Lopez, Paola

Article

Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Lopez, Paola (2021) : Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 10, Iss. 4, pp. 1-29, https://doi.org/10.14763/2021.4.1598

This Version is available at:
https://hdl.handle.net/10419/250397

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 4

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems

Paola Lopez *University of Vienna*

DOI: https://doi.org/10.14763/2021.4.1598

Published: 7 December 2021
Received: 30 October 2020 **Accepted:** 17 May 2021

Funding: The author has received project funding from the Gender Studies Association Austria (ÖGGF) for this article.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Lopez, P. (2021). 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

Keywords: Artificial intelligence, Machine learning, Bias

Abstract: This paper introduces a socio-technical typology of bias in data-driven machine learning and artificial intelligence systems. The typology is linked to the conceptualisations of legal anti-discrimination regulations, so that the concept of structural inequality—and, therefore, of undesirable bias—is defined accordingly. By analysing the controversial Austrian "AMS algorithm" as a case study as well as examples in the contexts of face detection, risk assessment and health care management, this paper defines the following three types of bias: firstly, purely technical bias as a systematic deviation of the datafied version of a phenomenon from reality; secondly, socio-technical bias as a systematic deviation due to structural inequalities, which must be strictly distinguished from, thirdly, societal bias, which depicts—correctly—the structural inequalities that prevail in society. This paper argues that a clear distinction must be made between different concepts of bias in such systems in order to analytically assess these systems and, subsequently, inform political action.

This paper is part of **Feminist data protection**, a special issue of *Internet Policy Review* guest-edited by Jens T. Theilen, Andreas Baur, Felix Bieker, Regina Ammicht Quinn, Marit Hansen, and Gloria González Fuster.

Introduction

Algorithmic systems, artificial intelligence and machine learning as socio-technical phenomena are currently experiencing a critical moment of social and political discussion. Critique is positioned in multi-layered constellations of accusations of discrimination, surveillance and reinforcement of inequalities (Angwin et al., 2016; Apprich et al., 2018; Benjamin, 2019a; Buolamwini & Gebru, 2018; Eubanks, 2017; Korinek & Stiglitz, 2021; O'Neil, 2016; UN Special Rapporteur, 2019; Zuboff, 2019). In the context of political, activist and academic discourses, the potential of discrimination, as well as the general tendency towards an automated reinforcement of inequalities by means of algorithmic systems, is closely linked to discourses on algorithmic and data bias (Angwin et al., 2016; Buolamwini & Gebru, 2018; Criado-Perez, 2019; Obermeyer et al., 2019). This paper argues that a clear distinction must be made between different types of bias in data-based[1] systems in order to analytically assess and politically critique these systems.

This paper proposes an analytical framework of three types of data bias: technical data bias, socio-technical data bias, and societal data bias (see Figure 1 below). The typology is intrinsically linked to legal anti-discrimination frameworks and is, therefore, to be applied to an algorithmic system that is situated in its specific context of use and national legal context.

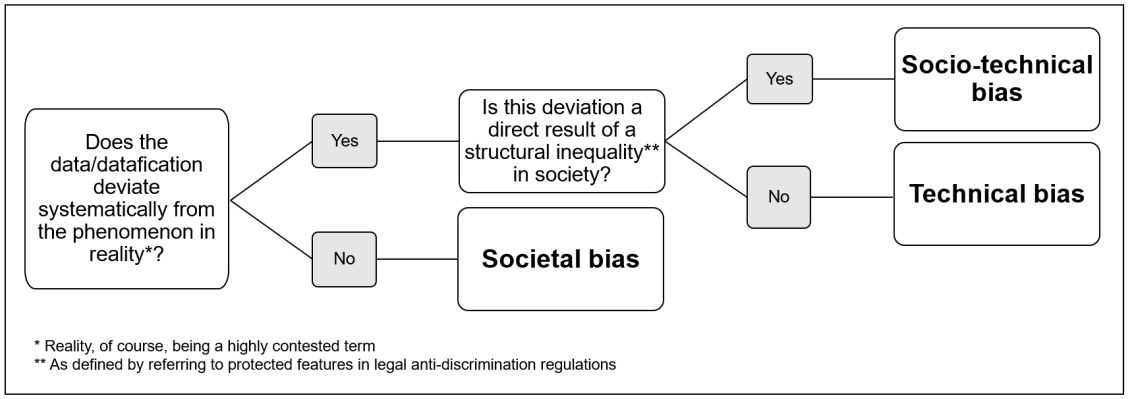

FIGURE 1: Bias scheme

1. See Section 1 for the use of terminology regarding "data-based" systems as opposed to "data-driven" systems.

A crucial differentiating factor between the three types is whether and to what extent what one wants to measure and datafy differs from what is actually measured and datafied, and if so, whether it is a discrepancy due to structural inequalities in society or a conceptual error. However, it is apparent at all times that such typologies can only be simplifications, just as any form of datafication and categorisation.

This paper introduces the proposed bias typology along an Austrian case study of an algorithmic classification system for the unemployed, the AMS algorithm, that has been discussed widely due to its potential of discrimination (Allhutter et al., 2020; Kayser-Bril, 2019; Lopez, 2019; Szigetvari, 2018a; UN Special Rapporteur, 2019; Wagner et al., 2020; Wimmer, 2018b). Looking at this particular algorithmic system, it becomes especially clear that one has to strictly differentiate between two concepts I call socio-technical bias and societal bias. Whereas socio-technical bias can theoretically be repaired, as it will be argued below, societal bias calls for political and activist action in order to transform the context of use—and potentially, to ban the system altogether.

Much scholarly work has already been written on categorising bias: Friedman and Nissenbaum (1996) introduced their iconic typology describing "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). What is meant by "unreasonable" or "inappropriate", though, is not defined in a precise way.

Friedman and Nissenbaum's typology differentiates biases in computer systems according to their source: technical bias stems from the specific affordances and constraints of technology and technological systems: "e.g., a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favoured systematically for a match over individuals displayed on later screens" (Friedman & Nissenbaum, 1996, p. 334). Pre-existing bias has its roots in social institutions, practices and attitudes. It is introduced into a computer system either via a prejudiced individual, consciously or unconsciously, or via society at large, one example being "gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls" (Friedman & Nissenbaum, 1996, p. 334). Emergent bias in a computer system arises in interaction with real users: either by the system's inability to adapt to new knowledge, "e.g., a medical expert system for

AIDS patients has no mechanism for incorporating cutting-edge medical discoveries that affect how individuals with certain symptoms should be treated" (Friedman & Nissenbaum, 1996, p. 334) or by a mismatch between users and system design regarding expertise or values.

In recent years, much fruitful research has emerged on categorising, measuring and analysing bias—updating Friedman and Nissenbaum's work to contemporary data-based algorithmic systems. The work in this field is as diverse as the applications of data-based systems.

For example, Thiem et al. (2020) recently found that, in the context of social research, applying a certain algorithm from electrical engineering (the Quine-McCluskey algorithm) in order to analyse data with the Qualitative Comparative Analysis method for purposes of causal inference can lead to significant algorithmic biases due to the algorithm's incompatibility with causal inference.

Focusing more on social platform data at the core of social research itself, Olteanu et al. (2019) have reviewed a variety of research that uses social data in order to understand or influence phenomena specific to, as well as beyond, social software platforms, and the biases that can arise within that data: some of these biases are specific to social (platform) data, such as behavioural biases that describe "systematic distortions in user behaviour across platforms" (Olteanu et al., 2019, p. 7), or normative biases that "are a result of written or unwritten norms and expectations of acceptable patterns of behaviour on a given online platform or medium" (Olteanu et al., 2019, p. 11). Other bias categories can be applied to analyse research with general data, such as population biases that describe "systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population" (Olteanu et al., 2019, p. 6). Or, more generally, data collection biases that arise via "selection of data sources, or by the way in which data from these sources are acquired and prepared" (Olteanu et al., 2019, p. 13).

