingly international media fora transgressing different cultures and cultural classification systems, an American standard would also be blind towards Indian, Egyptian or Korean classes inferred from the specific history and culture in these societies. Classes are also ubiquitous and can create ''cumulative mess'' (DiPrete and Eirich, 2006; Strauss et al., 1985). When we approach classes ecologically, we find that individuals, for instance, can be classified as many things at the same time, and some of the classes can (from the standards we work with) seem opposite or working against each other, leading to confusion for the algorithmic outcome. Such contradictory classes may be due to classes deriving from a certain context and not taking into consideration that the person can interact differently in different contexts as we saw with the chatbot Tay; that the classes are layered, textured, and tangled (Strauss et al., 1985), and that ''one size does NOT fit all!'' (Gasser, 1986). We need to work with parallel or multiple representational forms (Bowker and Star, 1999).

Statistics ''is immanently a science of classification'' (Farr, 1985: 252) and classification is another word for generalization. AI builds on statistics and other mathematical principles, and the main interest for platforms is often to create personalization and amplification to maintain and monetize user attention, presenting users with the norm trained on this specific person's own data traces paired with, for instance, relational data to find the right person to target with the right posts or ads. ''Ground Truth'' (Jaton, 2017) forms the apodictic basis that is used to train the learning algorithm to improve the prediction that is otherwise based on the historical behavior of the user (posts, likes, shares, comments, group memberships), similar users and the user's network as a prerequisite for what the person wants to do or to see in the future news feed. Still, such ground truths create problems if we turn to the classification theory of ubiquitous classification and cumulative mess; the basic assumption behind such ground truth being that; (1) there is a ground truth (disregarding confirmation bias); (2) scores can indicate whether ground truth has been reached; (3) we will repeat our past actions and preferences in the (as it has come to be) broad context of social media, ranging from different contextual uses of social media as in itself a ubiquitous service transgressing locations and incentives for use; (4) future predictions will match the ground truth registered for past interactions without model and system developers influencing the users and user behavior/incentives in question.

As pointed out in the late 1990s, if we work with too few categories, the information is not useful, and if we work with too many categories, the result will be increased bias, or randomness, on the part of those filling out the form: the Goldilocks zone is well described by Ashby in his ''law of requisite variety'' (1956). However, in the late 2010s, the information is not filled out in the system manually. This does not reduce the problems of too many categories; in fact, we argue that we see increased biases and randomness in actions built on top of multi-categorical processing. Categories are not a priori constructed, but highly context sensitive, following the cultural context of the person categorizing (Bowker and Star, 1999: 107) as well as being subjective, case- and site-specific as already suggested by Roth in the 1960s (1966). We assume that such qualitative findings in the 1960s also apply to the labelling industries in the 2010s (e.g. Mechanical Turk) that are widely used to identify features and other labels in training data; but what are the political consequences of such subjective processes and can system designers and developers apply AI without classification?

One site for the alleged disappearance of classification is the programing procedure of assigning ''dynamic proxy classes''. In object-oriented terms, this means that if you have a given classification built in (say humans as an example of species) then you can on the fly extend the category (to include, say, Neanderthals, as some would argue) by assigning Neanderthals automatically to the proxy class for humans—so that the object ''Neanderthal'' would act like the object ''human''. Through this technique, you can generate substantial drift from an initial category set. Significantly, however, we are still talking about classification work all the way down—the only issue is how visible and how a priori that work is. Instead of discussing lack of classification we argue that we need to account for how classes and social categorization arise in the design process as deliberate and unintentional consequences of decisions made.

