universität wien

Exposé

Arbeitstitel der Dissertation

Data-based algorithmic systems and individuals

Verfasserin

Mag.[a] Paola Lopez

Betreuer*innen

Univ.-Prof. Dr. Elisabeth Holzleithner

Ass.-Prof. Dr. Ben Wagner (TU Delft)

Angestrebter akademischer Grad: Doctor of Philosophy (PhD)

Studienrichtung laut Studienblatt: Doktoratsstudium Interdisciplinary Legal Studies

Studienkennzahl: UA 794 242 101

Wien, März 2022

Table of Contents

Introduction

Artificial Intelligence (AI), Machine Learning, and "algorithms" as socio-technical phenomena are currently experiencing a critical moment of political and social discussion. Enthusiastic standpoints hope for AI and related techniques to transform a variety of processes in which human performance is considered to be flawed. The potential areas of deployment of AI are as diverse as: the early diagnosis of illness and, therefore, the prevention of severe courses of disease (see, e.g., Goldenberg et al., 2019; Jiang et al., 2017), "combat[ting] climate change" (Cowls et al., 2021, p. 1), realizing projects of sustainable smart cities (Singh et al., 2020), and revolutionizing manufacturing (Li et al., 2017). AI can be found in official innovation agendas (Wiesmüller et al., 2018, p. 1), and data-driven methods are supposed to help eliminate unconscious human biases, as well as "noise" (Kahneman et al., 2021), in human decision-making. The striving for algorithmic innovation, however, is met with severe accusations of discrimination (Benjamin, 2019b), opacity (Pasquale, 2015), surveillance (Zuboff, 2019), and the exacerbation of social inequalities (Eubanks, 2017).

In several cases, algorithmic systems have been banned or replaced by other modes of decision-making due to their disadvantageous effects: In the Netherlands, for example, the Dutch welfare authorities deployed the system SyRI (short for "system risk indication") in order to detect fraud regarding welfare benefits. SyRI cross-referenced numerous kinds of personal data of citizens from different databases – e.g., data about work, fines, taxes, properties, housing, education, retirement, debts, benefits, allowances, subsidies, permits, and more – and it compared the large data masses to individual citizens with the goal of finding "discrepancies" and "unlikely citizen profiles" that lead to further investigation (Vervloesem, 2020). Its lack of transparency – targeted citizens were not informed and could not be aware that they were being under investigation – as well as its de facto primary use in socio-economically disadvantaged neighborhoods lead to massive criticism, and in early 2020, the Dutch Supreme Court has determined the system's incompatibility with Article 8 of the ECHR (DutchNews, 2020).

Another example for an algorithmic mode of knowledge production that has been banned in one city is the concept of "predictive policing" that uses data-driven analytics to determine potential hotspots of future crime. The allocation of police resources according to the forecast disproportionally affects, as critics argue, economically disadvantaged neighborhoods and communities of Color through over-policing and racial profiling (see, e.g., ACLU, 2016; Alexander, 2019; Benjamin, 2019a), as well as intensifies "feedback loops" (Ensign et al.,

2018): allocating police resources to crime hot spots leads to the production of more crime-related data due to the presence of police forces. This data is fed-back into the predictive system, which leads to the reinforcement of crime predictions. In 2020, the city of Santa Cruz, California, was the first city to officially ban the use of "predictive policing technologies" (Sturgill, 2020), due to widespread criticism.

These two very different examples have in common that the algorithmic system affects the respective individuals – welfare beneficiaries or criminal suspects – in sensitive areas of their lives. Especially when algorithmic systems are developed for and deployed in the context of state-action in constellations in which the individual is relatively powerless and dependent, considerations of potential disadvantageous effects – and the question of who will be harmed – are at the very heart of questions of justice.

Research questions

In this doctoral thesis, I pursue an interdisciplinary approach combining the perspectives of legal philosophy, mathematics, and intersectional feminist theory to examine the (in)justice of deploying data-based algorithmic systems in the context of state-action with regard to individuals in dependent constellations, as well as the legal remedies against disadvantageous effects. This main research question will be examined in the following chapters:

As a first step, a short introduction is provided into the mathematics behind data-based algorithmic systems that is as broad as necessary, while still being mathematically precise. The chapter then focuses on elaborating a workable concept of the situational constellation that accounts for the specificities of individuals being subject to algorithmic decision-making.