Expanding the focus of use cases of data-based systems from research to decision-making that affects individuals and groups, Barocas and Selbst (2016) provide a taxonomy of mechanisms through which the use of data-based algorithmic systems may have discriminatory effects: the target variable, i.e., the outcome(s) of interest, can be defined in a way that systematically disadvantages certain groups of people: defining a target variable entails making a real-world phenomenon "computable" and, therefore, necessarily leads to simplifications that are based on subjective choices of individuals. Biases can be introduced into the data that is used

to build the algorithmic system due to: manual labelling or in the process of data collection itself, or during the feature selection (in which, similarly to defining the target variable, choices are made about which kind of data is being looked at). If the use of an algorithmic system discriminates against certain groups in a hidden way, meaning that the implemented criteria are not discriminatory *per se* but, in fact, systematically disadvantage certain groups—Barocas and Selbst refer to "masking" if it is done on purpose (Barocas & Selbst, 2016, p. 692), and proxies if not (see below).

Zooming into the technical details of a machine learning system's life cycle, Suresh and Guttag (2020) described various issues that can introduce bias into a system: historical bias, representation bias, measurement bias, aggregation bias, learning bias, evaluation bias and deployment bias. Some of these types of data bias can only be identified through extensive knowledge and close examination of the development process of a particular system including the underlying data used to build the system. This knowledge is often neither available to the public nor to critical scholarship (Burrell, 2016). Even if an algorithmic system as well as its underlying data are made transparent for research purposes, the example of face recognition shows that measuring and quantifying bias is a complex endeavour: while there has been research on "How (not) to measure bias in face recognition networks" (Glüge et al., 2020, p. 1), Cavazos et al. (2021) stress the importance of a multi-faceted and case-based analysis: "a general assessment of bias for face recognition algorithms [is] unfeasible" (Cavazos et al., 2021, p. 108). They conclude that "race bias must be measured for each particular scenario, algorithm, race, and dataset" (Cavazos et al., 2021, p. 109).

Coupling the quest to analyse biases in algorithmic systems with the aim to mitigate their harmful effects, Simon, Wong and Rieder (2020) give a fruitful account of the discussion around bias in machine learning, and they expand on the benefits of the Value Sensitive Design methodology in designing, developing and deploying algorithmic systems: through conceptual-philosophical, empirical and technical investigations, Value Sensitive Design aims to "provide […] a constructive tool that enables and supports the realisation of specific desired values in the design and development of new technologies" (Simon et al., 2020, p. 5). They stress the importance of viewing algorithmic systems and their biases as embedded in "the broader socio-technical system in which [they] are situated" (Simon et al., 2020, p. 11). In the same vein, the bias typology introduced in this paper below, together with the question of whether biases can be "fixed" are viewed in their respective context of use: in the section on societal bias, for example, it will be dis-

cussed that an algorithmic system built on even a perfect datafication can reinforce inequalities—depending on its context of use.

The data bias typology introduced in this paper aims to contribute to the multifaceted discourse in several ways: firstly, this typology centres vulnerable humans and the concepts are defined accordingly. Secondly, the concept of structural inequality, and the associated concept of "undesirable bias" is linked to the conceptions of legal anti-discrimination regulations (Scherr et al., 2017; Atrey, 2019), and thus, defined accordingly. For example, the fact that the AMS algorithm differentiates according to an unemployed person's education will, in the following typology, not be viewed as an undesired differentiation, because one's status of education is not a feature that is protected by anti-discrimination law (Lopez, 2019). A differentiation on the basis of an unemployed person's gender entry with potentially harmful effects, on the other hand, is viewed as undesirable, because the applicable Austrian anti-discrimination law protects the feature gender. Algorithmic systems are, thus, viewed as being situated in their respective national legal context and context of use. Lastly, applying the proposed typology does not require deep knowledge of the inner workings of an algorithmic system. Put differently, this paper aims to provide a typology that is simple enough so that it can be applied widely, and complex enough to be useful in analysing an algorithmic system with regard to the overarching question: can a given biased algorithmic system—theoretically—be fixed? Applying this typology will make it clear that "de-biasing the AMS algorithm", for example, would not be a solution to the problem of potential discrimination.

Section 1: Data-based algorithmic systems

Various aspects of designing, developing and deploying an algorithmic system, such as the choice of the underlying model used, the ways in which performance metrics are defined, the organisational embedding of a system etc. are crucial for the socio-technical analysis of algorithmic systems (see Barocas & Selbst, 2016; Cavazos et al., 2021; Olteanu et al., 2019; Suresh & Guttag, 2020). This paper focuses on one specific aspect: data and data bias. Data is the mathematical materiality, and therefore the fundamental epistemic fabric of many contemporary algorithmic systems, as will be laid out in the following.

State-of-the-art algorithmic systems and research on artificial intelligence (AI) and machine learning differ in several significant points from the beginnings of AI research. At the beginning of the development of AI systems, the goal was to develop computer programmes that simulate a cognitive, human-like thought process

(Crevier, 1993; McCarthy et al., 1955; Press, 2016). The underlying assumption was that “every aspect of [...] intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al., 1955). This approach did not lead to satisfactory performance, as human inference and cognition processes are too complex to be explicitly transferred to a synthetic system.

Modern AI, however, is built on methods of data-based machine learning (Alpaydin, 2016). In this modern paradigm, it is not rule-based cognitive processes that are being emulated. Machine learning systems are based on probabilistic, statistical methods (Hastie et al., 2009). The strength and the crucial element of modern AI functionality are large amounts of data: increased processing capacities now make it possible to "teach" a computer programme what is the "right" (in the sense of: desired) outcome by processing huge amounts of data, so that modern AI programmes, at least with supervised machine learning (Hastie et al., 2009), are trained by the sheer number of examples to take the correct action accordingly. Patterns found in the respective training data become rules that will be applied to new input data.

In this way the terms artificial intelligence and ‘Big Data’ are related: ‘Big Data’ is the "material" of an artificial intelligence system. It requires a large quantity to recognise supposed patterns (Bishop, 2006). Epistemically, data-based algorithmic systems can only "see" the mass and not the individual. The algorithmically generated knowledge can then be used to compare the individual to the statistical mass.

In data science, one uses the term “data-driven” systems which speaks to the paradigm shift in which: “Data starts to drive the operation; it is not the programmers anymore but the data itself that defines what to do next” (Alpaydin, 2016, p. 11). From a perspective informed by Science and Technology Studies (STS), this paper has to reject the inevitability and naturalisation of claims and notions of “data itself” driving anything.

Firstly, because data is not *per se* existent in the world as “raw material” one has to merely “mine” (as the term “data mining” suggests (Aggarwal, 2015)) but is itself always produced, as STS scholarship has established (Gitelman, 2013; Mol, 2002). Secondly, because it is and always has been a variety of socio-technical actors who “define what to do next”. In an attempt to nevertheless capture the paradigm shift from the former AI research to machine learning techniques, this paper introduces the term “data-based algorithmic systems”. This refers to algorithmic tools, from logistic regression to machine learning and artificial intelligence methods such as deep learning (Goodfellow et al., 2016; Harrell, 2015; Hastie et al., 2009).

