Democratic Frontiers

Algorithms and Society

Edited by
Michael Filimowicz

ISBN13: 978-1-032-00267-5

First published 2022

Chapter 3

From Big to Democratic Data

Why the Rise of AI Needs Data Solidarity

Mercedes Bunz and Photini Vrikki

The chapter DOI: 10.4324/9781003173427-3

The Research for this article has been funded by the Wellcome Trust Grant n° 213552/Z/18/Z

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

3 From Big to Democratic Data

Why the Rise of AI Needs Data Solidarity

Mercedes Bunz and Photini Vrikki

Digital technologies and their processing of data have transformed our cultural, social and working lives through expansive digital connections and networks, allowing us to undertake social, cultural and economic transactions that shape global and local communities. This digital space is a sphere in which users interact, thereby creating data, which is then collected and analyzed shaping their societal possibilities through recommendations or algorithmic decision-making. Yet, paradoxically, in spite of the ubiquitous reach of our digital condition, the *political force* within data shaping our societies is only in parts understood. One reason for this is that the notion of "big data" at the beginning of the 21st century has been conceptualized *by* businesses and *for* the business world, as Puschmann and Burgess (2014) have shown. Given the significance of data in our public and everyday lives, many find the strong, confining link between data and business alarming; this is even more so, since data has gained societal and political importance through further technical developments in areas such as artificial intelligence (AI). As we will show in this text, recent advances in AI, particularly in the area of machine learning (ML; in which systems are trained on huge datasets), have opened up new possibilities for data analysis that have further strengthened the societal role of data in our political and social lives. This is why data needs to be understood more than ever not just as an economic opportunity but also as a *democratic frontier.*

When discussing data from the perspective of democracy, next to the rights of the individual and the effect of data on the individual, the effect of data on the collective, i.e., the shaping of a society, comes into view. Recently, a range of scholars have started to explore this *collective value of datasets* systematically and have shown that value for populations can be gained from insights into data relations emerging between individual data entries (Viljoen, 2021). This point is important,

DOI: 10.4324/9781003173427-3

as it highlights the power datasets have to drive benefits for societies (and not just companies), widely known as "data for the public good", which some argue could be governed by independent data trusts (see Delacroix, Pineau & Montgomery, 2021), a construct that is somewhat linked to the notion of digital commons (Dulong de Rosnay & Stalder, 2020). Such research into data trusts or digital commons stresses the collective value of data and calls for revisiting the principles of data governance, i.e., the processes that manage the *availability*, *usability* and *security* of data. Among these three aspects, it was the latter, the aspect of *security* and loss of privacy leading to a growing surveillance (Zuboff, 2019) that, at the beginning of the 21st century, gained most public attention with some positive effects. A variety of governments have tackled this issue by legislation amendments, one of the most far-reaching being Europe's General Data Protection Regulation (GDPR). The principles of *availability* and *usability*, however, were likewise discussed beyond experts and data science. Both principles have gained the attention of data activists, non-governmental organizations (NGOs) and even politicians – an attention that is now newly required. In his excellent genealogy of Open Data, Jonathan Gray (2014) has shown the wide range of initiatives Open Data has surfaced, from neoliberal takes to widening civic participation. Among them, we find calls for:

- opening data in a push for transparency to hold the public sector to account;
- reducing government by transforming it to a platform service;
- making data available that could be useful for businesses fostering economic growth and innovation;
- allowing citizens to reuse their data and/or to make their data portable from one platform to another; and
- making use of data to advance societal issues through civic hacking.