Two concepts within AI deserve to be outlined and discussed in continuation of the weak and strong AI discussion: the concepts of supervised and unsupervised learning (Alpaydin, 2016). The concepts describe algorithms that work with classifiers or labels to generate predefined outputs (supervised) or algorithms that do not have predefined outputs (unsupervised). In unsupervised learning data is placed in clusters or other pattern recognition outputs according to the structure in the data, here conceptualized as *inductive classifiers* or *data inferred classification*. ''Inductive inference'' was debated critically by, for instance, Slezak in the late 80s: "these programs constitute 'pure' or socially uncontaminated instances of inductive inference, and are capable of autonomously deriving classical scientific laws from the raw observational data'' (1989: 563). Slezak, among others, questions how this computational method can distinguish ''mere contingent co-occurrence from

causal connection" (p. 565). Like Searle, Slezak was also skeptical about the ability to develop strong programs. Through a critique of Bloor's strong program that highlights social contexts as important to inferences and causal explanations (the sociological turn), he argues that such explanations would lead to insufficient explanations of the context following the critique of radical relativism, and instead argues for a turn towards social interests that are best outlined through cognitive science. The focus of cognitive science, however, has been on human brain processing and rational decisions (Hayles, 2017). What is rational to computers is not necessarily rational to humans because depiction of the situation and context differs, and the definition of the task may be more narrowly interpreted by algorithms, not taking into consideration an ecological approach to the consequences on other tasks such as societal inclusion and social cohesion—is inclusion a rational choice (e.g. in the case of Cambridge Analytica)? In this way, algorithmic ecology plays a role in governing AI as will be illustrated in the case studies of the next section.

Hidden layers of knowledge production in AI

As we have pointed out, there are several problems with data inferred classification. Despite the ability to modernize existing classes, they can be highly spurious and self-fulfilling. Furthermore, they can be difficult to backtrack logically, and the consequences of acting on top of such classes can be highly problematic if classes prove to follow a different social code than human society allows (Caliskan et al., 2017; Elish and Boyd, 2018). The lack of algorithmic ecology makes machine learning task oriented and not necessarily aware of the larger consequences of a closed decision process and the influence on the larger socio-economic and political climate.

We exemplify this empirically by outlining our work with two different models on Facebook data from one of our labs, deploying a participant observation inspired methodology (Hammersley and Atkinson, 1995; Sandvig et al., 2014) to the applied algoritmic work. Sandvig et al. (2014) suggest how audit studies can be used to detect discrimination in algorithms when researchers do not have access to the algorithm itself as is the case with the proprietary algorithm of Facebook (Bruns et al., 2018). We supplement this software and algorithmic centric approach (Manovich, 2013; Sandvig et al., 2014), with a work process centric approach (Bowker and Star, 1999) to account for choices made specifically in relation to classification. With an inspiration in Barocas and Selbst's different discriminatory approaches (2016), Diakopoulos and Koliskas (2017)

Table 1. Steps outlined in the process of applying AI.

Lehr and Ohm (2017)	Diakopoulos and Koliska (2017)	Barocas and Selbst (2016)	Bechmann and Bowker
Definitions and terminology Problem definition		Defining target variable and class labels	Defining task and outcome/target variables and number
Data collection Data cleaning Summary Statistics review Data partitioning	Data (quality, sampling, variables, provenance, volume, assumptions, personal data)	Training data (labeling & data collection)	Data (sample quality, volume, delimiting the datasets (e.g. choosing variables and training/test data splits)
Model selection	Model (input variables, features, target variables, feature weight, model type, software modeling tools, source or pseudo-code, human influence and updates, thresholds)	Feature selection	Model selection (new or pre-trained model, b(l)acklist some outcome variables, set thresholds, pixel or feature selection, feature weight)
Model training (tuning, assessment, feature selection)	Inference (existence and types of inference made, benchmark for accuracy, error analysis, confidence values)	Proxies (repressive correlations)	Data preparation (normalizing data, standardizing data according to model and training setup)
Model deployment	Interface (algorithmic signals, on/off, tweakability of inputs, weights)	Masking	Model training and deployment (Interpret probability scores/accuracy (e.g. cluster semantics, quality of correlations or false negatives/ false positives, reset thresholds and weights, backlist outcome variables, retrain, add targeted or balanced datasets)

and Lehr and Ohm's (2017) accounts of phases in the design process of AI and machine learning, we illustrate how the design of AI and the choices made go through at least five steps, not necessarily in the same order. Every step can loop into a new iteration when results are not satisfying the researcher or developer (Table 1).