Much has already been written about data and datafication, not just in the digital context: That datafication is not a mere representation of a phenomenon, but rather entails ontological interference (Mol, 2002). That algorithmic systems and datafication function as a mode of regulation (Yeung, 2018). That science might be entering a paradigm of statistical exploration and data mining (Kitchin, 2014), and that data and evidence-based policy are becoming increasingly central (Rieder & Simon, 2016). There is also a lot of work already written on updating the scholarship on data to data-based algorithmic systems and automated decision-making, taking into account the specific affordances of digital technologies with respect to their ubiquity, scalability and supposed efficiency (Benjamin, 2019a; D'Ignazio & Klein, 2020; Prietl, 2019). Datafication turns a heterogeneous, complicated reality into a supposedly homogeneous microcosm with phenomena that seem classifiable, comparable and controllable.

Section 2: The AMS algorithm as a case study

In autumn 2018, the Austrian Public Employment Service (*Arbeitsmarktservice* in German, in short: AMS) announced a large-scale digitisation project which attracted a lot of media attention (Szigetvari, 2018a). The project entails a predictive algorithmic classification system, which became known by the name AMS algorithm, that is designed to categorise the unemployed into three groups with differing access to welfare support resources (Wimmer, 2018b).

The algorithmic system receives as input various data of unemployed individuals. On the basis of these data the system calculates their so-called "chances on the labour market" and according to the predicted chances the system produces as output a placement in one of the three categories: the category of unemployed 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 classification, the unemployed are supposed to have different support resources available to them (Szigetvari, 2018b).

Since the announcement of this project, there has been much criticism, both from the scientific community (Cech et al., 2019) and activists (Kurier, 2020). The manifold criticism is centred around three focal points: firstly, it became known in a published method documentation of the algorithmic system that the personal data entry "Gender: Female" results in an automatic "deduction of points", which translates to the fact that unemployed individuals can be assigned to a less eligible group solely on the basis of their gender (Holl et al., 2018; Wimmer, 2018b). Further potential point deductions according to personal data entries, such as age,

childcare responsibilities, disability and nationality, can lead to an intersectionally compounded effect (Lopez, 2019; Wagner et al., 2020): the group of job-seekers with the lowest “chances'' for job placement (according to the predictive model), group C, should not get full access to all AMS support resources, the rationale behind that being efficiency (Allhutter et al., 2020; Szigetvari, 2018a).

Secondly, and connected to the first issue, there was criticism of the fact that group C is supposed to be transferred out of the AMS internal system and relocated to external agencies. The aim of these external agencies is to "increase their chances" in order for them to be placed in group B—which is supposed to have full access to all support resources, and an evaluation spoke of the potential danger that these external formats might be a "one-way street" (Auer et al., 2019, p. 73; Lopez, 2019).

Thirdly, there has been and still is criticism concerning the so-called “decision support” that the system is supposed to provide. The algorithmic system itself is not designed to automatically classify, but to be used as a support system for the case workers in their daily work (Wimmer, 2018a). Researchers have expressed their concerns: as studies show that algorithmic results, in practice, are often accepted far too easily by users, they fear that the same effect might occur with AMS case workers accepting and confirming the algorithmic classifications in their daily work (Wimmer, 2019). In August 2020, the Austrian Data Protection Authority banned the planned use of the algorithmic system for three reasons: the missing legislative basis for this type of profiling, the lack of measures taken against the system turning into a *de facto* decision-making automatism (which, in the planned form without further protection measures, is not compatible with Art. 22 of the GDPR), and due to a missing data protection impact assessment according to Art. 35. of the GDPR (Staudacher, 2020). In December 2020, the AMS appealed the decision successfully (Szigetvari, 2020). As of August 2021, the Austrian Supreme Administrative Court has yet to decide on the algorithm’s fate (Fanta, 2021). The legal, as well as the overall political response to this system can be seen as a landmark case for data-based systems in the European welfare state context, especially regarding so-called “decision support” tools.

In sum, the heart of the criticism is thus a severe potential for *de facto* automated intersectional discrimination—because if, due to an intersectional convergence of various “point deductions” (the sensitive data entries being age, gender, disability status, childcare responsibilities, nationality) corresponding to one’s intersectional positionality, a person is outsourced to group C and does not receive certain support resources for this reason alone, then this is at the very heart of the concept of

intersectional discrimination: only the intersection of several axes of vulnerability and therefore only the combination of several "point deductions" leads to a placement in the disadvantaged group C, whereas the existence of "only one" disadvantaged feature implies a placement in the most eligible group B. Put differently, the system allows for individuals that belong to "only one" or "only a few" disadvantaged groups to get access to all AMS support resources, which, *per se*, can be a societally desirable goal. If, however, a person has too many data entries belonging to disadvantaged groups (e.g., female, above 50, with childcare responsibilities, and with a non-EU citizenship)—this person might be classified below the threshold with her "low chances'' and, consequently, be placed in group C. This is an algorithmically explicit manifestation of what feminist theory has studied and criticised for decades: the AMS classification system focuses its resources on those that are "a little" disadvantaged (and, therefore, might be classified into group B), and outsources those that are multiply disadvantaged: "This focus on the most privileged group members marginalizes those who are multiply-burdened [...]." (Crenshaw, 1989, p. 140; see also Crenshaw, 1990)

This becomes highly relevant especially in welfare contexts, as vulnerable individuals might be faced with automated decisions which can have unprecedented consequences (Dencik & Kaun, 2020; UN Special Rapporteur, 2019). Vulnerability, in this paper, is being centred and seen as a "universal, inevitable, enduring aspect of the human condition that must be at the heart of our concept of social and state responsibility" (Fineman, 2008, p. 8). Intersectional feminist theory specifies this in that "structures make certain identities the consequence of and the vehicle for vulnerability" (Crenshaw, 2016, n.p.).

Section 3: Three types of data bias

This paper argues that the proposed bias typology can be made fruitful when assessing a data-based algorithmic system that might produce biased results and, as a consequence, enable discriminatory practices. The three types of bias defined in this paper differ along the question whether the corresponding data or datafication differs systematically from what is supposed to be measured or datafied, and if so, whether that discrepancy is rooted in structural inequalities in society, see Figure 1 for a scheme of the bias types.

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 typol-

ogy, 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: in the EU, the different EU Directives, such as the Gender Directive 2004/113/EC, the Race Directive 2000/43/EC and the Framework Directive 2000/78/EC have been implemented into respective national law, such as the *Gleichbehandlungsgesetz* (GlBG) in Austria, or the *Allgemeines Gleichstellungsgesetz* (AGG) in Germany (see e.g. Holzleithner, 2017). In the USA, there is Title VII of the Civil Rights Act 1964 (which prohibits employment discrimination on the basis of sex, race, colour, national origin and religion), the Age Discrimination in Employment Act 1967, and the Americans with Disabilities Act 1990 (see, e.g., Atrey, 2019).

The respective legislator recognises that there are structural inequalities in society and deems them undesirable, thus prohibiting discriminatory treatment on the basis of this structural disadvantage (Givens & Evans Case, 2014). In Austria, for example, discrimination based on gender, parenthood, age, disability and various other so-called attributes in employment contexts is prohibited by law according to the GlBG which applies to the context of use of the AMS algorithm (Wagner et al., 2020).