While the benefits and drawbacks of the above points are still being discussed, the focus on opening up data has recently shifted. This shift is an effect of two, at times overlapping, strands of research transforming data analysis profoundly: (1) the growing body of critical research into the bias of datasets and (2) the development of data analytics through the method of ML. Both strands put new attention on the *quality* of datasets, which has not only become essential but also opens up room for the creation of datasets as a societal tool with strong political potential, which is the focus of this chapter. And while there is a growing body of ethics codes in different domains (Stark &

Hoffmann, 2019) as well as calls for "data infrastructure literacy" (Gray et al., 2018), computational science has far too long neglected to focus on questions about the creation, composition and processing of data. In other words, despite calls to move toward critical data studies (Iliadis & Russo, 2016), much of our data practice, particularly regarding ML, has been kept invisible. Our chapter will show how this *invisibility*, which endangers the quality of data, could be challenged if we deployed *data solidarity* as a principle of governance for the creation of datasets; a principle that could help governments and corporations understand datasets not just as economic opportunities but also as democratic resources that offer possibilities to advance the public good.

On the Link between Data Quantity and Data Quality

Ever since digital technologies have transformed data to become what has been called "big data" (Kitchin, 2014) – i.e., extremely large data sets that can be analyzed computationally to reveal patterns, trends and associations – new opportunities but also profound challenges regarding the quality of datasets arose. Data has become a resource of social life leading to digital technology and sociality becoming tightly interwoven, at times inseparable (Marres, 2017: 7–44). With this, substantial problems around the quality of datasets and their handling became apparent and have started to be discussed by a wide range of scholars. Contributing to critical data science, danah boyd and Kate Crawford (2012: 666 and 668) have, e.g., shown that bigger data is not automatically better data and that early "claims to objectivity and accuracy" were misleading. Ruha Benjamin (2019: 127) has pointed out that datasets are often "naturally occurring" within digital industries and are therefore taken from contexts that "reflect deeply ingrained cultural prejudices and structural hierarchies". The far reach of those ingrained prejudices was further elaborated by Wendy Chun (2021: 17), who showed in her excellent study *Discriminating Data* that even when ML algorithms do not officially include race as a category, unbalanced datasets embed whiteness as a default. Besides racial bias, the digital sphere is also haunted by class (Schradie, 2011) and gender gaps, the latter exposed by Caroline Criado Perez (2019) describing the discrimination against women through data as systemic as there is an invisible bias with a profound effect on women's lives (e.g., there have long been life-threatening knowledge gaps within medical data about women's heart attacks which manifest in slightly different symptoms from men's on whom the research of this disease was long

focused). Catherine D'Ignazio and Lauren F. Klein (2020) have also made a strong case showing example after example how profoundly data science needs feminism. Many of the above studies are interdisciplinary, drawing on important works within Computer Science such as the critical study into word embedding in natural language processing (Bolukbasi et al., 2016) or into the bias of large language models (Bender et al., 2021).

Being aware of such problems when gathering datasets or working with data is even more important in the face of ML developments advancing the capabilities of AI, which has widened the societal reach of data analysis. While cookies and other digital data traces allow for the predictive modeling of user data, i.e., informing conclusions and making predictions about those users, ML goes a step further. It can make predictions about users from *indirect* information, i.e., it is less dependent on data directly left about and by users. This is because of its new analytic capacity to process language, images or other symbols. Computational approaches to analyze these had long failed to succeed until ML using so-called "deep neural networks" allowed a breakthrough regarding the "calculation of meaning" (Bunz, 2019; Cantwell Smith, 2021), meaning which users accidentally leave behind when speaking, writing or appearing in photos or videos. Processing these formats and calculating meaning signified by them is a new capacity of data analysis that substantially widens the data pool as it allows reaching out much wider in the analysis of user information. The effect of ML is therefore a profoundly deeper reach of digital technology into the fabric of our societies, thereby affecting its social and political processes.