The second step will be to explore in what way disadvantageous effects of algorithmic systems are embedded in their underlying mathematical architecture. The idea of this second chapter is to zoom into the mathematics of algorithms and establish respective interdisciplinary methodological frameworks.

Against this background, chapter three will introduce two case studies of data-based algorithmic systems in constellations of dependency that will be analyzed via the established frameworks. The first case study is the "AMS algorithm" in Austria. The second case study, as of March 2022, has not yet been decided upon.

Based on the findings of chapter three, in its fourth chapter, the thesis is going to focus on the philosophical issue of developing an adequate approach to justice. The chapter is going to ask

to what extent anti-discrimination law and data protection law can foster the capabilities of individuals when they face potentially disadvantageous effects of data-based algorithmic systems. Finally, the question will be raised whether it is just for a state to subject individuals in certain constellations to data-based algorithmic systems. In the remainder of the exposé, the four chapters of the thesis are laid out briefly.

Chapter 1 – Scope of thesis

The first chapter establishes the four-fold scope of the thesis: This thesis focuses on (1) disadvantageous effects of (2) "data-based algorithmic systems" that are (3) deployed in contexts of state-action with respect to (4) individuals that are in a position that renders them especially prone to disadvantage.

The mathematical techniques behind the term "Artificial Intelligence" have changed significantly since its early research. In the abstract of "A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence", John McCarthy in 1955 described as the underlying assumption of Artificial Intelligence research "that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1955, p. 1). Since then, this *rule-based* paradigm of generating knowledge by explicitly inscribing the underlying processes into a computer program has transformed: The extensive digitalization and datafication of many areas of human life, as well as the increasing technological capacities to produce, collect, process, and analyze large quantities of data have led to a paradigm shift in which "[d]ata starts to drive the operation" (Alpaydin, 2016, p. 11).

In order to, firstly, account for the shift from former rule-based AI research to current Machine Learning techniques and to secondly, also include more elementary modes of knowledge production via traditional statistical methods, I introduce the term "data-based algorithmic system". This notion encompasses algorithmic tools, from statistical logistic regression (see, e.g., Harrell, 2015) to more complex Machine Learning and Artificial Intelligence methods such as Deep Learning and neural networks (see, e.g., Bishop, 2006; Goodfellow et al., 2016; Hastie et al., 2009). This definition is not a technical or methodical definition, but an epistemic definition asking the question of *which kind of knowledge* is produced.

Chapter 2 – Establishing methodical frameworks

In this chapter, I establish methodical frameworks that surface potential sources of disadvantageous effects and harm that are embedded in the mathematical architecture of data-based algorithmic systems.

2.1. Classification via prediction: Three building blocks

This subchapter examines a class of data-based algorithmic systems: classification via prediction. The main substance of this section has already been published as a peer-reviewed paper (Lopez, 2019) which was the first scientific publication on the Austrian "AMS algorithm" and that also serves as a case study within this thesis. Firstly, the mathematics of classification via prediction systems is set forth and explained for a non-mathematical audience. In a classification via prediction, first, a prediction is made regarding some phenomenon, e.g., whether a patient with certain health-related data entries will suffer from a heart attack in the near future. One might aim to allocate special medical resources and care to those at risk. The classification groups are, thus, *patients at risk* and *patients not at risk*. Depending on the classification, the patient receives extra medical resources, or not. The target variable, in this case *a heart attack in the near future*, must be precisely quantified to be understandable by a mathematical system. E.g., one might aim to predict whether a patient will suffer from a heart attack within the next four weeks. The outcome of a prediction is always a probability and, thus, a number between 0 % and 100 %, mostly coded as a number between 0 and 1. Such a continuous prediction becomes a classification by introducing thresholds, for example, by establishing that a patient will be classified as a *patient at risk* if the probability of suffering from a heart attack within the next four weeks is 45 % or higher. This percentage, 45 %, is a threshold that builds the classification. This example shows that there are three realms of decisions that constitute the classification: the data chosen to infer predictions, the quantified target variable(s), and the thresholds – the three building blocks.

This subchapter proceeds to generalize the three building blocks, each with a conceptual as well as a concretely implemented dimension: Firstly, the specific view on the past as a conceptual dimension that finds its realization in the underlying data, secondly, the particular outlook on the future implemented as the target variable(s), and thirdly, the classification thresholds are the numerical cut-off points that transform a continuous prediction into a discrete classification and, therefore, determine the final decomposition of classification groups. This thesis argues that the three building blocks can be operationalized as an analytical framework to identify

biases (see below) and predisposed assumptions that can have disadvantageous effects to vulnerable individuals.