It should be noted at this point that in the intersectional coalescence of structural vulnerabilities, by far not everything is seen through the mono-axially designed law, which differentiates according to supposedly separate categories (Atrey, 2019; Crenshaw, 1989; Holzleithner, 2010; Uccellari, 2008). In this way, legal anti-discrimination frameworks ask: “What caused the damage?” instead of “Who is suffering?” (Crenshaw, 1989, 1991). Another obvious shortcoming of using the legal anti-discrimination framework is the fact that economic inequality is not captured—suffering from poverty is not a protected feature in the EU directives, or in the legal anti-discrimination regulations in the USA mentioned above. Another issue with legal frameworks is that they are always embedded in their national (or supra-national) context. Accordingly, applying the typology proposed in this paper always follows the national legal context in which the algorithmic system in question operates. Algorithmic systems and their potentially harmful effects on vulnerable individuals and groups have to be viewed as situated in their context of use—and this situatedness is as dynamic and ever-changing as legal regulations can be. In short, linking the bias typology to legal anti-discrimination regulations entails its own set of invisibilities.

Nevertheless, the legal anti-discrimination regulations provide a precise instru-

ment of analysis with regard to which one can conceptualise "undesirable bias" in accordance with structural inequalities, so as to not having to resort to undefined or unclear concepts of fairness (the opposite of which being bias), or morality (Friedman & Nissenbaum, 1996; Suresh & Guttag, 2020). The huge field of fairness in machine learning shows that fairness can have numerous possible definitions and metrics (see e.g. Verma & Rubin, 2018). Defining "undesirable bias" according to the concept of fairness, thus, also depends on first choosing an applicable definition of fairness.

The first type of bias in the typology that I propose is purely technical bias, which I define quite broadly so that it includes any kind of technical or conceptual mismeasurement and misconception. A deviation exists here between what one wants to depict or measure and what is depicted or measured. However, this deviation is not based on an underlying structural inequality.

The second type of data bias I call 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 direct result of structural inequalities. This includes cases where disadvantaged groups are less visible, overly visible or wrongly depicted because of the way data is produced. Several cases of data bias have been discussed in the media and beyond, as in Angwin et al. (2016) and Criado-Perez (2019).

The third type is societal bias. The crucial aspect here is that societal bias is not a deviation of the datafication of a phenomenon from reality—acknowledging that reality is, obviously, a highly contested concept that always follows a normative decision. According to the proposed typology, societal bias arises when structural inequalities are reflected in the respective data, albeit correctly. The underlying data of an algorithmic system depicts—in a correct way—that society structurally discriminates against certain groups. The fact that, in the AMS algorithm, the data entry "Gender: Female" has a negative effect on the algorithmically predicted "chances" on the job market is, in that sense, a correct, albeit simplified, reflection of the structural inequality that marginalises women in the Austrian labour market—an instance where the statistically aggregated individual biases of decision-makers, such as recruiters and employers, become visible. More details and examples can be found below in the respective sections.

All three types of bias can cause enormous damage reinforcing structural inequalities in society. The impact of an algorithmic system, of course, always depends on how a system is built, how it is used, as well as on what it is supposed to do, and

what it is *de facto* doing. Algorithmic systems whose underlying data is subject to technical bias can produce completely incorrect and indeed nonsensical and meaningless results, so that the people who suffer are ultimately still vulnerable individuals who are already disadvantaged. The "correction" of nonsensical algorithmic results can require a lot of financial resources, knowledge, patience, legal advice et cetera (Eubanks, 2017; UN Special Rapporteur, 2019), so that a datafied error—even if it does not emerge in a structurally inequality-related manne—can have structurally different effects depending on the positionality of the affected individual. This assumes, though, that it is always desirable for the datafied version of something to be as correct as possible. However, this can be highly ambivalent, as will be discussed below.

Algorithmic systems that render disadvantaged groups along a socio-technical bias appear less visible, wrongly visible or all too visible can cause serious damage. A well-known and often discussed example of hypervisibility is that of neighbourhoods in the USA that are being over-policed due to racially biased law enforcement data and corresponding allocation of law enforcement resources due to predictive policing systems (Benjamin, 2019a; Ensign et al., 2018; Wang, 2018), corresponding to the fact that crime itself is conceptualised in a racialised way (Alexander, 2019; Butler, 2017; Wang, 2018). Underrepresentation in databases can lead to disadvantageous invisibility, which has been written about in health care contexts (Obermeyer et al., 2019).

Algorithmic systems that are subject to societal bias, which means that they depict society with all its structural intersectional inequalities—albeit correctly—can also, depending on the context of use, be disproportionately harmful to vulnerable individuals, if the depicted reality is regarded, and used, as both the normative status quo and a naturalised fact (Lopez, 2019). More details will be given below under the respective types of bias.

In any case, with any algorithmic system it depends on how it is used and what it is used for. Thus, a system is never harmful or beneficial in itself, but must be viewed the way it is situated in its specific context.

Section 4: Purely technical data bias

The first part of the proposed typology refers to instances in which the produced data—and thus the foundation of a data-based algorithmic system—deviates systematically from the phenomenon in reality that this data is supposed to describe, acknowledging, of course, that reality is a highly contested term that always re-

quires a preceding normative decision. The systematic deviation in the case of technical bias can be the result of a conceptual error made by a human actor. The concept of technical data bias differs from the concept of validity of data (see e.g. Olteano, 2019) as it points to conceptually emphasising the potentially harmful effects on vulnerable individuals or groups. Technical data bias, as opposed to the methodological concept of validity, therefore centres vulnerable humans.

The crucial feature of this type of data bias is the fact that these deviations are not directly rooted in a structural inequality that is prevalent in society but still may have effects that are structured around social inequalities. One simple illustrative example is a thermometer that is supposed to measure the temperature in a room, but that is placed very close to a heating unit so that the temperature measured deviates systematically from the average room temperature in a way that is systematically biased towards higher temperatures. If this biased measurement of temperature potentially has a systematic negative effect on disadvantaged individuals (e.g., as a heating regulating device), this paper speaks of technical bias.

In the case of the AMS algorithm, and, in fact, of any predictive data-based algorithmic system, the predictive model uses data from the past in order to predict outcomes in the future. The AMS algorithm predicts the individuals' "chances" for job placement on the labour market. The data-based predictions that are produced are based on the analysis of past data starting in 2015 (Holl et al., 2018, p. 4). Thus, in order to estimate the chances of an unemployed person, the aggregated statistical analysis of unemployed people with similar data entries from the past years is used.

In order to derive the predictive model, the statistical analysis asked two questions: "Which people with which data entries have achieved job placement for at least three months within seven months?". This was called the short-term goal. And: "Which people with which data entries have achieved job placement for at least six months within 24 months?". This was called the long-term goal (Holl et al., 2018, p. 6). Answers to these questions can be found in data on previous cases: the underlying model, namely a logistic regression model (Holl et al., 2018, p. 3), operates under the assumption that these "chances" can be estimated sufficiently well by analysing what was recorded in the past by the convoluted, aggregated AMS case histories and their data, combining all cases and respective case histories of the previous years (Holl et al., 2018, p. 4). Answers to the questions of which job-seekers with which data entries re-entered employment achieving the short-term or the long-term goal can be seen in the past case histories.

What makes this statistical analysis of the past a “prediction” is the active assumption that the future will behave in the same way that the past has behaved. In other words, there is no “time parameter” in logistic regression (Hastie et al., 2009). The mathematical architecture of this model, in this sense, knows no future. This future is being created by assuming factual, and in this case labour market related, continuity with regard to time.