To gain this reach, large datasets featuring our audio, video, photos or written texts are used to train ML systems, whereby the configuration and quality of data plays an essential role to train them correctly. At the same time, there has been a lack of attention regarding datasets due to the fact that in computer science, their creation (such as, e.g., the ImageNet dataset; Deng et al., 2009) has long been valued less than the making of algorithms or the building of models. The reason for this is that gathering or acquiring a dataset is, strictly speaking, not a computational procedure. Many introductory books teaching ML in computer science assume datasets as already available (e.g., Alpaydin, 2020: 154; Flach, 2012; Witten et al., 2011) making their creation an "invisible practice". However, acquiring a dataset for training is *fundamental to the development of machine learning model*s, which is why critical knowledge about the quality of data needs to become a standard in practices, from the conception to deployment of ML. While data is not

a computational procedure, the actual workflow when constructing a neural network to perform ML begins with the acquisition of a dataset, as Jaton showed in great detail in his ethnographic study of a computer science laboratory (2021: 54): for ML models, obtaining a dataset is part of "the practical processes that enable them to come into existence" (11). In other words, datasets are essential to train ML models; an observation that in 2021 led Andrew Ng, Professor at Stanford University Department of Computer Science and Electrical Engineering, to call for a more "data-centric AI" (Ng 2021). High-quality data, however, is not sufficiently publicly available to ML developers, and this is often highlighted as one of the biggest issues in the field. The essential role described here for datasets and their quality regarding ML, and with that the even bigger importance datasets have come to play in the technical and political realities of our overdeveloped world, creates the need for a different approach toward data: an approach that needs to engage with the issues of critical data science (Iliadis & Russo, 2016) in face of the fact that processing data creates and deprives opportunities. By revealing the absences, differences and disconnects within datasets, we can address some of the sociocultural problems they create. These issues show why a critical conceptualization that aims to make data more fair, transparent, available and accountable for the community is needed so we can think of "data as a public good".

The concept of "data as a public good" has been developed as a response to the massive deployment of data analytics by technology companies such as Google or Palantir. As Lane et al. (2014) point out in the introduction of *Privacy, big data, and the public good*, one of the first books on this topic:

> Much has been made of the many uses of (…) data for pragmatic purposes, including selling goods and services, winning political campaigns, and identifying possible terrorists. Yet big data can also be harnessed to serve the public good in other ways: scientists can use new forms of data to do research that improves the lives of human beings; federal, state, and local governments can use data to improve services and reduce taxpayer costs; and public organizations can use information to advocate for public causes, for example.
> (Lane et al., 2014: XI)

However, in an increasingly datafied world, the systemic and structural inequities we described earlier are intensified and exacerbated by narrow conceptions of how datasets are produced, reproduced, combined and shared. Data structures and data processes such as the

building of new datasets through other datasets, the combination of data etc. (see Roberts et al., 2021) are *invisible processes* that impact every decision that is taken based on their analysis. And these *invisible data processes*, mounded on existing systemic and structural inequities, can have profound societal consequences. In other words, *invisible data processes*, such as non-accessible, non-structured, non-available or misrepresented, incomplete or biased data often impact specific populations and countries, and are a threat to the health and safety of the global public. As Roberts points out, invisibility is "a metaphor that figures a state of being that comes into existence when others refuse to see us, to acknowledge our existence, to accept our presence as making a contribution to a world of meaning" (Roberts, 1999: 121). He goes on to argue that invisibility is not just created systemically and structurally, but it is also sustained through the complicity of those who are invisibilized – and this is why data solidarity, as we are going to show, is so important. Applying this logic to the invisibilization of data, it becomes clear that if we act as if data processes are visible, we perpetuate this invisibilization and sustain the power structures that suppress and marginalize data and their societal impact. How can we balance the fears of data/public control from Big Tech with the significance of data for the betterment of sectors such as healthcare? A challenge that translates into: how can we do good with better and more data? By now, several definitions aim to conceptualize this different political approach to data ranging from data justice (Dencik et al., 2016), responsible data (van der Aalst et al., 2017) to the call for data trusts (Delacroix, Pineau & Montgomery, 2021). To this, we would like to add the concept of *data solidarity* and the need to overcome the invisibility of data practice. In the following, we will demonstrate a need for this through a case study.