This section proceeds to further examine the first building block – data – that is not only relevant to classification via prediction systems, but that is, obviously, central to all data-based algorithmic systems. In order to understand data, I borrow from the disciplinary perspective of Science and Technology Studies (STS), where it has been established that data and datafication does not yield a mere representation of a phenomenon, but rather entails ontological interference (Mol, 2002), and is never per se existent in the world as "raw data" (Gitelman, 2013, p. 1). Regarding classification and establishing categories per se, Bowker and Starr (2008) have pointed out that "the act of classification is … both organizational and informational, always embedded in practice" (Bowker & Star, 2008, p. 320). Critiquing the scientific ideal of objectivity itself, Haraway (1988) has coined the "god trick of seeing everything from nowhere" (Haraway, 1988, p. 581) which can be applied to datafication and "Big Data" (Prietl, 2019). In general, STS scholarship on technology and, especially, on data and datafication has been adapted to digital technologies and algorithmic systems (see e.g. Benjamin, 2019a; Prietl, 2019; Vertesi & Ribes, 2019).

2.2. The bias typology

Taking up the thread from the previous chapter with its focus on data, this section enters the broad discourse around biases in data. The main substance of this chapter has already been published as a peer-reviewed paper (Lopez, 2021b; see also Lopez, 2021a). Much scholarly work has already been written on bias (see, e.g., Angwin et al., 2016; Buolamwini & Gebru, 2018; Chouldechova, 2017; Criado-Perez, 2019; Hildebrandt, 2019; Obermeyer et al., 2019): Friedman and Nissenbaum (1996) introduced their widely cited typology on "Bias in computer systems" as early as 25 years ago. A computer system is biased, in their definition, if it discriminates unfairly and systematically, meaning that "it denies an opportunity or a good or … it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate" (Friedman & Nissenbaum, 1996, p. 332). Friedman and Nissenbaum (1996) introduce a taxonomy of biases according to their source, and since then, much work has emerged on categorizing biases (see, e.g., Barocas & Selbst, 2016; Olteanu et al., 2019; Suresh & Guttag, 2020)

I introduce a typology of three types of data bias in data-based algorithmic systems that is intrinsically linked to legal anti-discrimination conceptions. A crucial differentiating factor

between the three types is whether and to what extent the underlying data (the *training data* that was used to build the system) differs from the phenomenon that is supposed to be represented by the data, and if so, whether it is a discrepancy due to structural inequalities in society or a conceptual error. The concept of structural inequality in this paper is linked to the legal anti-discrimination regulations that are applicable in the respective context. Anti-discrimination regulations concern individuals who share a so-called feature that is legally protected against discrimination in certain contexts. In the proposed typology, algorithmic systems, just as instances of potential discrimination, are viewed as situated in their respective legal context. Accordingly, the typology presented in the following must always be considered together with a respective legal anti-discrimination framework.

The first type of bias I term purely *technical bias* which is defined quite broadly so that it includes any kind of technical or conceptual mismeasurement and misconception: There is a deviation between what is supposed to be depicted (or measured) and what is actually depicted (or measured) in the data. However, this deviation is not based on an underlying structural inequality. The second type of bias is *socio-technical bias*. In this case there is a discrepancy between what is to be represented and what is being represented, and this discrepancy is a result of structural inequalities. This includes cases where disadvantaged groups are less visible, overly visible, or wrongly depicted because of the way the underlying data is produced. The third type is *societal bias*. The crucial aspect here is that societal bias is not a deviation of the datafication from the phenomenon in reality – acknowledging that reality is a highly contested concept that always requires a normative decision. Societal bias arises when structural inequalities are reflected in the respective data, albeit correctly. The underlying data of a data-based algorithmic system depicts – in a correct way – that society structurally disadvantages certain groups. The subchapter continues to elaborate how all three types of biases can have disadvantageous effects to individuals.