This is the source for technical bias, in that the labour market that is depicted in the AMS data from the past several years is very different from the present COVID-19 job crisis and its related labour market situation (Der Standard, 2020). In Austria, the employment sectors tourism, retail, temporary employment, production, and construction have drastically fluctuated in the course of the pandemic (APA-OTS, 2021). These are sectors that rely heavily on precarious workers, so that a misrepresentation of these industries in the underlying data might have systematic effects on vulnerable individuals. The way unemployment during, and probably after, the COVID-19 crisis presents itself therefore cannot be modelled with data from before COVID-19. The data deviates systematically from the present labour market, which renders the corresponding predictions useless (Wimmer, 2020a). Using pre-COVID-19 data in order to predict a COVID-19 or post-COVID-19 job placement therefore constitutes a conceptual error that is an example of technical bias: the data used to make a prediction is systematically biased towards the past, which, in fact, all data-based predictive systems are (see also Lopez, 2021).

Section 5: Socio-technical data bias

In the case of socio-technical data bias, there is, too, a systematic divergence between the data and the phenomenon that is supposed to be depicted by the respective data. In contrast to technical bias from above, this discrepancy is rooted in a systematic over-, under- or misrepresentation or misdatafication of disadvantaged groups or aspects corresponding to marginalised groups in the data due to a structurally biased way the data is produced.

This deviation is due to and thus reveals a structural inequality that prevails in society. As mentioned above, this bias framework in this paper builds on the legal anti-discrimination framework applicable in the respective context. Structural inequality is thus defined as inequality along a so-called "feature" which is legally protected against discrimination in the national legal context of use of the algorithmic system in question.

One example that has been much discussed is face detection and face recognition

software based on machine learning. As mentioned above, (supervised) machine learning in this case functions by providing the computer programme with large amounts of training data in order to optimise the model parameters. A machine learning programme therefore requires many samples of visual data to find patterns in how a face is to be detected, classified or recognised.

The way of and extent to which faces are detected and recognised by such an algorithmic system, therefore, depends on the visual data that has been used to train this system. The work of Joy Buolamwini and Timnit Gebru has shown that several widely used systems for gender classification are subject to intersectional racial bias. These systems recognise faces with light skin colour and male faces better and poorly recognise faces with dark skin colour, and especially faces of Women of Colour. Buolamwini and Gebru show in their work that two benchmark data sets used to evaluate such systems for accuracy are biased and contain disproportionately few faces of People of Colour. They conclude that a system that performs poorly on faces of People of Colour and performs well on faces of white people will perform well with regard to these standard benchmark data sets (Buolamwini & Gebru, 2018; see also Cavazos et al., 2021; Glüge et al., 2020).

With regard to face detection, in which a camera determines whether or not a face is within the camera scope, Buolamwini, during an impressive Ted Talk, showed a video of how her own face is initially not recognised by a widely used face detection software, but then is recognised as a face as soon as she puts on a white plastic mask (Buolamwini, 2017).

These two instances are directly visible translations of the fact that these face recognition, classification and face detection programmes have primarily learned what faces look like by looking at white faces, specifically white male faces. Thus, this face detection system is subject to socio-technical bias, which means that the underlying data of the system (a collection of visual data of faces) differs from the actual range of faces to be captured.

This kind of biased visual design is hardly new either, as early on in the development of analogue photography decisions were made in the chemical processes of standardised photo development that favour light faces and make dark faces look undifferentiated, bad and invisible, as Ruha Benjamin puts Lorna Roth's works into new contexts (Benjamin, 2019a; Roth, 2009). A structural and systematic racial bias thus results in a technical divergence—the outcome is socio-technical data bias.

Vulnerability of marginalised groups intersects with socio-technical data bias in

face recognition software when low software performance leads to wrongful arrests: widely used face recognition technologies in law enforcement that perform poorly on faces of Black people already have had consequences that are devastating for the individual affected: several cases are known where Black men were wrongfully arrested and incarcerated due to a mismatch in a face recognition tool (Hill, 2020a, 2020b).

Another example of what this paper classifies as socio-technical bias, is an algorithmic system in the USA situated in health care contexts which is designed to predict which patients are at high risk and therefore need attention and medical care. This system is widely used to guide decisions on the allocation of health resources and attention to millions of people, and a study has found out that this risk prediction system systematically predicts that Black patients have less medical need than white patients, resulting in them receiving less medical care (Obermeyer et al., 2019; Obermeyer & Mullainathan, 2019).

The study showed that this systematic underrepresentation of medical need is due to the fact that a so-called proxy variable was used to calculate the risk and the medical needs of patients. A proxy variable is a feature that is supposed to determine what is to be measured (in this case the severity of the health condition) by means of another variable that often is easier to quantify. In this case the proxy for medical need was chosen to be health care costs. Thus, what is actually being predicted by the system, is not risk, but health costs, and this prediction, as mentioned above, is merely an analysis of past health cost data. The study argues that disproportionately little money is spent on Black patients, which demonstrates a structural system of racial inequality in the USA (Benjamin, 2019b).

This leads to the following situation: a white patient and a Black patient, who both have the same algorithmically predicted risk and should therefore, guided by the algorithmic prediction, receive the same amount of medical care, differ systematically in that the Black patient has a much more serious health condition (Obermeyer et al., 2019). Put differently, in the case of a Black and a white patient with the same degree of actual medical need, decisions guided by this system will allocate more resources to the white patient. This naturally leads to an allocation of resources and medical attention that is immensely harmful to the vulnerable Black patient in need of medical care.

In the case of the AMS algorithm, an example of socio-technical bias is the fact that the corresponding gender entry only knows two options, male and female, even though in Austria, the corresponding gender scheme is not binary (Allhutter

et al., 2020; Wagner et al., 2020). The complexity of the, at least, legally possible gender variety, is not being captured by the possible categories in the AMS database. This is a simplification due to the fact that additional options for a gender entry are not being considered as a priority high enough to change the categorisation system in the database.

It can also be argued compellingly that the binary gender scheme fits into the category of societal bias (see below) as it depicts the reality of a binary societal gender regime.

This shows that the bias typology introduced in this paper has blurry boundaries between the different bias concepts—categorising biases in specific algorithmic systems is always also a normative act: after all, one has to decide on a version of ”reality” that the datafication deviates from (or not). Striving for more diverse representation, one will reject a binary gender scheme: however, the consequences of adding possible data entries to the binary scheme, or removing the data category “gender” altogether from the AMS algorithm are difficult to estimate. From the perspective of wanting to improve the capacity of a system to represent gender variety, adding as many possibilities of gender entries as possible may be desirable. However, and especially in view of “surveillance capitalism” discourses (Zuboff, 2019), it becomes clear that representation and (datafied) visibility is always highly ambivalent. Removing the “gender” category, on the other hand, will mathematically lead to a lower performance of the algorithmic system (see below). Depending on one’s perspective, this can be desirable or not: a low performance of the AMS algorithm may lead to its political legitimation to disappear, which may lead to the system’s abolition (which may be an activist goal).

The examples of this section have shown that socio-technical bias in data-based algorithmic systems can appear in diverse forms and at differing stages of the development of an algorithmic system. With regard to face detection, the underlying visual data is unbalanced due to an underrepresentation of faces of women of Colour (Buolamwini & Gebru, 2018). With regard to the predictive health risk system, what is supposed to be measured, namely the actual need of a patient, is being approximated by the predicted cost in health care, revealing that Black patients have been and are being under-treated (Obermeyer et al., 2019). With regard to the gender categories in the AMS system, the possible entries and therefore the degree to which complexity is enabled, are restricted to a binary gender scheme (Wagner et al., 2020).