In the following, we will outline the five phases and the associated classification questions and potential manipulative/discriminatory processes in two case studies of famous AI models on Facebook data (Bechmann, 2019). On social media classification happens on people or groups of people with the purpose of profiling and subsequently targeting them with advertisement and tailored content (as was the case with Cambridge Analytica). We test two famous algorithms in the setting of such profiling. First, we are interested in identifying the topic of the content people are exposed to in the Facebook News Feed, and secondly, we are interested in predicting uploader gender from a random picture users will upload (not portrait). Both case studies in this article serve as illustrative cases in the setting of classification and not as empirical findings on its own right (Bechmann, 2018; Bechmann and Nielbo, 2018).

Classification in LDA model application processes

The Latent Dirichlet Allocation (LDA) model Text2vec (text2vec.org) is a (seemingly) unsupervised model and a standard model for semantic analysis in textual data. As such we are interested in accounting for how the model can be applied to understand what content people are exposed to in the Facebook news feed to subsequently tailor messages that go viral in a certain population (Figure 1).

Working with LDA models – a case study

Defining task and outcome variables: We want to find the most frequent topics in the Facebook news feed in order to understand what the participants are exposed to. Yet, to do so we need the clusters to be meaningful to us on a semantic level, otherwise the result is useless. Even though we do not set predefined outcome variables, we classify by choosing frequent topics over other topics, force the model to adhere to the logics within a certain (albeit weighted) number of clusters.

***Data*:** We choose a balanced dataset of a sample mirroring the national Danish Facebook population on age, gender and education and their total news feed for 14 days in 2014. We manipulate social categories by using only representative subjects on designed socio demographic categories, albeit e.g. usage patterns could be relevant as a sample category. To adhere to our need for semantic meaningful clusters we need to delimit the dataset significantly as news feed posts can contain status updates, links, comments, shares, likes and photos. We choose to only focus on status updates and links in order not to get too much 'noise' by including comments that will not contribute to the topical recognition. Yet, by doing so we might neglect to include classes of people that prefer to communicate in other types of content such as photos and topical shifts in the conversation (e.g. frame setting) – only taking the initiator into consideration.

Model selection: As we have a clear goal of meaningful clusters, we test the performance of different kinds of LDA models on our data to initially see if clusters make sense to us. As researchers and developers our understanding of 'meaningful' classifiers thereby guide our model selection even though the model itself may be unsupervised. Classes therefore are informed by our a priori understanding of the data.

Data preparation: We are not satisfied with the result of any of the models, but we choose the model that provides most coherent clusters (to us). In order to increase our understanding we clean the dataset. We test whether converting multilingual language into English improves the result. But by doing so we fail to treat nuances in languages on equal terms, thereby favoring the representation of English speaking classes of participants. We test whether reducing the dataset into only nouns that contain topical clues provide more meaningful clusters (to us) which was the case. However, this reduction again favor class representation of participants that use many nouns and back-list people where the topic of the conversation on word-level is not immediately clear (e.g. relational conversations that could be a proxy for females).

Model training and deployment: Because we want to convert high probability scores on words that define this cluster into meaningful labels for the cluster in question, this need to be done by humans. Showing pretty cluster diagrams with labels may look nice but it surely contains a lot of interpretative work to identify more or less similar words into a coherent label or parent class. Doing so means trying to think computationally and to understand why the model chooses to cluster these particular words together and at the same time we tend only to use words that are actually meaningful, otherwise we return to the data preparation phase. The political power that lies in interpreting the probability scores and labeling the clusters is enormous and as a consequence is setting the agenda for what is actually inferred from the data.

Figure 1. A case study of human choices and potential discrimination made through classification when working with LDA models.