Case Study: On the Role of Datasets for Machine Learning Research

To understand the importance of data processes and cut through their invisibility, we studied the role datasets have for ML research in healthcare, particularly the usage of patient data to train ML systems. Taking advantage of the abundance of ML models being trained and developed within healthcare, we conducted a systematic literature search focused on medical diagnosis on arXiv; arXiv, hosted by Cornell University, was chosen as it plays a central role for the publication of research by the ML community (Balki et al., 2019). Established in 1990, the repository is generally a popular place of prepublication for

science, technology, engineering and mathematics (STEM) disciplines as it has a fast publishing turnaround getting papers out before peer review (Delfanti, 2016); the pace in which ML research develops created the need for researchers to get their findings out quickly. Our systematic literature research focuses on a very specific area – that of ML models assisting with medical diagnosis. On arXiv, 82 relevant studies were identified by searching "machine learning", "medical", "diagnostics". One duplicate was removed with the use of reference management software. The remaining papers were included if they met the criteria of describing a ML experiment in a scientific paper that involved processing medical data entries. This led to a corpus of 62 papers published between 2009 and 2021 that were analyzed in detail regarding their usage of data. Our aim was to learn more about the medical datasets used when training and validating ML models, a process that is in parts invisibilized – while datasets are mentioned, their creation is often treated as negligible. The focus was therefore on the origin and the creation of a dataset, including the gathering and (in some cases for supervised learning) on the labeling of data, information that at times is communicated in the margins (through acknowledgements, affiliations, etc.). Cleaning of existing datasets was not taken into account. Datasets mentioned in the papers were coded according to three categories: Code N for *newly created datasets*; code L for *datasets that had to be labeled* by medical experts to allow for supervised learning; code P for *publicly available datasets*.

We found that over half of the experiments, 33 papers, worked with publicly available datasets, i.e., medical datasets that have been published to foster research such as the National Institutes of Health Chest X-Ray Dataset published by the National Library of Medicine in the US, or the Alzheimer's Disease Neuroimaging Initiative (ADNI). Six further experiments used datasets of mixed status, i.e., some were publicly available while others were specifically created for the study. This procedure reflects the process of training a ML system, which runs through two or three interlinked phases each needing separate datasets – the phase of training the ML model (1) and of testing the model (2b); some also validate the model with a step in the middle adjusting parameters further (2a). About a third, 21 experiments, created their own dataset from the ground up; all but one through a close collaboration with a medical institution.

Even though the findings of this systematic review are not representative, they clearly show a strong tendency within ML research: The majority of experiments, 33 out of 62 papers, used publicly available datasets. Adding the six experiments that made use of available

datasets while enriching them with newly created ones, one could come to the conclusion that 39 papers, i.e. 63% of the papers we reviewed, worked with available datasets. Given the fact that publicly available datasets are rare, this clearly shows the extent to which datasets incentivize and influence the conducting of ML research – they are obviously needed. And this is the case for academic research as well as for businesses. Among our body of 62 papers were six in which businesses led the research or were part of it – some big ones such as Google Brain or Microsoft Research plus a range of less well-known, smaller companies. Most of them were working with available datasets: among the 39 papers using publicly available datasets, five were conducted by businesses or in collaboration with businesses. Only one paper, for which academics collaborated with the British company Babylon Health, used a newly created dataset, most likely one Babylon Health held internally.