Section 6: Societal data bias (or: prediction as depiction)

Societal data bias occurs when structural societal inequalities are reflected by data, albeit in a correct way. Societal bias in data-based algorithmic systems is therefore not a false representation or a deviation, but a manifestation of the aggregated individual biases that are structured along social inequalities due to gender, race, age, disability (the features depending on the legal context and its anti-discrimination regulations). In the following, I argue that algorithmic systems that depict societal bias can serve as an emancipatory tool of diagnosis, or as a normative reinforcement of structural inequalities, depending on the context of use.

The AMS algorithm in Austria is an example of an algorithmic system rendering societal bias visible. The underlying assumption behind the development of this predictive algorithmic system is that the target probability ("job chances") can be estimated sufficiently well by the data included (Harrell, 2015; Holl et al., 2018). The statistical finding that the feature "Gender: Female" (as well as: age over 30, age over 50 more severely so; non-EU citizenship; disability; childcare responsibilities) has a negative impact on the probability of job placement thus shows, correctly, that there is a structural disadvantage in the labour market: two unemployed individuals with otherwise completely identical data entries have statistically different success rates with regard to job placement. The "Gender: Female" feature alone with otherwise unchanged data has a negative effect.

The AMS algorithm depicts societal bias which means that it shows the aggregated individual biases of institutions and decision-makers in the realms of the labour market that are structured via intersectional inequalities. The model therefore reflects (only up to a certain degree of simplification, of course) the structural situation on the labour market with which this person will be confronted when searching for a job (Lopez, 2019).

This knowledge could potentially open up an emancipatory moment in the use of data-based algorithmic systems. As an analysis of the Austrian labour market and its discriminatory tendencies, this model with its embedded statistical findings could thus be an insightful tool for combatting inequalities and allocating resources starting at the group with the lowest predicted "chances". The current use of the model does the opposite, however, in that individuals are subjected to the collective disadvantage of their non-voluntary memberships to groups, formed via datafied categories, that are discriminated against structurally (Lopez, 2019).

As mentioned above, the lines between socio-technical data bias and societal bias that is rendered visible in data are blurry. One example is risk assessment in the realms of prediction of recidivism in the criminal justice system. Very well-known are the findings of Angwin et al., who showed that the accuracy of the COMPAS[2] system varies very much among different races (Angwin et al., 2016). This is, therefore, bias that is reflected in the way how "well" the system works. Non-white people are systematically classified as more dangerous and are subject to stricter sanctions in various areas. Hence in the proposed typology of this paper this varying accuracy can be seen as a case of socio-technical bias.

This section discusses a different risk assessment tool, namely the Level of Service/Case Management Inventory (LS/CMI) risk assessment and case management tool. It is claimed to be "The Most Widely Used and Researched Risk/Needs Assessment" (MHS Public Safety, 2020, p. 1) which is designed for the "management of offenders in justice, forensic, correctional, prevention, and related agencies" (Andrews et al., n.d.). This section of the paper looks at a different aspect from that of the machine bias debate initiated by Angwin et al. (2016), namely the extent to which such tools are also a mirror of aggregated societal bias.

The items for calculating the risk of recidivism include "Any prior youth convictions?", "Arrested or charged under age 16?", "Currently unemployed?", "Frequently unemployed?", "Never employed for a full year?", "Financial problems", "Less than regular [school] grade 10?", "Suspended or expelled [from school] at least once?", "Criminal family/spouse?", "Some criminal acquaintances" (Andrews et al., 2004). Crime has been shown to be defined in a way that targets communities of Colour (Alexander, 2019; Benjamin, 2019a; Butler, 2017; Wang, 2018). The way in which racialised definitions of crime interact with the partly privatised systems of incarceration in the USA result in "carceral capitalism" (Wang, 2018). There is a diagnostic moment in the way this type of crime is predicted. What the items above show—even if one assumes that the "predictions", and actually correlations, are more or less accurate (which, as Angwin et al. (2016) showed for the COMPAS system, is highly questionable)—is that there is a high convergence between racialised poverty, and the way crime is defined.

2. The COMPAS system is an algorithmic system that predicts criminal offenders' risk of recidivism in order to inform further measures, such as the severity of punishment and the possibility for parole. Angwin et al. (2016), in their widely read and cited study, showed that this system is subject to a systematic racial bias: the rate of false positives with Black offenders is disproportionally high; complementary, the rate of false negatives with white offenders is disproportionally high. This means that, within the COMPAS system, Black offenders are much more likely to be falsely predicted to have a high risk score, whereas white offenders are much more likely to be falsely classified with a low risk score.

The context of use, as well as the epistemic embedding of these societally biased algorithmic tools serve as a litmus test for the world view of the responsible actors: if political actors see the world around them as fundamentally neutral, even natural, and as a given mere fact, then they will understand the corresponding data-based predictions as merely neutral facts that do not require fundamental transformation. An example of this kind of thinking is Lawrence Mead, an academic "leading expert on poverty" who was very influential in the welfare reform in the USA in the mid-1990s (Ramesh, 2010). In a by now retracted commentary (Flaherty, 2020), he claims that racial inequality is a result of "difference of culture", by which he means: "Fifty years after civil rights, their main problem is no longer racial discrimination by other people but rather that they face an individualist culture that they are unprepared for" (Mead, 2020, p. 3). This type of denial of structural oppression will regard the datafied version of an oppressive phenomenon as a neutral and individualised fact that does not require fundamentally transformative action.

From an intersectional feminist perspective, the world and the inequalities as they exist is a created world with inequalities that are constantly being re-created and perpetuated. The fact that people are entering the labour market at different rates and with differing degrees of permanence simply because of their gender, as depicted in the AMS algorithm, is a result of structural inequalities in the labour market (Lopez, 2019). Similarly, the fact that racialised poverty can be linked to the corresponding definition of crime is a tragic diagnosis of, on the one hand, the type of crime that is punished and, on the other hand, the socio-economic and racialised conditions that correlate with that definition of crime (Benjamin, 2019a).

Discussion

In this paper a typology was introduced which classifies bias in data-based algorithmic systems (see Figure 1). This typology can be applied, even if technical details of an algorithmic system are not made available transparently: discrepancies and biases can be seen from an outcome-oriented perspective that doesn't open the algorithmic black box. Furthermore, the typology is linked to the conceptualisations of the corresponding anti-discrimination legislation, so that, firstly, the concept of structural inequality is defined accordingly, and secondly, the algorithmic systems addressed are considered to be embedded in their legal context and context of use: algorithmic systems and their decision-making that may have effects on vulnerable individuals and groups are viewed within the scope of the respective legal anti-discrimination regulations.

The three types of bias were elaborated through examples from face recognition, health care management, and risk assessment, with an emphasis on the Austrian AMS algorithm. In summary, technical bias is a systematic and conceptual distortion in the underlying data of an algorithmic system; socio-technical bias is a systematic deviation due to structural inequalities and must be strictly distinguished from societal bias, which depicts—correctly—the structural inequalities in society. Applying the bias framework proposed above can inform the question of how to address these biases, as I lay out in the following.

Technical data bias and socio-technical data bias can, theoretically, be fixed by improving the data (or the target variable, or the model itself). However, as mentioned above, the quest to eliminate biases in order to be "seen" correctly by technology, is highly ambivalent, and must always be evaluated and assessed together with the respective context of use: with regard to the binary gender scheme embedded in the AMS algorithm, adding gender categories, or deleting the gender entry altogether can have side-effects that are not easy to anticipate. The "dilemma of difference" coined by Martha Minow in 1987—which states that both seeing and not seeing difference can lead to an amplification of that difference and, thus, inequality (Minow, 1987)—can be updated to data-based algorithmic systems, and points to said ambivalence. As mentioned above, deleting the gender entry would lead to a lower performance of the algorithm (in that its predictions would be less accurate) which might lead to a loss of legitimation for the project itself, as the algorithm has been praised for high accuracy rates (Der Standard, 2019). This can, depending on the perspective, be a desirable political goal. On the other hand, one can aspire to be represented more accurately in one's gender identity that is situated in a larger variety than a binary gender scheme affords. In that case, pressing for the addition of more possible gender entries will be the desirable action.