As the case shows, a seemingly unsupervised model becomes extremely supervised due to classification work such as setting number of topics, cleaning data in a particular way with an a priori understanding of "meaningful" clusters and interpreting clusters with parent classes manually. Due to the high level of human control in natural language processing models (NLP) already marginalized groups are potentially underrepresented (see also Duarte et al., 2018).

Classification in CNN models application processes

Convolutional Neural Networks (CNNs) are supervised models designed to predict identified outcome classes (here female and male uploader), the seemingly opposite of LDAs and thus interesting to include here. CNNs are standard models for image recognition/processing in Big Data (Bechmann, 2017; Burrell, 2016; Shelhamer et al., 2017). As a social media profiling tool we are interested in training the model to predict uploader gender, feeding the model with only random Facebook pictures labeled with uploader gender. What would the accuracy be with such sparse data and what do the falsely categorized pictures tell us about the classification work of such models? (Figure 2).

As the case study shows in Figure 2 (next page) false negatives *exclude* people from a class, whereas false positives *include* people into the wrong class, e.g. categorizing people as gorillas (Sweeney, 2013). The study is a good example of how developer's choices may discriminate according to stereotypical categories of gender construction because the data in itself show these categorical patterns, and alternative classes and sub-classes are not hardcoded into the algorithm (Caliskan et al., 2017). Both cases show that we can use the steps in the analytical framework to scrutinize for classification choices made in the work processes of applied AI, employing them as a way to account for the human role in seemingly automated knowledge production and potential discrimination on classifiers. Table 2 summarizes some of the classification dispositifs exemplified in the case studies.

Discussion and conclusion: Classification and AI governance

Our analytical framework has shown how seemingly mundane classification processes carry potential discriminatory consequences in a type of *hyper-discrimination* that combine "zero sense discrimination" (between quantities that might appear similar, being able to spell out a meaningful difference between apples, oranges and pears) and "discriminating against" by utilizing persistent patterns of social injustice (masking, redlining, data inferred biases). If returning to the introductory case of Cambridge Analytica such hyper-discrimination is also important to take into consideration here. When we are discriminated into smaller and smaller groups, experimented on through A/B testing, and then acted upon down to a very granular level such hyper-discrimination becomes highly political. Either because such targeting might become a proxy (Barocas and Selbst, 2016; Kroll et al., 2017) for discrimination against race and education level, or for people highly susceptible to manipulation that fall out of our democratic defined protected classes, yet undermines the integrity of the democratic process by providing an unequal approach to elections.

But how do we govern these fundamental issues? The speed by which processing takes place, often in combination with A/B testing, makes it difficult to govern these algorithms and services favor effective and fast models and processes over standards, balancing tests and documentation (see also Kroll et al., 2017). Furthermore, algorithms are protected by proprietary rights. In this article, we have suggested that we need to focus less solely on access to the models and algorithms as technical constructs by themselves and more on documenting the human choices made in the *work process* surrounding AI in order to act sustainable in relation to shared democratic values, avoiding masking, redlining, discriminating biases, and voter discrimination specifically. Such deliberate discrimination is already regulated against but needs to be governed effectively in the application of AI models. This supports the work of Dwork (2006) and Hardt (2011) that distinguishes discrimination as "blatant explicit discrimination", "discrimination based on a redundant coding", and "redlining".

Many critical algorithmic scholars (Bostrom, 2017; Gillespie, 2014; Rogers, 2009; Sandvig et al., 2016) have discussed the issue of algorithms being opaque, arguing that we need to govern the algorithm through a larger degree of transparency so that we know how it sorts information. Kroll et al. disagree, suggesting that it is not a solution to make the rules or source code open and transparent (supported by Ananny and Crawford, 2018). Testing them through audits (Sandvig et al., 2014) and simple random tests would only create ex post analysis, but algorithms, environments, and populations change too quickly between tests in order to use such tests as a benchmark for measuring discrimination. They criticize black-box evaluation of systems as the least powerful of available methods to understand algorithmic

Working with CNN models – a case study

Defining task and outcome variables: We are looking to predict gender in a binary form, knowing that gender negotiations exist (albeit in minority) in many different forms and do not represent a binary output variable. However, choosing to model with, for instance, five output variables would decrease our success in predicting. Already in setting the task, we classify and discriminate against people not defining themselves as males and females.