The data bias typology introduced in this chapter contributes to the existing discourse in two ways: firstly, the concept of structural inequality, and the associated notion of "undesirable bias" is linked to the respective applicable legal anti-discrimination regulations, and thus, defined accordingly. Secondly, applying the proposed typology does not require deep knowledge of the inner workings of a data-based algorithmic system, as algorithmic systems are often opaque in several ways (Burrell, 2016, p. 3). The typology is simple enough so that it can be applied widely, and complex enough to be useful in analyzing an algorithmic system with the overarching question: Can a given biased data-based algorithmic system – theoretically

– be modified to technically remove the discriminatory parts using de-biasing methods established in the field of Fair Machine Learning (see, e.g., Zehlike et al., 2020), or is it mathematically impossible?

Chapter 3 – Case studies

In this chapter, the methodical frameworks established in the previous chapter will be applied to two case studies. The first case study is the Austrian "AMS algorithm" by the Austrian Public Employment Service (Arbeitsmarktservice, in short: AMS). The second case study is, as of March 2022, not yet decided upon.

3.1. The AMS algorithm in Austria

In autumn of 2018, the Austrian Public Employment Service announced a large-scale digitalization project which attracted a lot of media attention (see, e.g., Al-Youssef, 2021; Staudacher, 2020; Szigetvari, 2018a; Wimmer, 2018a; Wimmer, 2018b). The project entails a data-based algorithmic system, which became known by the name "AMS algorithm". The algorithmic system is supposed to be used by the respective AMS case worker, and it receives as input various data of the unemployed individuals. On the basis of these data the system calculates their *chances on the labor market* and according to the predicted chances the system produces as output a placement in one of three categories: the category of unemployed individuals with predicted high chances (group A), those with medium chances (group B) or of those unemployed with predicted low chances (group C) (Holl et al., 2018). Depending on the algorithmically supported classification, the unemployed are supposed to have differing access to support resources (Szigetvari, 2018b). Group A is supposed to not get access to support resources, as the individuals classified into this group will probably enter employment without support; group C is supposed to be transferred to external agencies, as providing internal resources is considered to be too inefficient; internal resources will, instead, be focused on those that are classified into group B (Allhutter et al., 2020, p. 11). This subchapter examines the AMS algorithm with the previously established frameworks in order to answer the question whether and in which way potentially emerging disadvantageous effects are embedded within the mathematical architecture of the algorithmic system.

The AMS algorithm is a data-based algorithmic system, as it produces knowledge about the individual by way of statistically comparing their data to the previously recorded and analyzed aggregated AMS case data (Allhutter et al., 2020, p. 24; Holl et al., 2018, p. 4). It is deployed

with regard to individuals that are in a position of dependency: the unemployed depend on the welfare resources and, therefore, have to comply with the AMS.

3.2. Zooming into the mathematics of the AMS algorithm

The AMS algorithm is a system that classifies via prediction. The three building blocks established previously are, firstly, the data, secondly, the target variables and, thirdly, the thresholds. Using these building blocks, this subchapter will explore the question of where human decisions were interwoven into the mathematical architecture of the AMS algorithm. Examining the building blocks will make it possible to locate biases, which directly leads to applying the data bias typology established above. The subsequent subchapter applies the bias typology (*technical bias – socio-technical bias – societal bias)* to the AMS algorithm to examine the question which biases might be removed and which not.

Chapter 4 – Questions of (in)justice

There is an entire sub-field of research within Computer Science that is devoted to the question of how to design Machine Learning systems that decide fairly: *Fair Machine Learning*[1] studies given Machine Learning systems and tests them for fairness and potentially discriminatory outcomes. Researchers in this field draw from philosophy and theories of justice (see, e.g., Shah et al., 2021), as well as from anti-discrimination law (see, e.g., Zehlike et al., 2020), and transform ideas of justice to explicit fairness metrics (see, e.g., Binns, 2018; Lundgard, 2020). These metrics, then, can be mathematically implemented to ensure that a Machine Learning system decides fairly – according to the implemented metric. For example, a Machine Learning system might be tested with regard to "demographic parity", meaning that for each – previously defined – demographic group, the algorithmic outcome will be equal, i.e., independent of belonging to that demographic group (Yee et al., 2021, p. 7). Another way of measuring and, thereby defining, fairness is "accuracy equity", which examines whether the Machine Learning system in question works equally well for all demographic groups, deeming it unfair, if it systematically works worse for certain demographic groups (Angwin et al., 2016). There are many different fairness metrics (Verma & Rubin, 2018; Wachter et al., 2021), and researchers have shown that it is mathematically impossible for a Machine Learning system to suffice all

[1] See, e.g., the annual Conference on Fairness, Accountability, and Transparency of the Association for Computing Machinery: https://facctconference.org/

possible fairness metrics (see, e.g., Chouldechova, 2017). Whether a given Machine Learning system decides fairly, thus, depends on the chosen metric.