Within the health resources allocation system, it will most likely be desirable for vulnerable patients in need of adequate medical care to be represented by the system with one's actual medical need: after improving the system, the rate of Black patients receiving more medical care afterwards increased significantly (Obermeyer et al., 2019). In other cases of socio-technical bias it highly depends on the context of application of a specific system whether one wants to be seen and "captured" by algorithmic systems, and this will also differ for different individuals and is not a group-based need or wish: In the case of unbalanced training data sets used to develop face detection or face recognition systems, one solution to the problem of socio-technical bias (if one wants to improve algorithmic performance in order for the software to become better at detecting and recognising all kinds of

faces) is considered to be producing "better" data sets, meaning data sets that are more intersectionally balanced. This leads to a search for better visual data, i.e. of more diverse visual face data. This "search" for previously marginalised faces can be, too, a highly ambivalent endeavour, especially in the context of technology improvement in order to cater an industrial product to a broader market:

The promotion video of the unlocking function of the Google Pixel 4 showed a Black woman in a dark room unlocking her phone automatically via face recognition, emphasising the improved face recognition capacity of the camera (Barbello, 2019) and thus, implicitly responding to intersectional bias and varying degrees of accuracy of above-mentioned face detection systems. However, in striving for better software performance by "repairing" the socio-technical bias, Google and its contractors were accused of specifically targeting homeless Black people in order to "scan" and "collect" their faces for their improved training data set (Nicas, 2019) which points to the highly adaptive power of capitalist interests.

If an algorithmic system is subject to societal bias, I argue that "repairing" the system is not an adequate quest: instead, to mitigate the potentially harmful effects of the algorithmic system, the very context of use needs to be transformed. In the case of the AMS algorithm, the underlying model assumes correlations between the variables and the predicted "chances" (Hastie et al., 2009). It is mathematically built on the very differentiation between the datafied groups structured along gender, age, nationality, disability status, childcare responsibilities, and others. It is its foundational mathematical functionality to discern between these groups. I argue that removing these variables, or striving to balance the data would result in the model collapsing: mathematically and statistically, it needs these categories in order to make accurate predictions. The "point deductions", in that sense, can be seen as a tool of diagnosis: as mentioned above, they reflect the fact that certain groups have been disadvantaged in the past (which is reflected in the data). Any predictive system that is built for that purpose will entail similar biases. The harmful effects in this case stem from its specific context of use:

Both in the AMS algorithm, as well as in the LS/CMI risk assessment tool, a prediction is made that, in a first step, does not "predict" but actually depicts the discriminatory reality of structural oppression as a datafication. Then, in a second step, it normatively reinforces this discriminatory reality as a supposedly objective fact and finally, in a third step, returns it to the social sphere by means of the resulting measures:

Group C of the AMS algorithm—those with predicted "low chances" on the labour

market—shall be transferred to external agencies, in order for the AMS to focus its own resources on group B—those unemployed individuals with predicted "medium chances". This might have beneficial effects for those placed in group B, but unforeseen and harmful effects on the "outsourced" unemployed individuals of group C (which are, as discussed above, those individuals that are already multiply disadvantaged to begin with).

In the LS/CMI system, offenders with a high predicted risk of recidivism receive systematically stricter treatment (Angwin et al., 2016; O'Neil, 2016; Wang, 2018). In this case, as well as in the AMS algorithm case, what is actually an issue of group-based structural inequalities is being individualised via the resulting "prediction" of the "risk" or "chances" of an individual. As discussed above, data-based systems only "see" the masses and not the individual: data-based predictions are made on the basis of past data that has nothing to do with the actual individual who will then have to suffer from the individualised consequences. Instead of changing the algorithmic systems, the very contexts of use, as well as their embedding, need to be transformed: outsourcing group C with the predicted "low chances" could be transformed to using the algorithmic system as a tool of diagnosis and, thus, as a vehicle for structural transformation by especially focussing on group C.

Data-based algorithmic systems, especially those that are deployed with resulting consequences for vulnerable individuals or groups, are interwoven with existing forms of intersectional oppression. Thus, not only the question whether biases can and need to be "repaired", and what that might entail in a specific context, but also the modes in which this endeavour is undertaken are always ambivalent and require critical examination.

References

Aggarwal, C. C. (2015). *Data Mining: The Textbook* (1st ed. 2015). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-14142-8

Alexander, M. (2012). *The new Jim Crow: Mass incarceration in the age of colorblindness* (Revised edition). New Press.

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, *3*, 5. https://doi.org/10.3389/fdata.2020.00005

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

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004a). *Level of Service/Case Management Inventory*. Global Institute of Forensic Research. https://storefront.mhs.com/collections/ls-cmi

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004b). *Level of Service/Case Management Inventory QuikScore TM Form*. Global Institute of Forensic Research. http://faculty.uml.edu/jbyrne/44.203/documents/LSCMIblankpaperversion.pdf

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

A.P.A.-O.T.S. (2021, July 1). Kocher: Mit 3,8 Millionen Beschäftigten wieder das Beschäftigungsniveau von vor der Krise erreicht. *APA-OTS*. https://www.ots.at/presseaussendung/OTS_20210701_OTS0058/kocher-mit-38-millionen-beschaeftigten-wieder-das-beschaeftigungsniveau-von-vor-der-krise-erreicht

Apprich, C., Chun, W. H. K., & Cramer, F. (2018). *Pattern discrimination* (H. Steyerl, Ed.). University of Minnesota Press.

Atrey, S. (2019). *Intersectional Discrimination* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198848950.001.0001

Auer, E., Tamler, P., Weber, F., Hager, I., Krüse, T., & Reidl, C. (2019). *Evaluierung des Betreuungsformates für Personen mit multiplen Vermittlungshindernissen (BBEN)* [Report]. http://www.forschungsnetzwerk.at/downloadpub/2019_BBEN_BBEN-ams_final.pdf

Barbello, B. (2019, July 29). (Don't) hold the phone: New features coming to Pixel 4. *The Keyword*. https://www.blog.google/products/pixel/new-features-pixel4/

Barcoas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. 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

Bishop, C. M. (2013). *Pattern recognition and machine learning*.

Buolamwini, J. (2017, March 27). *How I'm fighting bias in algorithms* [YouTube]. https://www.youtube.com/watch?v=UG_X_7g63rY

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Butler, P. (2017). *Chokehold: Policing black men*. New Press.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2021). Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(1), 101–111. https://doi.org/10.1109/TBIOM.2020.3027269

Cech, F., Fischer, F., Human, S., Lopez, P., & Wagner, B. (2019, October 3). *Dem AMS-Algorithmus fehlt der Beipackzettel*. Futurezone.

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, *1989*(1), 139–167.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241. https://doi.org/10.2307/1229039

Crenshaw, K. (2016, March 12). *On intersectionality. Keynote at the Women of the World Festival 2016* [YouTube]. https://www.youtube.com/watch?v=-DW4HLgYPlA

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. Basic Books.

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

Dencik, L., & Kaun, A. (2020). Datafication and the Welfare State. *Global Perspectives*, *1*(1), 12912. https://doi.org/10.1525/gp.2020.12912

Der Standard. (2019, January 18). Arbeitslose nach Chancen eingeteilt: OECD lobt AMS-Algorithmus. *Der Standard*. https://www.derstandard.at/story/2000096564832/teilt-arbeitslose-nach-chancen-ein-oecd-lobt-ams-algorithmus

Der Standard. (2020, September 1). Im August 422.910 Personen arbeitslos, 452.499 in Kurzarbeit. *Der Standard*. https://www.derstandard.at/story/2000119711670/im-august-422-910-personen-arbeitslos-452-499-in-kurzarbeit

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway Feedback Loops in Predictive Policing. *ArXiv:1706.09847 [Cs, Stat]*. http://arxiv.org/abs/1706.09847

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

Fanta, A. (2021, January 28). *Jobcenter-Algorithmus landet vor Höchstgericht*. NetzPolitik.org. https://netzpolitik.org/2021/oesterreich-jobcenter-algorithmus-landet-vor-hoechstgericht/

Fineman, M. A. (2008). The Vulnerable Subject: Anchoring Equality in the Human Condition. *Yale Journal of Law & Feminism*, *20*(1), 1–23.

Flaherty, C. (2020, July 28). U.S. and 'Them.' *Inside Higher Ed*. https://www.insidehighered.com/news/2020/07/28/leading-voice-welfare-reform-accused-racism

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. (2013). *"Raw data" is an oxymoron*. The MIT Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2517806

Givens, T. E., & Evans Case, R. (2014). *Legislating Equality: The Politics of Antidiscrimination Policy in Europe*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198709015.001.0001

Glüge, S., Amirian, M., Flumini, D., & Stadelmann, T. (2020). How (Not) to Measure Bias in Face Recognition Networks. In F.-P. Schilling & T. Stadelmann (Eds.), *Artificial Neural Networks in Pattern Recognition* (Vol. 12294, pp. 125–137). Springer International Publishing. https://doi.org/10.1007/978-3-030-58309-5_10

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

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer.

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

Hill, K. (2020a, August 3). Wrongfully Accused by an Algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hill, K. (2020b, December 29). Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. *The New York Times*. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). *Das AMS-Arbeitsmarktchancen-Modell* [Concept Paper]. SYNTHESISFORSCHUNG. http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Holzleithner, E. (2010). Mehrfachdiskriminierung im europäischen Rechtsdiskurs. In U. Hormel & A. Scherr (Eds.), *Diskriminierung* (pp. 95–113). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92394-9_5

Kayser-Bril, N. (2019, October 6). Austria's employment agency rolls out discriminatory algorithm, sees no problem. *Algorithm Watch*. https://algorithmwatch.org/en/story/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Korinek, A., & Stiglitz, J. E. (2021). Covid-19 driven advances in automation and artificial intelligence risk exacerbating economic inequality. *BMJ*, n367. https://doi.org/10.1136/bmj.n367

Kurier. (2020, June 25). Petition gegen AMS-Algorithmus gestartet. *Kurier*. https://kurier.at/wirtschaft/petition-gegen-ams-algorithmus-gestartet/400950995

Lopez, P. (n.d.). *Reinforcing Intersectional Inequality via the AMS Algorithm in Austria*. https://doi.org/10.3217/978-3-85125-668-0-16

Lopez, P. (2021, March 25). Artificial Intelligence und die normative Kraft des Faktischen. *Merkur*, 42–52.

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

Mead, L. M. (2020). RETRACTED ARTICLE: Poverty and Culture. *Society*. https://doi.org/10.1007/s12115-020-00496-1

M.H.S.Public Safety. (2020). *Measure and Predict Recidivism in Adults with the NEW Digital LS/CMI Assesment and Case Management System*. Global Institute of Forensic Research. https://issuu.com/mhs-assessments/docs/ls-cmi.lsi-r.brochure_insequence

Minow, M. (1987). Foreword: Justice Engendered. In M. Minow & D. C. Langevoort (Eds.), *The Supreme Court, 1986 Term* (Vol. 101, p. 7). Harvard Law Review. https://www.jstor.org/stable/1341224?origin=crossref

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

Nicas, J. (2019, October 4). Atlanta Asks Google Whether It Targeted Black Homeless People. *The New York Times*. https://www.nytimes.com/2019/10/04/technology/google-facial-recognition-atlanta-homeless.html

Obermeyer, Z., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 89–89. https://doi.org/10.1145/3287560.3287593

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

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

Press, G. (2016, December 30). *A Very Short History Of Artificial Intelligence (AI)*. Forbes. https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/#3b14f51a6fba

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

Ramesh, R. (2010, June 16). Does getting tough on the unemployed work? *The Guardian*. https://www.theguardian.com/society/2010/jun/16/lawrence-mead-tough-us-welfare-unemployed

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 1–6. https://doi.org/10.1177%2F2053951716649398

Roth, L. (2009). Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication*, *34*(1). https://doi.org/10.22230/cjc.2009v34n1a2196

Scherr, A., El-Mafaalani, A., & Yüksel, G. (Eds.). (2017). *Handbuch Diskriminierung*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-10976-9

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Staudacher, A. (2020). *Einsatz von AMS-Algorithmus wird untersagt. Futurezone*.

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

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

Szigetvari, A. (2018b). *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

Szigetvari, A. (2020). *Gericht macht Weg für umstrittenen AMS-Algorithmus frei*. Der Standard. https://www.derstandard.at/story/2000122684131/gericht-macht-weg-fuer-umstrittenen-ams-algorith

mus-frei

Thiem, A., Mkrtchyan, L., Haesebrouck, T., & Sanchez, D. (2020). Algorithmic bias in social research: A meta-analysis. *PLOS ONE*, *15*(6), e0233625. https://doi.org/10.1371/journal.pone.0233625

Uccellari, P. (2008). Multiple Discrimination. How Law can Reflect Reality. *The Equal Rights Review*, *1*, 24–49.

U.N. Special Rapporteur. (2019). *Report of the Special Rapporteur on extreme poverty and human rights*. https://www.ohchr.org/en/issues/poverty/pages/srextremepovertyindex.aspx

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

Wagner, B., Lopez, P., Cech, F., Grill, G., & Sekwenz, M.-T. (2020). Der AMS-Algorithmus.: Transparenz, Verantwortung und Diskriminierung im Kontext von digitalem staatlichem Handeln. *Zeitschrift für kritik - recht - gesellschaft*, *2*, 191. https://doi.org/10.33196/juridikum202002019101

Wang, J. (2018). *Carceral capitalism*. Semiotext(e).

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

Wimmer, B. (2019, October 18). AMS-Sachbearbeiter erkennen nicht, wann ein Programm falsch liegt. *Futurezone*. https://futurezone.at/netzpolitik/ams-sachbearbeiter-erkennen-nicht-wann-ein-programm-falsch-liegt/400147472

Wimmer, B. (2020a, August 24). "AMS-Algorithmus sollte ganz abgedreht werden". *Futurezone*. https://futurezone.at/netzpolitik/ams-algorithmus-sollte-ganz-abgedreht-werden/401009924

Wimmer, B. (2020b, September 24). AMS beruft gegen Algorithmus-Stopp durch Datenschutzbehörde. *Futurezone*. https://futurezone.at/netzpolitik/ams-algorithmus-ams-beruft-gegen-stopp-durch-datenschutzbehoerde/401042806

Yeung, K. (2018). Algorithmic regulation: A critical interrogation: Algorithmic Regulation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/rego.12158

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

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

cnrs centre internet et societe

University of Tartu Johan Skytte Institute of Political Studies