The demand for publicly available datasets clearly shows their potential. Datasets strongly incentivize both academic and commercial research. Despite the talk of big data, however, they are scarce – platforms such as Kaggle, which allows users to find and publish data sets and was bought by Google in 2017, lists 50,000 datasets for more than one million active users. This indicates that in 2021, too many users conducting data analysis research worked with the same datasets, which our analysis confirmed. A dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) was used five times in papers from Russia, France, US, Pakistan and China. Overall, publicly available datasets such as ADNI or the chest X-ray datasets published by the National Institute of Health and others led to multiple papers using them. Papers frequently mentioned that "progress has been hindered by a sparsity of available training data, commonly attributed to the difficulty of publishing datasets" (McManigle et al., 2020: 1) or noted that "in domains where data is highly regulated and expert time is rare, it can be exceedingly cumbersome to obtain new expert-labeled data sets every time a model needs to be improved" (Cai et al., 2019: 12). As Roberts et al. (2021) have also pointed out, the need for public data leads to serious issues for research. More and more datasets are "assembled from other datasets and redistributed under a new name". These "Frankenstein datasets" may inadvertently include overlapping or identical datasets, which, in turn, lead algorithms to wrong diagnoses and suggestions.

The scarcity of data and the invisible data processes that produce datasets lead to working with unbalanced datasets – an issue that impacts not just the medical but all sectors, and with it, society. While

data is abundant, the majority of datasets are proprietary and built for commercial reasons with no oversight. At the same time, as demonstrated by the high number of Kaggle users compared to the low number of public datasets, publicly available datasets are generally scarce. While the issue is known, the low regard for the creation of datasets, which, as we have shown, is often not seen as an act of computer engineering and not taught in introductory ML books, makes the need for public datasets pertinent.

This is where an approach foregrounding the democratic value of data and an initiative to create datasets making them publicly available out of a gesture of *solidarity* could help. This is even more the case, as in current debates, the focus on the *collective value* datasets have for society is often missing (Delacroix & Montgomery, 2020; Viljoen 2021). This is worrying as data analysis, driven further by ML, has become a process people experience directly or indirectly everyday: when shopping on the internet, when using government services or when applying for a loan or an insurance. As long as these data analysis decisions are based on commercial datasets without checks and balances and to which there is no alternative, there will be issues of bias and fairness leading to a lack of trust. This importance of taking the collective value of data into perspective has been demonstrated by Salome Viljoen (2021). In her in-depth report on the issue titled "Democratic Data", she correctly reminds readers:

> The data economy has resulted in massive collection of information regarding consumer purchasing preferences and social networks, for instance, but has contributed comparatively little to ongoing discussions concerning waste production, water usage, or how wealth from financial instruments flows globally.
>
> (649)

With the understanding of big data as something mainly useful for business, data to support our democratic public infrastructures needs further strengthening. Admittedly, the change needed here is not just infrastructural, it is also political. A democratic use of data could tackle bias in datasets and handle it more transparently; it could turn toward opportunities such as programmes to create datasets in under-researched areas that are socially relevant or help us understand niche issues that have been consistently ignored due to lack of corporate or government interest. As Viljoen points out: "Datafication is not only unjust because data extraction or resulting datafied governmentality may violate individual autonomy; datafication may also be

unjust because it violates ideals of social equality" (58). Viljoen calls for a shift in the understanding of data "from an individual medium expressing individual interests, to a democratic medium that materializes population-level, social interests" (54). This would also mean the following:

- data does not only need to be gathered where it naturally occurs, instead governments need to start initiating the collection of datasets to ensure democratic values;
- datasets could be used to allow citizens a better representation in the conditions and purposes of data production;
- issues with bias in datasets can be targeted or made transparent;
- datasets could be used to incentivize ML research in particular areas attractive from a societal and not commercial perspective.

These points, however, depend on the availability of data and the willingness of citizens to embrace data sharing for the public good. Naturally, the gaining of data for the public good operates differently from the commercial top-down approach leading to data extraction. Instead, it must revolve from a participatory understanding of data sharing and a belief in "data commons". This needs communicative work. As Dulong de Rosnay and Stalder have (2020: 16) pointed out:

> The constitution of data commons (…) needs to overcome the apparent contradiction between personal data and property, and between privacy and open access, as a personal data commons would not lead to sharing personal information, but to govern their reuse according to values of the digital commons.