Equality as a central facet of justice is regarded between individuals that are subject to algorithmic decisions: If individual A obtains the algorithmic outcome X_A, and individual B obtains the outcome X_B: Is this distribution of outcomes just? The locus of justice in the field of Fair Machine Learning is situated at the micro level: Fair Machine Learning examines single decisions and specific Machine Learning systems. In this doctoral thesis, I examine the question of justice at a different level by asking: Is it just to subject individuals in a position of dependency to algorithmic decision-making (and others not)? Or is it, on the contrary, even imperative for a state to deploy efficient algorithmic methods in order to efficiently manage its resources with respect to individuals?

Contribution of thesis

This thesis aims to give an interdisciplinary account on how to assess the development and deployment of Artificial Intelligence, Machine Learning, and algorithmic systems that, while they entail many promises, are not free from risks. In order to make sure that the risks of being harmed by new technologies do not affect those that are most vulnerable, this thesis centers a intersectional perspective in the inquiry of justice. By providing interdisciplinary frameworks of analysis, this thesis aims to contribute to discussions in the realms of digitalization and innovation management. Proactively assessing from a theoretical perspective of justice the impact of technologies to vulnerable individuals allows us as a society to decide whether we would like to implement a new technology or not.

References

ACLU. (2016). Statement of Concern About Predictive Policing by ACLU and 16 Civil Rights Privacy, Racial Justice, and Technology Organizations. https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

Alexander, M. (2019). The new Jim Crow: Mass incarceration in the age of colorblindness. https://www.overdrive.com/search?q=705ED541-07AB-4970-81C9-9C1E76AABF9E

Allhutter, D., Mager, A., Cech, F., Fischer, F., & Grill, G. (2020). Der AMS Algorithmus – Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS) (0xc1aa5576_0x003bfdf3). self. https://doi.org/10.1553/ITA-pb-2020-02

Alpaydin, E. (2016). Machine learning: The new AI. MIT Press.

Al-Youssef, M. (2021, April 15). Regeln für künstliche Intelligenz: AMS-Algorithmus auf dem EU-Prüfstand. Der Standard. https://www.derstandard.at/story/2000125884443/regeln-fuer-kuenstliche-intelligenz-ams-algorithmus-auf-dem-eu-pruefstand

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., & Selbst, A. (2016). Big Data's Disparate Impact. https://doi.org/10.15779/Z38BG31

Benjamin, R. (2019a). Race after technology: Abolitionist tools for the new Jim code. Polity.

Benjamin, R. (2019b). Assessing risk, automating racism. Science, 366(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of Machine Learning Research, 81, 149–159.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Bowker, G. C., & Star, S. L. (2008). Sorting things out: Classification and its consequences (1. paperback ed., 8. print). MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 1–15.

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

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3804983

Criado-Perez, C. (2019). Invisible women: Data bias in a world designed for men. Abrams Press.

DutchNews. (2020, February 5). Government's fraud algorithm SyRI breaks human rights, privacy law. DutchNews.Nl. https://www.dutchnews.nl/news/2020/02/governments-fraud-algorithm-syri-breaks-human-rights-privacy-law/

Ensign, D., Friedler, S. A., Nevilla, S., Scheidegger, S., & Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. Proceedings of Machine Learning Research, 81, 1–12.

Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330–347. https://doi.org/10.1145/230538.230561

Gitelman, L. (Ed.). (2013). 'Raw data' is an oxymoron. The MIT Press.

Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). A new era: Artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology, 16(7), 391–403. https://doi.org/10.1038/s41585-019-0193-3

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. The MIT Press.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies, 14(3), 575. https://doi.org/10.2307/3178066

Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed). Springer.

Hildebrandt, M. (2019). The Issue of Bias. The Framing Powers of Ml. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3497597

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). Das AMS-Arbeitsmarktchancen-Modell. Dokumentation zur Methode. http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment (First edition). Little, Brown Spark.