***Data*:** We use the same population as in the previous LDA case study but now their entire private Facebook photo albums. We deliberately choose to mine only one type of data in order to see if we could create high accuracy from as little as possible instead of including metadata, filenames, geo-location data, demographic data etc. in the first iteration. Hence, discriminating against classes of people that do not upload many images (a total of 340,000 images were collected). We need a sample that is as balanced as possible. This means that we have to know that as many different people as possible are represented in the sample that we choose; further, we need to choose a sample where the potential population of the platform is broad measured against the national census data; even then, such data only take into account differences against specific classical variables and we will not see differences measured against other variables such as skills and interests. We also sort out people without a Facebook account; approx. 20% of the population in Denmark (Bechmann, 2019). Also, we can see from the data that people do not upload images with equal frequency, making the prediction in favor of frequent posting classes.

Model selection: We start testing the models on our data to understand what model we want to move forward with in order to optimize it for this specific task: three of the most widely used open source pre-trained neural networks on semantic data was tested (Shelhamer et al., 2017): Alexnet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), and GoogLeNet (Szegedy et al., 2015). In this way, we deliberately avoid creating a model of our own that has not seen any type of data before, instead employing models that have been trained on massive amounts of images with 1000 output labels (Generic images from the database Imagenet), based on the assumption that a model that has seen images before is better than one that has never seen an image. We forced the model to use prior knowledge from the Imagenet training, but built new knowledge on the basis of our new training data from Facebook to predict binary labels instead (male and female). We do not specify which patterns to look for and the maximum of clusters to coin before deriving at the binary prediction. In this way, the experience of the machine is defined through Imagenet and thereby may discriminate towards patterns and classes not present in this Facebook dataset that is very different from Imagenet.

Data preparation: We need to clean the data and make it balanced in order not to create overfitting or over-dispersed results**.** We make images the same size, but the dataset shows different uploading patterns ranging from people having uploaded only one image apart from the profile picture, to people with 13,125 pictures. We choose to work with data from participants with at least one image upload apart from the profile picture, and the same number of females and males in order not to over-represent females in the dataset. This provides us with 397 unique participants in each category. However, if we had not cleaned the data in this way, we would have had an over-representation of females in the dataset, which may in turn have led the AI to be inclined towards a better understanding of 'female' patterns than 'male' patterns.

Model training and deployment: We run a test on the three algorithms to measure performance on the data after having cleaned for duplicates, providing accuracy of around 66%. In order to increase the accuracy we initially look at the false negatives to see if the pictures deviate from the stereotypical understanding of male and female lifestyles and thereby create an unequal and discriminating understanding of females and males. There are more baby pictures in the male false negative category than in the male true positives, and there are more alcoholic beverages in female false negatives than in female true positives. The result is that we need to adjust the algorithm in order to understand that such 'deviant' subclass is part of the parent class. Interestingly, in doing so, we need to be aware of cultural differences because these false negatives might not be the same across different cultures. For instance in cultures where males do not look after babies and females do not drink. Adjusting according to these false predictions may cause an overfitting to the Danish context and provide a weaker AI when used for the same task but in a different country.

Figure 2. A case study of human choices and potential discrimination made through classification when working with convolutional neural networks.

behavior, but instead suggest that the algorithm ex ante is designed for governance and accountability in a type of what could be labeled value-accountability-by-design.

We believe that the analytical framework suggested in this article with a focus on the human nexus in knowledge production could also be a good starting point for governing against counter-productive

Table 2. Classification in the applied AI and machine learning process.