This brings the importance of solidarity, exercised by giving data to support the community, to the fore.

Toward Data Solidarity

In order to develop not just fair and transparent but also democratic and visible data processes, we propose that we need to cultivate and sustain a culture of solidarity in data sharing processes. Solidarity has functioned as a key principle in democratic struggles of the past, such as the labor union Polish Solidarność of the 1980s, the mid-19th-century French workers' fight against oppression (Wilde, 2013) and in the most recent past, in social movements such as Occupy (Vrikki, 2018). In social movements, solidarity visibilizes and materializes values such as

trust, openness and common principles (Pavan & della Porta, 2020). In the data era we currently experience, living with others and the social construction of our societies have given solidarity a wider role that does not just hold political importance, it can also be perceived as a form of caring and protecting others (Chatzidakis et al., 2020). At the same time, this can build on interpretations of solidarity in social theory where one finds, on the one hand, interpretations that perceive solidarity as the sum of norms contributing to social cohesion, e.g., in the works of Emilé Durkheim (1984, 2001), and on the other hand, one finds interpretations that deduce solidarity as a relationship between members of a group with common interests, referring to the works of Marx (1906) and Weber (1978). Beyond social theory, political philosopher Scholz (2008) has identified three kinds of solidarity: social solidarity (describing the relationship between the group), civic solidarity (referring to the relationship between citizens and the state) and political solidarity (expressing the commitment and morals of the individual), which divide solidarity based on the relationships onto which it depends on. The variety of approaches within social and political theory shows how ingrained solidarity is in our social, political and cultural lives that in everyday life gets often translated as the process of supporting the vulnerable, as acts of public caring such as education, welfare and healthcare and as the primary care relations we build and sustain through friendships, households and families (Lynch, 2007).

Building on these interpretations and approaches, we identify *data solidarity* as an articulation of *visibilizing data processes* for the benefit of public good. The proposition here is to perceive data solidarity in a productive opposition to current hierarchical data structures as well as to the latent processes of the neoliberal market, personal responsibility and individual agency (Cohen, 2010). This is pertinent to the conceptualization of data processes as a set of democratic norms that together reinforce the capacity of communities to produce collective goods for the public benefit (Laitinen & Pessi, 2014). Recent critical studies into the democratisation of AI, for example by Himmelreich (2021), have stressed that the matter is complicated and that there is no simple administrative panacea to the injustices that are perpetuated by AI. Attention towards ways, in which democratic governance of AI can be initiated and structured, are still underdeveloped. Informed by these reasons, we propose 'data solidarity' as a value supporting a process to enhance our AI futures in the same way solidarity between working class and farmers resulted in the establishment of a universal pension system (Baldwin, 1990). Data solidarity can advance the

inclination of corporate and public data stakeholders to share both the risks and the benefits of data access, production and sharing. The term solidarity is "sometimes used as a nebulous concept" (Stjernø, 2009: 2), but data solidarity can most conducively be defined as the willingness to share datasets and resources with others while acknowledging the invisible processes that take place during the creation, production and sharing of datasets. Visibilizing those processes and their flaws that may result in marginalizations such as racism, sexism and classism accentuate the need for a collective action that will be based on the values or solidarity.