Li, B., Hou, B., Yu, W., Lu, X., & Yang, C. (2017). Applications of artificial intelligence in intelligent manufacturing: A review. Frontiers of Information Technology & Electronic Engineering, 18(1), 86–96. https://doi.org/10.1631/FITEE.1601885

Lopez, P. (2019). Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. Critical Issues in Science, Technology and Society Studies. Conference Proceedings of the STS Conference Graz 2019, 289–309. https://doi.org/10.3217/978-3-85125-668-0-16

Lopez, P. (2021a, April). Artificial Intelligence und die normative Kraft des Faktischen. Merkur, 863, 42–52.

Lopez, P. (2021b). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review, 10(4). https://doi.org/10.14763/2021.4.1598

Lundgard, A. (2020). Measuring justice in machine learning. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 680–680. https://doi.org/10.1145/3351095.3372838

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Mol, A. (2002). The body multiple: Ontology in medical practice. Duke University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data, 2, 13. https://doi.org/10.3389/fdata.2019.00013

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

Prietl, B. (2019). Big Data: Inequality by Design? Weizenbaum Conference. https://doi.org/10.34669/wi.cp/2.11

Shah, K., Gupta, P., Deshpande, A., & Bhattacharyya, C. (2021). Rawlsian Fair Adaptation of Deep Learning Classifiers. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 936–945. https://doi.org/10.1145/3461702.3462592

Singh, S., Sharma, P. K., Yoon, B., Shojafar, M., Cho, G. H., & Ra, I.-H. (2020). Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city. Sustainable Cities and Society, 63, 102364. https://doi.org/10.1016/j.scs.2020.102364

Staudacher, A. (2020, August 20). Einsatz von AMS-Algorithmus wird untersagt. Futurezone. https://futurezone.at/netzpolitik/einsatz-von-ams-algorithmus-wird-untersagt/401006765

Sturgill, K. (2020, June 26). Santa Cruz becomes the first U.S. city to ban predictive policing. Los Angeles Times. https://www.latimes.com/california/story/2020-06-26/santa-cruz-becomes-first-u-s-city-to-ban-predictive-policing

Suresh, H., & Guttag, J. V. (2020). A Framework for Understanding Unintended Consequences of Machine Learning. ArXiv:1901.10002. http://arxiv.org/abs/1901.10002

Szigetvari, A. (2018a, October 10). AMS bewertet Arbeitslose künftig per Algorithmus. Der Standard. https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus

Szigetvari, A. (2018b, October 10). AMS-Vorstand Kopf: ‘Was die EDV gar nicht abbilden kann, ist die Motivation’. Der Standard. https://www.derstandard.at/story/2000089096795/ams-vorstand-kopf-menschliche-komponente-wird-entscheidend-bleiben?ref=article

Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness - FairWare ’18, 1–7. https://doi.org/10.1145/3194770.3194776

Vertesi, J., & Ribes, D. (Eds.). (2019). DigitalSTS: A field guide for science & technology studies. Princeton University Press.

Vervloesem, K. (2020, April 6). How Dutch activists got an invasive fraud detection algorithm banned. Algorithmwatch. https://algorithmwatch.org/en/syri-netherlands-algorithm/

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3792772

Wiesmüller, M., Hegny, I., Triska, M., Banfield-Mumb-Mühlhaim, A., Prem, E., & Dachs, B. (2018). Artificial Intelligence Mission Austria 2030. Die Zukunft der Künstlichen Intelligenz in Österreich gestalten. Bundesministerium für Verkehr, Innovation und Technologie (BMVIT). https://www.bmk.gv.at/themen/innovation/publikationen/ikt/ai/aimat.html

Wimmer, B. (2018a, October 12). AMS-Chef: 'Mitarbeiter schätzen Jobchancen pessimistischer ein als der Algorithmus'. Futurezone. https://futurezone.at/netzpolitik/ams-chef-mitarbeiter-schaetzen-jobchancen-pessimistischer-ein-als-der-algorithmus/400143839

Wimmer, B. (2018b, October 17). Der AMS-Algorithmus ist ein „Paradebeispiel für Diskriminierung". Futurezone. https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421

Yee, K., Tantipongpipat, U., & Mishra, S. (2021). Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–24. https://doi.org/10.1145/3479594

Zehlike, M., Hacker, P., & Wiedemann, E. (2020). Matching code and law: Achieving algorithmic fairness with optimal transport. Data Mining and Knowledge Discovery, 34(1), 163–200. https://doi.org/10.1007/s10618-019-00658-8

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power (First edition). PublicAffairs.