AI and machine learning design phases	Classification and discrimination exemplified
Defining task and outcome variables	Selecting frequent patterns instead of marginal, setting number of clusters, setting limited number of stereotypic outcome variables
Data	Ignoring specific data as not containing meaning thus limiting model to be predisposed to only certain inputs, balancing sample ignoring classes on the margins and unconventional sampling classes, along with groups outside social media
Model selection	Have a subjective understanding of "meaningful clusters, use pretrained model with undocumented biased 'experience', sub-classifiers, weights and thresholds"
Data preparation	Translate into mono-language input and thereby limit native language nuances, reduce image information that potentially would correlate with marginal classes, taking out under/oversharers that could be of interest in terms of evaluating, e.g. validity of outcome variables or overall task
Model training and deployment	Interpret clusters manually and subjectively in dialogue with model logics, interpret false negatives and false positives to detect non-stereotypic sub-classes to include in retraining

democratic values. In order to make an *ex ante value-accountability-by-design* policy, we need for instance to re-inscribe anti-discrimination classes (see e.g. European Union, 2000; U.S. Equal Employment Opportunity Commission, 1964) beforehand so that it is possible to adjust for them and then bootstrap new categories to enable the machine to test for discrimination against these potential new protected categories. When the machine is "race-blind", "gender-blind" or "income-blind", and the categories deliberately omitted from the processing such as in the Cambridge Analytica case of unsupervised learning, any discrimination against such categories or proxies for such classes cannot be adjusted for in the process (for tests using categories see Kim et al., 2018; Kusner et al., 2017). This is especially important in A/B testing as the fundamental social test of our time in the data-driven society, and such tests are practically ungoverned at the moment (Leese, 2014).

In general, the role of amplification in the algorithms needs to have more attention in the policy work as these principles are nearly entirely unregulated at the moment. Encouraging algorithmic workers (in a broad understanding) through both education and regulation to test for discrimination on, for instance, anti-discriminatory classes moves us away from a populistic programmed consensus truth where discriminatory progressiveness is given towards questions of programmed anti-discrimination as a standard for inclusion. In the race to pursue the "right" and most effective solutions, we need a fair game in which protected classes are in fact protected against correlations of the best predict variable(s). We also need to use AI as a way to protect against potential new discrimination and new, as yet unknown, rising suppressive classes.

Acknowledgments

The authors would like to thank Userneeds, participants who shared their data for the purpose of research, DATALAB developers and assistants Peter Vahlstrup, Ross Kristensen-McLachlan, Henrik Pedersen and Anne Henriksen, Judith Gregory for comments on earlier version of the paper, special issue editors, and anonymous reviewers for their insightful suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Aarhus University Research Foundation (grant number AUFF-E-2015-FLS-8-55) and part of the research was performed while A Bechmann visited Evoke Lab at University of California Irvine.

ORCID iD

Anja Bechmann http://orcid.org/0000-0002-5588-5155

References

Alba D (2016) It's your fault Microsoft's teen AI turned into such a jerk. Available at: https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk/ (accessed 1 May 2018).

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

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*.

Ashby WR (1956) *An Introduction to Cybernetics*. Reprint 1957. London: Chapman & Hall Ltd.

Barocas S and Selbst AD (2016) Big data's disparate impact. *104 California Law Review* 671.

Bechmann A (2017) Keeping it real: From faces and features to social values in deep learning algorithms on social media images. In: *Proceedings of the 50th Hawaii international conference on system sciences*, 4–7 January, Hilton Waikoloa Village, Hawaii, 2017, pp.1793–1801.

Bechmann A (2018) *The Epistemology of the Facebook News Feed as a News Source*. ID 3222234, SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=3222234 (accessed 27 November 2018).

Bechmann A (2019) Inequality in Facebook posting behavior over time: A large-scale data-driven case study of Danish Facebook users. *Nordicom Review*.

Bechmann A and Nielbo KL (2018) Are we exposed to the same "news" in the news feed? *Digital Journalism* 1–13. Epub ahead of print 2018. DOI: 10.1080/21670811.2018.1510741.

Bechmann A and Vahlstrup PB (2015) Studying Facebook and Instagram data: The Digital Footprints software. *First Monday* 20(12): 1–13.

Bostrom N (2017) Strategic implications of openness in AI development. *Global Policy* 8(2): 135–148.

Bowker G and Star SL (1999) *Sorting Things Out*. Cambridge, MA: MIT Press. Available at: https://mitpress.mit.edu/books/sorting-things-out (accessed 28 February 2018).

Bowker GC (2014) Big data, big questions| the theory/data thing. *International Journal of Communication* 8(0): 5.

Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.

Bruns A, Bechmann A, Burgess J, et al. (2018) Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. Available at: https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786 (accessed 9 May 2018).

Buchanan BG (2005) A (Very) brief history of artificial intelligence. *AI Magazine* 26(4): 53.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12. DOI: 10.1177/2053951715622512.

Cadwalladr C (2017) The great British Brexit robbery: How our democracy was hijacked. *The Guardian*. Available at: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (accessed 1 May 2018).

Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Cheney-Lippold J (2017) *We are Data: Algorithms and the Making of Our Digital Selves*. New York: NYU Press.

Citron DK and Pasquale FA (2014) *The Scored Society: Due Process for Automated Predictions*. Rochester, NY: Social Science Research NetworkID 2376209, SSRN Scholarly Paper. Available at: https://papers.ssrn.com/abstract=2376209 (accessed 9 May 2018).

Dewey J (1927). *The Public and its Problems*. Reprint 1946. New York: Holt.

Diakopoulos N and Koliska M (2017) Algorithmic transparency in the news media. *Digital Journalism* 5(7): 809–828.

DiPrete TA and Eirich GM (2006) Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology* 32(1): 271–297.

Duarte N, Llanso E and Loup A (2018) Mixed messages? The limits of automated social media content analysis. In: *Conference on fairness, accountability and transparency*, 21 January 2018, New York University, NYC, pp.106–106. Available at: http://proceedings.mlr.press/v81/duarte18a.html (accessed 26 September 2018).

Dwork C (2006) Differential privacy. In: *Proceedings of the 33rd international conference on automata, languages and programming – Volume part II*, 10–14 July 2006, Berlin, Heidelberg, 2006, pp.1–12. ICALP'06. Springer-Verlag.

Elish MC and Boyd D (2018) Situating methods in the magic of Big Data and AI. *Communication Monographs* 85(1): 57–80.

Eubanks V (2017) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

European Union (2000) *Charter of Fundamental Rights of the European Union*. C 364/01. Available at: http://fra.europa.eu/en/charterpedia/article/21-non-discrimination (accessed 1 May 2018).

Farr W (1985) *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr, M.D., D.C.L., C.B*. London: Offices of the Sanitary Institute.

Foucault M (1971) *The Order of Things*. New York, NY: Pantheon Books.

Gasser L (1986) The integration of computing and routine work. *ACM Transactions on Information Systems* 4(3): 205–225.

Gillespie T (2014) The relevance of algorithms. In: *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press.

Hammersley M and Atkinson P (1995) *Ethnography: Principles in Practice*. London: Routledge.

Hardt MAW (2011) *A study of privacy and fairness in sensitive data analysis*. PhD Thesis. Princeton University, Princeton, NJ, USA.

Hayles NK (2017) *Unthought*. Chicago, IL: The University of Chicago Press.

Howard PN (2005) *New Media Campaigns and the Managed Citizen*. Cambridge: Cambridge University Press.

Jaton F (2017) We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811–840.

Kanda T, Ishiguro H, Imai M, et al. (2004) Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE* 92(11): 1839–1850.

Kim MP, Ghorbani A and Zou J (2018) Multiaccuracy: Black-box post-processing for fairness in classification.

arXiv:1805.12317 [cs, stat]. Available at: http://arxiv.org/abs/1805.12317 (accessed 27 November 2018).

Kosinski M, Stillwell D and Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15): 5802–5805.

Krizhevsky A, Sutskever I and Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25* (eds F Pereira, CJC Burges, L Bottou, et al.), 2012, pp.1097–1105. Curran Associates, Inc. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (accessed 9 May 2018).

Kroll JA, Huey J, Barocas S, et al. (2017) Accountable algorithms. *University of Pennsylvania Law Review* 165: 633–699.

Kusner MJ, Loftus JR, Russell C, et al. (2017) Counterfactual fairness. Available at: https://arxiv.org/abs/1703.06856 (accessed 16 November 2018).

Latour B (2002) Gabriel Tarde and the end of the social. In: Joyce P (ed.) *The Social in Question: New Bearings in History and the Social Sciences*. London: Psychology Press.

Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 495–511.

Lehr D and Ohm P (2017) Playing with the data. *UC Davis Law Review* 51: 653–717.

Levin S (2016) A beauty contest was judged by AI and the robots didn't like dark skin. *The Guardian*. Available at: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people (accessed 1 May 2018).

Lomborg S and Bechmann A (2014) Using APIs for data collection on social media. *The Information Society* 30(4): 256–265.

Manovich L (2013) *Software Takes Command*. New York, NY: Bloomsbury Academic.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1st ed. New York, NY: Crown.

Polanyi M (1966) *The Tacit Dimension*. (Reprint 2009). Chicago, IL/London: University of Chicago Press.

Rieder B (2017) Scrutinizing an algorithmic technique: The Bayes classifier as interested reading of reality. *Information, Communication & Society* 20(1): 100–117.

Rogers R (2009) Post-demographic machines. In: *Walled Garden*. Amsterdam: Virtueel Platform, pp. 29–39.

Roth JA (1966) Hired hand research. *The American Sociologist* 1(4): 190–196.

Russel S and Norvig P (2010) *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Pearson Education Inc. / Prentice Hall.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. In: *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Seattle, WA, 2014, pp.1–23.

Sandvig C, Hamilton K, Karahalios K, et al. (2016) When the algorithm itself is a racist: Diagnosing ethical harm in the basic components of software. *International Journal of Communication* 10(0): 4972–4990.

Searle JR (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–424.

Shelhamer E, Long J and Darrell T (2017) Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4): 640–651.

Shirky C (2005) Ontology is overrated: Categories, links, and tags. Available at: http://shirky.com/writings/herecomeseverybody/ontology_overrated.html (accessed 1 May 2018).

Simonyan K and Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. In: *arXiv:1409.1556 [cs]*, 2014. Available at: http://arxiv.org/abs/1409.1556 (accessed 9 May 2018).

Slezak P (1989) Scientific discovery by computer as empirical refutation of the strong programme. *Social Studies of Science* 19(4): 563–600.

Star SL (1990) Power, technology and the phenomenology of conventions: On being allergic to onions. *The Sociological Review* 38(S1): 26–56.

Strauss AL, Fagerhaugh S, Wiener C, et al. (1985) *Social Organization of Medical Work*. Chicago, IL: University of Chicago Press.

Sweeney L (2013) Discrimination in online ad delivery. *Communications of the ACM* 56(5): 44–54.

Szegedy C, Liu W, Jia Y, et al. (2015) Going deeper with convolutions. In: *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, 7–12 June, Boston, MA, 2015, pp.1–9.

Tufekci Z (2014) Engineering the public: Big data, surveillance and computational politics. *First Monday* 19(7) Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4901.

Turing AM (1950) Computing machinery and intelligence. *Mind* LIX(236): 433–460.

U.S. Equal Employment Opportunity Commission (1964) *Title VII of the Civil Rights Act of 1964*. Pub. L. 88-352. Available at: https://www.eeoc.gov/laws/statutes/titlevii.cfm (accessed 1 May 2018).

Walford A (2014) *Performing data: An ethnography of scientific data from the Brazilian Amazon*. PhD Thesis, IT University Copenhagen, Denmark.