Conclusion: Moving from Big to Democratic Data

In the same ways in which our political and financial systems have determined so much of our behaviors and societies, data analytics are and will keep stretching our cultures and democracy. In this chapter, we aim to answer this challenge by making the political force of data practices visible. Our argument positions itself as an addition to the ongoing debate about *critical data practice*, which aims "to account for, inventively respond to and intervene around the socio-technical infrastructures involved in the creation, extraction and analysis of data" (Gray et al., 2018: 8). Our research also builds on recent insights into collective aspects regarding datasets (Delacroix, Pineau & Montgomery, 2021; Viljoen, 2021), insights that (a) are gained *from the collective*, i.e., from relations between data entries, and could (b) be processed *for the collective* advancing the public good. To advance this, the tendency to shroud data practice in invisibility needs to end. To move from big to democratic data, we need to understand datasets and data infrastructure as democratic tools which can advance societal interests and assist with bringing forth elements of public good for populations. How influential publicly available datasets are, could be seen in our case study of medical diagnosis through ML systems trained on medical data. To encourage the building of such publicly available datasets, we need a new notion of data: next to the understandable fear about surveillance through the extraction of data, we need to stress the potential that data sharing has in public hands and move toward data solidarity. While there is no simple answer to the question "how can we do good with better and more data?", we know that ultimately it boils down to collective action. By deploying *solidarity as a principle of data governance* for the creation of publicly held datasets, we can start building trust and accountability. Digital technologies, AI systems such as ML and other advanced data

analytics can help us better our societies if we deploy principles of critical data practice that visibilize data processes and apply a critical approach to datasets aiming for the inclusion of different kinds of data. As we stand at the precipice of datafied democracy, now is an opportunity for a steady refocus on how data and data infrastructure can support inclusion. The data infrastructures we shape, shape us in return. The rise of AI has made these infrastructures even more important. To shape these infrastructures according to democratic values, the principle of data solidarity is essential.

We would like to express our thanks to Shuprima Guha, Jonathan Gray and Adam Bull for their useful comments and corrections.

Bibliography

Alpaydin, E. (2020). *Introduction to machine learning*. Cambridge, MA: MIT press.

Baldwin, P. (1990). *The politics of social solidarity: Class bases of the European Welfare State, 1875–1975*. Cambridge: Cambridge University Press.

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., Moody, A. R., & Tyrrell, P. N. (2019). "Sample-size determination methodologies for machine learning in medical imaging research: A systematic review." *Canadian Association of Radiologists Journal*, 70(4): 344–353.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610–623.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new jim code*. Cambridge: Polity.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv*:1606.06121.

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

Bunz, M. (2019). The calculation of meaning: On the misunderstanding of new artificial intelligence as culture. *Culture, Theory and Critique*, 60(3–4): 264–278.

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., & Terry, M. (2019, May). Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Cantwell Smith, B. (2021, April). The foundations and metaphysic of computing [draft]. NYU Digital Theory Lab.

Chatzidakis, A., Hakim, J., Littler, J., Rottenberg, C., & Segal, L. (2020). From carewashing to radical care: The discursive explosions of care during Covid-19. *Feminist Media Studies*, 20(6): 889–895.

Chun, W. H. K., (2021) *Discriminating data*. Cambridge, MA: MIT Press.

Cohen, C. J. (2010). *Democracy remixed: Black youth and the future of American politics*. New York: Oxford University Press.

Delacroix, S., Pineau, J., & Montgomery, J. (2021). Democratising the digital revolution: The role of data governance. In: Braunschweig, B. & Ghallab, M. (ed.) *Reflections on AI for Humanity* (pp. 1–15). Cham: Springer.

Delacroix, S., & Montgomery, J. (2020). From research data ethics principles to practice: Data trusts as a governance tool. *Available at SSRN 3736090*.

Delfanti, A. (2016). Beams of particles and papers: How digital preprint archives shape authorship and credit, *Social Studies of Science*, 46(4): 629–645.

Dencik, L., Hintz, A., & Cable, J. (2016). Towards data justice? The ambiguity of anti-surveillance resistance in political activism. *Big Data & Society*, 3(2): 2053951716679678.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. Cambridge, MA: MIT press.

Dulong de Rosnay, M., & Stalder, F. (2020). Digital commons. *Internet Policy Review*, 9(4): 1–22, http://dx.doi.org/10.14763/2020.4.1530

Durkheim, E. (1984 [1983]). *The division of labor in society*. Translated by Q. D. Halls with an Introduction by Lewis Coser. New York: Free Press.

Durkheim, E. (2001[1912]). *The elementary forms of religious life. A new translation by Carol Cosman*. Oxford: Oxford University Press.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Gray, J. Towards a genealogy of open data (September 3, 2014). The paper was given at the *General Conference of the European Consortium for Political Research in Glasgow*, 3–6th September 2014. http://dx.doi.org/10.2139/ssrn.2605828

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*. https://doi.org/10.1177/2053951718786316

Himmelreich, J. (2021). Against 'democratizing AI'. Forthcoming in *AI & Society*. https://johanneshimmelreich.net/papers/against-democratizing-AI.pdf

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716674238.

Jaton, F. (2021). *The constitution of algorithms: Ground-truthing, programming, formulating*. Cambridge, MA: MIT Press.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1) https://doi.org/10.1177/2053951714528481.

Laitinen, A., & Pessi, A. B. (2014). *Solidarity: Theory and practice. An introduction*. Minneapolis, MN: Lexington Books.

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement.* New York: Cambridge University Press.

Lynch, K. (2007). Love labour as a distinct and non-commodifiable form of care labour. *Sociological Review*, 54(3): 550–570.

Marres, N. (2017). *Digital sociology: The reinvention of social research.* Cambridge, UK: Polity Press.

Marx, K. (1906). *Capital*, Vol. 1. Translated by S. Moore, E. B. Aveling and E. Untermann. New York: Modern Library.

McManigle, J. E., Bartz, R. R., & Carin, L. (2020, July). Y-Net for Chest X-Ray preprocessing: Simultaneous classification of geometry and segmentation of annotations. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1266–1269). *IEEE*.

Ng, A. (2021) MLOps: From model-centric to data-centric AI. *Deeplearning.ai* https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf

Pavan, E., & della Porta, D. (2020). Social movements, communication, and media. In Lievrouw, L. A., & Loader, B. D. (Eds.). *Routledge Handbook of Digital Media and Communication* (pp. 307–318). New York: Routledge.

Perez, C. C. (2019). *Invisible women: Exposing data bias in a world designed for men*. London: Penguin Random House.

Puschmann, C., & Burgess, J. (2014). Metaphors of big data. *International Journal of Communication*, 8: 1690–1709.

Roberts, J. W. (1999). "… Hidden right out in the open": The field of folklore and the problem of invisibility 1998 American Folklore Society presidential address. *Journal of American Folklore*, 112(444) (Spring, 1999): 119–139.

Roberts, M., Driggs, D., Thorpe, M. et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Natural Machine Intelligence*, 3: 199–217.

Scholz, S. J. (2008). *Political solidarity*. Pennsylvania, USA: Penn State Press.

Schradie, J. (2011). The digital production gap: The digital divide and Web 2.0 collide. *Poetics*, 39(2): 145–168.

Stark, L., & Hoffmann, A. L. (2019). Data is the new what? Popular metaphors & professional ethics in emerging data culture. *Journal of Cultural Analytics.* https://doi.org/10.22148/16.036

Stjernø, S. (2009). *Solidarity in Europe: The history of an idea*. New York: Cambridge University Press.

van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering*, 59: 311.

Viljoen, S. (2021, November). A relational theory for data governance. *Yale Law Journal*, 131(2), 573–645.

Vrikki, P. (2018). The beginning of the end: Telling the story of Occupy Wall Street's eviction on Twitter. In Adi A. (ed.) *Protest public relations: Communicating dissent and activism* (pp. 76–93). London: Routledge.

Weber, M. (1978). *Economy and society: An outline of interpretive sociology* (Vol. 1). Berkeley: University of California Press.

Wilde, L. (2013). *Global solidarity*. Edinburgh: Edinburgh University Press.

Witten, I., Frank, E., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques*. Amsterdam: Elsevier.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs.