Bias as Boundary Object: Unpacking The Politics Of An Austerity Algorithm Using Bias Frameworks

Gabriel Grill
ggrill@umich.edu
University of Michigan at Ann Arbor
Ann Arbor, Michigan, USA

Fabian Fischer
fabian.fischer@uni-ak.ac.at
Department for Cross-Disciplinary Strategies, University of Applied Arts Vienna
Vienna, Austria

Florian Cech
Center for Technology and Society, TU Wien
Vienna, Austria

ABSTRACT

Whether bias is an appropriate lens for analysis and critique remains a subject of debate among scholars. This paper contributes to this conversation by unpacking the use of bias in a critical analysis of a controversial austerity algorithm introduced by the Austrian public employment service in 2018. It was envisioned to classify the unemployed into three risk categories based on predicted prospects for re-employment. The system promised to increase efficiency and effectivity of counseling while objectifying a new austerity support measure allocation scheme. This approach was intended to cut spending for those deemed at highest risk of long term unemployment. Our in-depth analysis, based on internal documentation not available to the public, systematically traces and categorizes various problematic biases to illustrate harms to job seekers and challenge promises used to justify the adoption of the system. The classification is guided by a long-established bias framework for computer systems developed by Friedman and Nissenbaum, which provides three sensitizing basic categories. We identified in our analysis "technical biases," like issues around measurement, rigidity, and coarseness of variables, "emergent biases," such as disruptive events that change the labor market, and, finally, "preexisting biases," like the use of variables that act as proxies for inequality.

Grounded in our case study, we argue that articulated biases can be strategically used as boundary objects to enable different actors to critically debate and challenge problematic systems without prior consensus building. We unpack benefits and risks of using bias classification frameworks to guide analysis. They have recently received increased scholarly attention and thereby may influence the identification and construction of biases. By comparing four bias frameworks and drawing on our case study, we illustrate how they are political by prioritizing certain aspects in analysis while disregarding others. Furthermore, we discuss how they vary in their granularity and how this can influence analysis. We also problematize how these frameworks tend to favor explanations for bias that center the algorithm instead of social structures. We discuss several recommendations to make bias analyses more emancipatory, arguing that biases should be seen as starting points for reflection on harmful impacts, questioning the framing imposed by the imagined "unbiased" center that the bias is supposed to distort, and seeking out deeper explanations and histories that also center bigger social structures, power dynamics, and marginalized perspectives. Finally, we reflect on the risk that these frameworks may stabilize problematic notions of bias, for example, when they become a standard or enshrined in law.

CCS CONCEPTS

• **Applied computing** → **Computing in government**; • **Human-centered computing**; • **Social and professional topics** → **Government technology policy**; • **Computing methodologies** → **Machine learning approaches**;

KEYWORDS

public employment services, job seeker profiling, algorithmic bias, infrastructure studies

ACM Reference Format:
Gabriel Grill, Fabian Fischer, and Florian Cech. 2023. Bias as Boundary Object: Unpacking The Politics Of An Austerity Algorithm Using Bias Frameworks. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594120

1 INTRODUCTION

Recent controversies around algorithmic decision-making systems (ADM) have led to increased public and scholarly attention toward the concept of "bias." This shifting sentiment, as many controversies turned attention onto algorithms, has also been described as the "techlash" [70, 74]. In this article, we closely investigate an ADM, colloquially called "AMS Algorithm", for profiling job seekers by the Austrian public employment services (AMS, short for Arbeitsmarktservice) that became controversial and marked as "biased." The AMS envisioned it as enabling a new regime for resource allocation based on risk scores but was heavily critiqued for justifying austerity politics, discrimination, and a strong lack of accountability, contestability, and transparency [2, 3, 45]. In this paper, we identify biases of this austerity algorithm using a long-standing framework developed by Friedman and Nissenbaum [26] as an analytical lens and discuss implications, benefits, and trade-offs of using such an approach to critique problematic ADMs.

Throughout the ADM controversy, bias ascriptions played a key role in deliberations about the possible impacts of the system as

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594120

various actors invoked them in different ways. As we will explicate later in more detail, we theorize that bias can act as a boundary object that facilitates exchange and collaboration [69] despite no clear consensus about it's meaning among involved actors. Describing an ADM as "biased" marks it as contentious, problematic, and political in some discourses. For example, in law and public debates, bias often refers to human decisions that discriminate against marginalized people, are overly partisan, or simply wrong due to the limitations of a decision maker's perspective [21]. In contrast, the term "bias" can have a more neutral or technical connotation in other contexts. For example, in machine learning, "bias" can refer to problems with data selection and collection.

Many social science scholars don't subscribe to this antagonism and hold that technologies can never be 'unbiased', similar to how technologies cannot be objective [21]. Since bias has such a strong hold in many public, scholarly, and professional spheres, some technology scholars have thus aimed at broadening how the term can be used and understood. For example, scholars highlighted how biases are not necessarily problematic and could also be intentionally included in ADMs to counteract historical injustices akin to affirmative action[56]. These conflicting valuations of bias highlight the fluidity of the concept, which we understand as a placeholder construct for some phenomena understood as a 'distortion' [56] or 'slant' [26]. Such biases become tangible usually in the form of either a calculated number or a description of the bias, its causes, and implications. There are many different ways biases could be classified, some may be more useful than others, but ultimately no absolute version exists [10]. Thus, we understand biases as socially constructed, which does not make them less real but highlights how they could be conceived of in many ways, e.g., by modelling or choosing different variables, theories, or proxies in measurement.

We understand biases as relational and situated, which highlights how they need to be contextualized to become tangible and debatable. For example, a recognized bias can be experienced as problematic or harmful by one social group while seen as an opportunity for others. This conflict is a common concern in risk assessment where the stable disadvantages of marginalized, recognizable groups can produce predictability that enables and justifies algorithms [29] used to enact and reproduce these inequalities, e.g., through triage or austerity politics [23]. We also understand biases as performative, which means once they are conceived and communicated as such, they become ways how people understand and feel failure, injustice, and unfairness. Some problematic biases are hidden and barely noticeable, but once they become known as such, they can turn into a bigger controversy[59]. This understanding of bias motivates our research into how bias can be operationalized in productive and emancipatory ways, such as to illustrate, diagnose, and problematize injustices perpetuated by ADMs. [1, 59] Such interventions have to be done with care since, as technology scholars have argued [7, 20], bias frames have a tendency to depoliticize controversies and render them 'merely' technical.

A contested concept closely related to bias is that of 'fairness,' sometimes problematically understood as an absence of bias[62]. Increasingly, various researchers and practitioners interested in improving the fairness of algorithms to address mounting criticism have begun to develop ethics checklists [47] to improve the reflexivity in development without requiring practitioners to be completely retrained or replaced. Some of these efforts have also been critiqued as techno-solutionist "band-aids" [20], and ethics-washing [75], as they may distract from necessary structural changes and regulation. Still, both the checklist approach and the concept of bias endure. In various fields of technology development, checklists and other related formalized protocols are standardized and trusted approaches and consequently remain appealing to engineers interested in dealing with problems of algorithmic systems. At this juncture, where some critical scholars argue against the use of the term bias and checklists, and practitioners continue to trust in them, the combination of both in the form of bias frameworks is becoming an important and popular genre that seeks to enable practitioners in seeing and acknowledging various forms of bias.

We understand bias frameworks as classification schemes that provide categories for possible causes of bias that can be used to guide analysis. We employ such a framework[26] to identify and construct biases of the previously mentioned job seeker profiling ADM, thereby problematizing the system and challenging its promised objectivity. In section 4, we use this analysis to reflect on using bias for critique and argue for its strategic use as a boundary object [69]. In the second part of that section, we discuss implications, issues, and benefits of using bias frameworks for analysis and provide recommendations for how to improve them so they can potentially aid in emancipatory projects. By comparing four bias frameworks and drawing on our case study, we illustrate how these frameworks are political by prioritizing certain aspects in analysis while disregarding others, vary in their granularity and how this can influence analysis, and favor explanations for bias that center the technology instead of social structures. We conclude with several recommendations for making bias analyses more just, arguing that identified biases should be seen as starting points for reflection on harmful impacts, questioning the framing imposed by the imagined "unbiased" center that the bias is supposed to distort, and seeking out deeper explanations and histories that center also bigger social structures, power dynamics, and marginalized perspectives.

2 RELATED LITERATURE

This paper mainly relates to two areas of research. First, our case study is situated in the context of public administration, specifically in public employment services. It thereby contributes to a growing body of research that critically engages with algorithms used to profile the unemployed and allocate support measures. We review literature on such systems in section 2.1. Secondly, we see our analysis of bias frameworks as contributing to the broader topic of algorithms and inequality, of which biases are framed as a source and consequence. We take a closer look at existing frameworks addressing algorithmic biases in section 2.2.

2.1 Profiling of the unemployed

The increasingly wide-spread use of profiling algorithms in public employment services (PES) are part of a so-called 'digital transformation' in public administration [25]. We conceptualize these algorithms as socio-technical assemblages increasingly woven into various everyday counseling and other administrative practices

[27, 33, 38, 50]. They are adopted to reap the promises of automation and 'evidence-based' algorithmic decision making such as efficiency and impartiality [58]. However, these systems are also a source for concern due to their potential to perpetuate inequalities and discrimination at scale [23, 72].

Nevertheless, public employment agencies increasingly use profiling algorithms to make predictions, frequently of a job seeker's likelihood of becoming long-term unemployed. This is often done to aid counselors in the assessment of prospects or to either focus or cut resources for certain risk groups. Usually algorithms are used to focus attention and resources on those deemed at highest risk of long-term unemployment[3]. Only a few cases like the Austrian AMS algorithm and a profiling algorithm in Poland[42, 51], which was scrapped after a legal challenge, follow a triage approach that seeks to cut spending on job seekers predicted to be least likely to find stable employment. Beyond algorithms used to determine unemployment risk, a related use case is the prediction of a job seeker's likelihood to exhaust unemployment insurance benefits in the USA and Canada [18, 32]. The algorithmic profiling of job seekers has a long history in public administration. Early statistical profiling systems were rolled out as early as 1993 in the USA and 1994 in Australia [46].

In addition to profiling systems used in real-world applications, the task of predicting long-term unemployment has also attracted researcher's interest to develop new and better approaches by applying long-established methods such as logistic regression [36, 40] or more sophisticated machine learning approaches, e.g., using ensemble machine learning [76], in their research.

Most of the statistical approaches in practice use some logistic or probit regression (e.g., Australia, Italy, Netherlands, Sweden and the USA), although random forest models (e.g., New Zealand, Flanders/Belgium), gradient boosting (New Zealand) and factor analysis (Latvia) are being utilized as well. More advanced machine learning approaches see little use (yet) [18]. Besides purely statistical or machine learning approaches, profiling can also be done by caseworkers and through rule-based approaches [46] (see also [32]). Frequently, a mix of methods is used, including systems that allow caseworkers to override rule-based or statistical results [18], which highlights the socio-technical nature of these ADMs.

The second crucial ingredient to profiling is the data used. Broadly speaking, two types of data sources can be identified. The first source encompasses administrative data that is collected by public employment agencies and other authorities primarily for purposes other than profiling [18]. The second type includes data collected through questionnaires and interviews specifically for the purpose of profiling. Other data sources, such as click behavior on websites, see use for profiling only in rare cases.

The general concern over algorithmic bias has become an important topic for job seeker profiling. Critical scholars from Science & Technology Studies, legal studies and computer science have recently articulated concerns that profiling systems are biased [2, 3, 45] and can lead to reinforced inequality [44]. Concerns over such bias and discrimination also called civil society to action, most recently in Austria [22] and Poland [42, 51].

Concerns over biases and 'fairness' have also led to recent research that aims to advance profiling methods capable of addressing these issues. Some researchers put a strong focus on the development of 'fair' profiling approaches [36] while others audit newly developed profiling approaches for biases with existing toolkits [76]. The topics of bias and 'fairness' have also been taken up by scholars who evaluate existing profiling systems [19] and advise policymakers [73]. A very promising line of research, advocates for participatory approaches that seek to include marginalized groups and focus on questions both on the utility of ADMs and more broadly how practices of PES can be remade for better and more just support [60]. We see our analysis of bias presented later in the paper as contributing to the growing body of research that critically engages with *algorithmic profiling of the unemployed*. These are algorithms in the public sector that affect vulnerable and marginalized people and thus questions of inequality, bias and discrimination are particularly relevant.

2.2 Bias, Frameworks and Inequality

As discussed in section 1, various critiques have been leveled against narrow and technical conceptions of bias, such as in machine learning where 'bias' often merely refers to issues of data quality and representativeness [13, 37, 53]. Scholars problematized how bias frames are captured by techno-solutionist discourse and encourage debiasing as a method for applying superficial 'fixes' mainly intended to silence critics instead of tackling deeper social issues [7, 30]. Some scholars called for abandoning the term bias altogether in favor of 'oppression' [20] as the latter emphasizes important stakes of marginalized communities, which otherwise would be branded merely as 'bias'. Scholars also argued that questions of shifts in power structures instead of bias should become central to work interrogating ADMs [35]. These interventions provide important insights that should be considered when using bias frames for different purposes. In this paper, we are particularly concerned with bias as a mode for critique and diagnosis of problems, which prior work has also identified as a potentially emancipatory intervention computing professionals could be involved in [1].

The task of how to bound and structure various identified biases to challenge the promises of a controversial ADM has motivated us to engage with so-called 'bias frameworks.' We understand them as classification schemes that provide categories that aid in identifying and bounding biases and their possible causes. They promise to bring order and comprehensibility to the messiness and variety of algorithmic failures and harms [4]. They can aid in diagnostic analysis by providing insights on where, what, and how problems can emerge in the life cycle of an algorithm from its inception to deployment.

Many frameworks are based on literature surveys that document common algorithmic failures, problems, pitfalls, and/or biases. Thereby, not all categories they list are referred to as bias, but we still consider them part of the same genre since bias is often only a shorthand for problem or distortion, as discussed in section 1. The frameworks are usually meant to aid researchers in becoming aware of known issues and how to appropriately deal with them. Most are written for an audience of researchers and potentially industry specialists. Additionally, there have been efforts (e.g., [16]) to synthesize published bias frameworks into frameworks specifically designed for use in the industrial sector. Many of these

frameworks are highly specialized. For instance, frameworks have been created for social (media) data research [52], machine learning based security research [5], computer vision [24], autonomous car development [17], and natural language processing [6, 63]. There are also frameworks that seek to be more general, in turn requiring more coarse categories. A germinal example is Friedman and Nissenbaum's bias framework for computer systems [26] which distinguishes between three major categories: technical, pre-existing and emergent bias.

More recent examples focus on the machine learning pipeline [48, 71] or algorithms in general [65]. Scholars have also developed classification schemes for the legal context, e.g. one framework [45] introduces three bias categories: technical, socio-technical (i.e., bias due to problematic modeling decisions), or societal (i.e., bias due to the reproduction of inequality in society). Bias frameworks have also been used for educational purposes[1] to illustrate the variety of sources of bias affecting algorithms and their adoption. This great variety of frameworks illustrates the current interest in better understanding biases and also how creating them to fit different contexts is seen as important to make them useful. We contribute to this literature about bias and frameworks addressing algorithmic biases by highlighting how they can be applied in an empirical case study on job seeker profiling and providing meta-level reflections and recommendations considering their utility, implications, and potential for emancipatory projects. With this work, we hope to ignite wider debate about the use and development of bias frameworks for critique and analysis.

3 BIAS FRAMEWORKS AT WORK: THE CASE OF THE AMS ALGORITHM

We present a case study of a controversial ADM, building on and extending previous work [2, 3, 12, 45]. We briefly describe the case, our methodological approach, and technical details of the system. Based on this we trace biases of the ADM using a bias classification framework [26]. The identified bias categories describe potential problems of the system when used as envisioned. The aim is to deconstruct advertised capabilities of the ADM, such as its promised objectivity, that justify its adoption and distract from the controversial austerity politics it enacts [2].

3.1 Case Study AMS Algorithm: Algorithmic profiling of the unemployed in Austria

Introduced in October 2018 by the Public Employment Service Austria (AMS) under a right-wing government,[2] the stated goals of the ADM were to decrease the influence of the subjectivity of caseworkers, increase the overall efficiency and speed of counseling, and improve the effectiveness of supportive measures (including, for instance, specialized job training, re-education, and other labor market reintegration measures). To achieve these goals, the system classifies job seekers into one of three categories based on their predicted "integration chance" (IC) into the job market:

- Group A: High short-term prospects[3]
- Group C: Low long-term prospects[4]
- Group B: Medium prospects (neither part of groups A or C)

This classification determines the levels of support available to the job seeker. Job seekers from group A are deemed most likely to find gainful employment soon without additional support measures. Subsequently, the frequency of mandatory visits with the AMS would also be reduced for these job seekers and they would not be offered certain support measures. Based on the assumption that expensive active labor market programs and other supportive measures would not substantially improve the (already presumably low) reintegration chances of group C, these job seekers would be referred to an external institution offering supervision and other 'stabilization' measures on a voluntary basis. Finally, group B, the residual category, would be offered all traditional supportive measures.

As such, the system was introduced as a semi-automated decision-support tool that promises to help distributing scarce resources efficiently. To make the necessary predictions in the form of the IC values, the system utilizes data reaching back four years in the form of both personal attributes of job seekers, as well as their history on the labor market and the performance of regional job centers. A nationwide implementation was planned for 2020. Due to a number of factors, including the heated public debate, the system gained public attention and was reviewed by the Austrian Data Protection Agency, which forbade its use in 2020. The agency argued that the processing of sensitive information requires a dedicated legal foundation according to the GDPR and the use for individual predictions is not covered by existing laws. A court case challenging this ruling is ongoing, and the system is currently not in use.

The planned operationalization of the so-called 'AMS algorithm" presented a number of tensions calling into doubt its capability to reach the stated goals of efficiency, effectiveness and reduction of human subjectivity and prejudice; for instance, the stated goal to streamline the job seeker / caseworker interactions stands in stark contrast to the assumption that caseworkers would be capable of acting as a 'human corrective' for any erroneous classifications, which puts additional strain on the already limited time with job seekers. The Data Protection Agency also argued that the proposed human oversight is insufficient and that the system should be classified as an automated decision making system according to GDPR Art. 22. This would results in stricter accountability regulations being applied that would have to be legally binding (as opposed to 'merely' internal guidelines).

3.2 Methodological approach

Our study of the AMS algorithm is based mostly on internal documents provided by the public employment agency and the contracted company as well as documents in direct response to a catalogue of questions we posed to these parties. We also engaged with civil society groups while conducting this work, such as job seeker interest groups. Many of these documents are not public, but

[1] Illustration of an educational bias framework published by digital culture scholar Felix Stalder: https://web.archive.org/web/20230207054852/https://pbs.twimg.com/media/FN39g0kXMAMQ1f-?format=jpg&name=4096x4096

[2] The early origins of the AMS algorithm date back to 2008 and early versions were envisioned as instead being used to focus support and attention on those at high risk of long-term unemployment. Planning and development with the current goal started in earnest in 2015. The ADM's deployment was originally planned for the end of 2016, but was stopped by the government at the time[3].

[3] Predicted likelihood $\geq$ 66% to have 3 months of gainful employment within 7 months

[4] Predicted likelihood $<$ 25% to have 6 months of gainful employment within 24 months

provided to the authors for a commissioned case study [3] with the permission to use them for further scientific research. The section 3.4 presents an in-depth classification of different of problematic biases of the AMS algorithm by reworking and extending results of the commissioned study.

To analyse the documents, we used qualitative document analysis as a method [9]. This involved close reading of the documents and annotating relevant sections. Our analysis draws on sensibilities from a feminist and constructivist tradition [14, 66] and seeks to follow a line of research concerned with the social study of algorithms [59, 61]. This means we understand these documents not as purely descriptive or objective, but instead as constructed texts emerging out of practices, situated in organizational cultures and responding to specific contextual demands.

Three kinds of information were annotated: First, information about the broader context out of which the AMS algorithm was created, such as what requirements guided development. Second, information that allowed us to understand the precise technical workings of the algorithm, which includes descriptions of variables and databases, and internal evaluations of the system. Third, how the AMS algorithm was supposed to be embedded into the actual processes of the public employment agency's counselling processes.

Based on identified technical details, we tried to reconstruct the workings of the AMS algorithm to assess its potential impacts on job seekers. We use the term reconstructing over reverse-engineering [39] because we were not able to recreate an implementation of system because we didn't have access to the data used for constructing and evaluating the models. We then identified possible biases of the ADM, and categorized and interpreted them using a chosen framework [26].

The public employment service was not transparent about the AMS algorithm, and this proved to be controversial. The initially published documents by the AMS turned out to be confusingly formulated, contradictory and selectively disclosed flattering statistics on error rates and precision in lieu of a detailed and balanced description [31]. One document described a logistic regression model that was not used for the classification. Against this backdrop, our analysis contributes to a better public and academic understanding because it can build on internal documents detailing the AMS algorithm's development and evaluation, providing crucial information that was previously unknown to the public.

3.3 Technical description

We found that the algorithmic system is comparatively simple and not a complex, modern statistical system, despite initially being portrayed as such. Before the categorization of job seekers into one of the three categories as outlined above, each individual's IC value is calculated based on a simple ratio between prior observations of job-seekers with the same or similar attributes: those that fulfilled either the long or short integration criterion versus those that didn't.

A matrix cross-relating a total of 13 variables models job seekers, including *gender*, *age*, *citizenship*, *education*, *health impairments*, *duties of care* for others, *job sector*, *assignment to a specific job center* and *prior employment history*. An individual is then compared to a historical group of job seekers with the same variable values. To explicate this procedure with a simplified example, a 35-year old woman with non-EU citizenship and no health impairments would be compared to all other 30 to 49 year-old women with non-EU citizenship and no health impairments within the prior four years. If, for instance, 83 out of a total of 100 persons with these same attributes did manage to find gainful employment for at least 3 months within the first 7 months of unemployment, the individual's short-term IC value would be given at 83%.

Given the many possible combinations of personal attributes and variable values, it is not surprising that–for a significant subset of job seekers–the number of comparative observations could be quite low, in some cases even less than 10. The system would then merge adjacent groups by joining certain variable values. The specific process of how these merges would occur was not disclosed. The variables were not only chosen due to the advertised goals of increasing accuracy, efficiency, and effectiveness. For example, they were meant to be convincing and explainable to different stakeholders and certain variables were omitted due to ethical concerns while others remained. Similarly, the thresholds applied to IC scores to produce the three risk categories were calculated using other constraints beyond the stated goals.

Further complicating the process is the fact that the total population of available observations was split further into those with a complete employment history within four years prior and those where this data was incomplete. The latter were again separated into individuals with a migration background, individuals under 25 years of age and the remaining individuals with incomplete employment history. This step was likely a measure to improve the accuracy / error rates of the predictions, and to make up for a lack of observations for certain sub-populations. Overall, the system was advertised as 80% accurate; however, this purported accuracy varies greatly depending on which subset of the populations individuals are assigned to, sinking as low as 69% for some combinations as indicated by the internal evaluations by creators of the AMS algorithm. There was no evaluation by external experts with access to the data used for constructing and evaluating the models, which calls the reported numbers into question. Prior work has also highlighted, how reductive performance indicators can hide problematic disparities disadvantaging marginalized groups and arbitrariness in classification [2, 29].

3.4 Analysis: Tracing Bias

We employ a long-established bias framework for computer systems by Friedman and Nissenbaum [26] to analyze the AMS algorithm and illustrate how its three basic categories for bias (technical, preexisting, and emergent) can be traced. We adapted the categories slightly to make them fit the context and added subcategories that we deemed useful for illustrating specific structural issues of the algorithm that could lead to harms, such as not receiving appropriate support measures when needed.

3.4.1 Technical Bias. This bias category is concerned with representational accuracy in light of "technical constraints or technical considerations" [26, p. 334] such as the abstraction, reduction, and decontextualization necessary in statistical modeling at scale [54]. This category provides a lens on flexibilities and tradeoffs decision makers have to consider, which highlights the non-innocence and

political character of their work [10]. It attributes accountability to human actors as design decisions are problematized. This can point to ways the system could be potentially improved [60] or how it has inherent fundamental issues that no design can remedy. In the context of the AMS algorithm, this flexibility refers to the decisions of involved actors such as on the data used and how categories are constructed, which influence which job seekers receive social support measures from the state. The following paragraphs categorize various biases and illustrate how long-standing problematic societal assumptions and established practices are reproduced through design decisions and thereby reify marginalization and enacting representational harms [8].

Rigidity and Coarseness of Variables: The variables assumed to influence job seeker chances are modeled based on a few, often dichotomous, categories. Consequently, these categories have to be coarse and encompassing to be convincing and allow for a classification scheme that can maintain the promise of capturing the complexity and variety of the Austrian labor market. For example, three age groups were used to categorize job seekers, which poses that people part of these groups have sufficiently similar integration chances to allow prediction. The variables fix fluid, continuous realities into supposedly stable and discrete categories. For example, the dichotomous health impairment variable absurdly oversimplifies disability and how it impacts possibilities for work in the labor market, disregarding difference, and representational justice in favor of supposed bureaucratic efficiency.

Rigid Categories: The categorization into risk groups is based on strict thresholds applied to calculated integration chance (IC) values. Thus, job seekers with IC values that only differ minimally may be assigned to completely different groups, which entail completely different treatment and support measures.

Uncertainty of Job Seeker Groups: As noted above, the AMS algorithm supposes that job seekers that share variable values have the same integration chance. Increasing the number of potential values also increases the number of groups, which is assumed to increase homogeneity within groups and differentiation between them. Yet since the number of job seekers is constant, when the number of groups increases so does the number of groups based on very few observations. According to the documentation, about 1900 groups are based on 50 or more observations and are thereby considered "statistically satisfactory," a term made up by the creators of the AMS algorithm. About 39% of job seekers are classified based on aggregated integration chances of groups that are considered "statistically not satisfactory." About 12% of job seekers are classified based on groups with less than 10 observations. Making predictions based on so few observations is concerning and a consequence of the chosen method. As a remedy, smaller ones are aggregated into bigger ones, but it was not made clear what criteria were used for merging. The system requires stable, historical data about job seeker groups in order to calculate integration chances. Since it only considers data collected over a period of 4 years, it cannot assess new job seekers that are part of groups that represent sets of values not encountered in that timeframe. For these groups no data would be available at all and, in turn, the integration chances for such job seekers would either have to be guessed or would need to be based on another group with available observations. Counselors don't know on how many observations a classification is based upon and thereby cannot consider this uncertainty in their assessment.

Privileged Perspectives: The data used to construct the model only represents a limited perspective on the labor market. It is only able to uncover correlations between essentialized characteristics of individual job seekers as recorded in governmental databases and periods for which job seekers use the services of the AMS. This partial perspective ignores, e.g., people that look for a job without being registered with the AMS. Furthermore, the correlations that can be uncovered in this setup only attribute unemployment to characteristics of job seekers. The data does not reveal, for instance, if certain industries or companies have problematic hiring practices, e.g., excluding women or other minoritized communities from open positions. The data does not contain any information on whether job seekers found a position they are happy with, either. Similarly, it is not able to uncover correlations that may point to problems with caseworkers at the AMS and the AMS management. The tool ultimately provides a perspective on the labor market that, based on calculated reasons for unemployment building on the available correlations, always insinuates that job seekers are the ones responsible for being unemployed.

Qualitative Factors: The AMS algorithm does not consider qualitative factors which potentially impact job seeker integration. For example, the intrinsic motivation of job seekers is ignored although it may impact reintegration chances significantly. Similarly, issues around appearance are sidelined, although it has been shown that they can be important factors in a job search and furthermore may result in discrimination [64].

3.4.2 Emergent Bias. This category seeks to capture bias that "arises in context of use [...] some time after a design is completed" [26, p. 335]. Put differently, this bias category foregrounds changing contexts, for example, due to the passage of time or use in spaces it was not intended for. This section identifies several bias categories we deemed relevant to the AMS algorithm. For example, how norms, laws, and representations are not stable over time but constantly change, which affects the reliability of the system and cannot be easily solved through updates as it may take time until these changes result in stable, measurable trends (if they ever do so). Consequently, job seekers are increasingly misrepresented over time. A desired "unbiased" state can never be reached, and the envisioned yearly updates cannot be considered a sufficient fix for these problems.

Disruptive events: The algorithm is updated yearly, and thereby, it assumes that the labor market does not change much in that year. Yet, recent examples, like COVID-19 or 2008 financial crash, highlight how disruptions happen regularly and often invalidate recent empirical labor market data as structures change quickly.

Changing Laws and Norms: Both laws and norms change, and can invalidate aspects of the system. For example, the historical databases don't account for the third gender option, which was recently legally recognized in Austria. This is also affects for the AMS Algorithm, as it only recognises a gender binary of male and female. With that, it either forces people not conforming to this binary into one of these categories or they simply cannot be processed. It is unclear how new versions of the system will handle this change because there are no historical records for people

identifying as neither male nor female to calculate their integration chance.

Interpretation and Interface: : The AMS algorithm's group determination likely deteriorates the quality of counseling, as caseworkers have very little time (sometimes only 10 minutes) [2] for engaging with job seekers and resolving issues with the calculated scores. The envisioned interface presents counselors a risk category and for job seekers with fully documented histories (about 70%) also the calculated integration chance can be accessed. These bits of information do not provide much room for interpretation and contestation for job seekers. In fact, internal instructional materials reveal that counselors are trained to convince job seekers of the objectivity of the algorithm when they question their own classification. After the introduction of the algorithm, counseling may center questionable classifications instead of job seekers' articulated needs. The interface is also meant to aid counselors in this convincing by providing a set of a few rudimentary explanations for certain scores that can be presented to worried job seekers. Most of these are not very descriptive and some of them even reproduce racist and sexist stereotypes, e.g., only women (but not men) with children are warned about their childcare responsibilities negatively impacting their integration chances. Previous work highlights how the discretion caseworkers have for interpreting the interface and possibly changing group assignments could also further reinforce disparities if caseworkers hold discriminatory beliefs [28].

3.4.3 Preexisting Bias. This form of bias has its "roots in social institutions, practices, and attitudes" and "exist[s] independently, and usually prior to the creation of the system" [26, p. 334]. In practice, it embodies prejudices, held values, established practices and current social orders that seep into algorithmic systems through actions of institutions or individuals. In our case, this analytical bias category sensitizes us to the ways that contemporary injustices and historical inequalities on the labor market and the AMS as a public institution are inscribed in the algorithm and reinforced when it is used in decision making. Thus, this analysis problematizes these inequalities, as the desired "unbiased" norm or center in this category would be an algorithm that does not allocate fewer resources to people based on historical injustices. Instead, a "less" biased algorithm according to this category would seek to remedy and repair these injustices, for example, by allocating more resources to marginalized people, so they don't negatively affect people's opportunities on the labor market.

Variables as direct/indirect Proxy for Inequality: The data used to construct risk scores also accounts for historical disadvantage and discrimination in the labor market and by the AMS. Thus, the calculated risk of long-term unemployment of marginalized groups affected by these mechanisms is higher, which means fewer resources for these groups with the danger of a self-reinforcing loop algorithmically optimized to exacerbate inequality. The effects of these injustices are accounted for in the model through variables that directly model marginalized groups, e.g., women are a category of the gender variable, and also indirectly, e.g., the area an unemployed person lives is also a variable and correlates with socio-economic opportunity.

Multiple Models as Proxy for Inequality: For some job seekers, the data for four prior years is not complete. For these individuals, the designers of the system created three separate models based on particular variable values. The category for "people with migration background," for example, contains people of foreign nationality, naturalized citizens, and people who have at least one parent of foreign nationality, three very different circumstances. It is not completely clear why these lines of separation were chosen. The separation concentrates many job seekers with calculated high risk of long-term unemployment in one model, which means that new job seekers classified as "people with migration background" are likely to be also classified as high risk. The effects of the segmentation also partially align with stated policy goals of the government under which the AMS algorithm was introduced, to focus less on the support and integration of refugees and more on other groups[3].

4 FRAMING BIAS: UNPACKING THE USE OF BIAS FOR ANALYSIS AND CRITIQUE

Based on our experience from the case study presented in the previous section, we reflect on the implications of using bias and bias frameworks to analyse and problematize algorithmic systems. We discuss how bias can act as a boundary object, enabling cooperation and deliberation across contexts. In section 4.3, we unpack benefits of using bias frameworks, potential issues, and discuss ways forward.

4.1 Criticism of bias

Various technology studies scholars have recently problematized the language of bias for its often narrow technical framing and argued instead for a focus on broader analyses of social structures. We agree with these calls in principle and emphasize the importance of critically engaging with the concept and its uses. In light of these debates around the usefulness of bias, we think it is instructional to look at how other fields dealt with similar issues. For example, surveillance scholars researched and problematized the capture of privacy by industry interests [15], highlighting an over-individualization of the concept focused on reductionist notions of consent. In response, scholars produced foundational critiques of privacy while also strategically using the concept and seeking to redefine it to center collective and public interests [15, 49]. In a similar manner, we advocate for foundational critiques of the concept and for strategic and critical engagement with bias.

The way bias can be meaningful for critique becomes visible when concerns of actors are examined that aim to hide problematic biases and their harms. For example, Meta reportedly banned employees from using the terms 'bias' and 'discrimination' when discussing their algorithms to avoid possible liabilities [41]. The example highlights how the articulation of a bias can pose such a challenge to problematic practices and narratives around the assumed functionality of systems [57] that companies may even try to preempt it. In the remainder of this section, we argue for politizing bias and point to its emancipatory potential by highlighting the possibilities of its strategic use as a boundary object [43, 69]. Furthermore, drawing on our analysis presented in the previous section, we discuss benefits and risks of using bias frameworks to aid in identifying, constructing, and explaining biases of algorithms and highlighting how they could be improved.

4.2 Bias as enticing Boundary Object

The articulation of bias can enable cooperation across community boundaries because bias can be meaningfully constructed to act as a boundary object [69]. Such boundary objects are defined as "adaptable to different viewpoints and robust enough to maintain their identity across them" [69] and thereby enabling "different groups to work together without consensus" [43]. This ability of bias frames to cross boundaries is also enabled by the regular usage of the term in many different professional and public communities and its authority through its association with scientific practice [21].

In the FAccT community, auditing highlights how bias can act as a boundary object. For example, auditors can bring algorithmic bias into existence as a tangible number through measurement. This bias can then be further contextualized, transformed, theorized, and problematized by interest groups with their own local knowledges and the intention to elicit change and challenge injustices. It may also be debated in courts where negotiations take place how it relates to existing law. As the bias identified and constructed by the auditors travels, it maintains some identity while its interpretative flexibility enables other groups to adapt meanings relevant to the contexts they engage with and thereby make it useful to them. The bias acts as a boundary object until closure is reached when deliberations such as around its explanation and significance are settled [55, 67]. This process of closure involves also power as different actors try to make their interpretations convincing and enduring, and, thus, may result also in unjust outcomes stabilizing problematic and harmful notions of bias or its absence.

The narrative around the AMS algorithm as a modern, impartial statistical tool for the allocation of scarce resources was initially not perceived as highly controversial in the public discourse. The controversy took off after coefficients in a publicized regression analysis were interpreted by interest groups and scholars. They pointed to the fact that being a woman in the system was a risk factor that increased the chances of long-term unemployment and categorization as part of group C. This reading undermined the promised objectivity of the system and encouraged and mobilized various interest groups, journalists, academics, civil rights organizations, gender equality activists, and legal professionals to further investigate and interpret the document, the biases it may show, and question the claims of the AMS. The numbers, not intended to be indicative of bias, were reinterpreted as such and then became a boundary object around which all these disparate groups came together to unearth issues such as discrimination and contribute their perspectives. This example highlights the potential of boundary objects for critique, and we hope it also shows how bias can be useful, if strategically and thoughtfully applied, as a boundary object for emancipatory endeavors.

The interpretative flexibility of bias can also foster controversies and conflict, especially when opposing conceptions and interests meet. For example, an understanding of bias as the opposite of 'objectivity' produces disagreements with those that question this antagonism altogether. Advocates of the first definition frame biases as merely technical and thus seek to fix them to reach some imagined 'unbiased' or objective state. In contrast, adherents to the second definition push back against such bias conceptions because they stabilize technological frames that make an imagined 'unbiased' state seem truthful and thereby incontestable.

Whether a bias or harm is recognized as such in different contexts is also a question of power[29]. As different actors influence debates to further their interests, those with a vested interest in the technology and the politics it enacts are often in positions of power. They are incentivized to keep certain biases and harms hidden, don't give them their attention, or push reductive notions of bias to stabilize the technology[29]. Such endeavours involve unmaking bias as a boundary object that acts as a site for critical and emancipatory debate and instead produces closure to end debates and stabilizes one perspective. In such cases, confrontations between opposing positions are necessary to challenge problematic frames that, e.g., make the adoption of a problematic ADM seem inevitable and foreclose alternatives. We think this is the case for the AMS algorithm and discuss this further below in our reflection on bias frameworks.

In other cases where needs and desires addressed by an algorithm are less contested, bias can also act as a useful boundary object for reformist projects. For example, bias in health care is often widely understood as distortion [56], which can be a useful frame to work across boundaries and suggest improvements to different people. We think in such cases a bias framing may be also useful because it is comprehensible to the general public and different scholarly communities. It may even be required in some cases to enable a public debate at all. Some communities ascribe to certain notions of objectivity and may not engage with other framings. In the next section, we reflect on using frameworks to identify and construct biases and the benefits and risks of this practice.

4.3 Reflections on using Bias Frameworks

We understand bias frameworks as classification schemes that name and exemplify potential sources of bias to aid auditors and other critical analysts in identifying issues with a technology. Yet, these frameworks also come with the danger of reifying and stabilizing certain bias categories developed with a certain stance and use cases in mind. We highlight how these frameworks are partial and political by comparing several bias frameworks [26, 48, 53, 71] and pointing to differences in how they conceive of biases and discuss how they could be rethought.

4.3.1 Explanations for Bias. Bias frameworks provide explanations for biases by, e.g., highlighting how and where they occur, and structure them into bias categories. The construction of such bias categories requires interpretative and normative work as it entails intentional decisions about, e.g., how a bias is bounded/framed (including the imagined norm/expectation from which it deviates), how it is explained, and what potential concerns it may entail. In turn, there are many different ways explanations for bias could be classified [55].

Friedman & Nissenbaum [26] introduced three coarse bias categories and several subcategories, some of which arguably are less relevant to contemporary critiques of algorithms (e.g., issues with randomization). This is not surprising given the time of publication and the different technological landscape then. More recent bias frameworks[48, 53, 71] also have a category for evaluation bias that

captures problems in testing, which is absent in the Friedman & Nissenbaum framework but would be an important extension.

We caution to simply take bias categories provided by frameworks uncritically because at times they provide misleading explanations. For example, the term 'technical bias' in [26] may help in identifying limits of the technologies used, but the term 'technical' tends to also make invisible the decisions that have to be made – and potential values carried by these decisions – when developing an algorithmic system. As Jaton [34] nicely shows, many technologists are "reluctant to make hesitations and uncertainties visible" (p. 8) that have an influence on the possible forms of an algorithmic system, most of which are (eventually) not realised. We consider 'socio-technical' bias, for example, as a more appropriate name for this category as it sensitizes analysts to see perceived technical limitations also as the result of human judgement.

We argue that frameworks should not focus solely on technical issues and trade-offs as these overemphasize decisions of the individual developers. As our study has shown, policy, social conditions, discourse, and organizational rules also put bounds on viable options. For example, the variables and their possible values had to be available in governmental databases and familiar to caseworkers and other actors within the Austrian public employment service. The frameworks we examined did not contain a category about bias that emerges due to decision making of higher-ups or other organizational boundaries, which may foreclose important paths toward more just outcomes. The frameworks also depoliticize biases by making them appear as though they often result from individual decisions, mistakes, or pitfalls [29]. This moves attention away from questions of power, injustice, positionality, limitations, and organizational responsibility.

Bias frameworks also imply different theories about how bias circulates. One framework [71] depicts biases as seeping into algorithms within a linear ML development pipeline that ends with deployment. The authors acknowledge that, in practice, development is more messy and iterative, involving feedback loops. In contrast, the bias framework by [48] identifies three interrelated sources of bias in machine learning: the data, algorithm and user interaction, and connects them through a loop. In the resulting system, biases travel in a feedback loop, from data to algorithm to user interaction and back to data, to illustrate how biases can be self-reinforcing. This model does not address how bias initially became part of the system, which may imply that bias is always there in some capacity. In such circular, homeostatic models, a theory of change and agency is absent. Lastly, the bias framework we chose [26] provides a list of categories, which we found most useful for our case. It provides more flexibility and comes with few assumptions about processes that are hidden to us as external researchers.

4.3.2 Granularity of Bias Frameworks. The Friedman & Nissenbaum framework [26] provided us with rough guidance for the analysis of the ADM at hand. As a consequence of following 'only' three broad categories of bias (cf. sections 3.4.1–3.4.3), it was not possible to simply *use* the framework in a template-like fashion, but it had to be significantly adjusted. We neglected the subcategories since many of them did not fit our case. The structure the three categories provided aided us in constructing different bias categories in an orderly manner, while the flexibility encouraged close examination of the different parts of the ADM to come up with nuanced subcategories that fit the context and case well.

Several of the more recent bias frameworks we looked at comprised more categories that were more nuanced. They were designed for specific contexts and technologies that were a bad fit for our case study. Furthermore, we found that some frameworks have bias categories with similar meanings but slightly different labels and emphases on certain aspects. For example, preexisting bias [26] is akin to historical bias [48, 71] as both problematize injustice that seeps into technology, but the latter is more specific and only concerned with data (generation). Sometimes multiple categories may represent one category in another framework, e.g., representation, measurement, aggregation bias [71] are similar to technical bias [26]. Representation bias could also be categorized as preexisting bias [26] if the under-sampling of one group is due to historical disadvantage. This further highlights how blurry these abstract categories that consequently require interpretative work by the analyst are. Finally, some frameworks lacked categories that were important to our analysis of the AMS algorithm, such as, e.g., emergent bias [71].

Bias frameworks that put a focus on high granularity and specificity can have obvious benefits: Fine-grained bias categories can aid in identifying these types of biases, and it can be particularly useful for novice researchers. It can be an opportunity to allow for a bias framework to account for more variability through specificity. By using more granular categories, researchers can also build on top of previous research more easily, e.g., by observing if certain types of biases frequently occur in a particular type of application or if they are hard to spot. However, we fear that the more nuanced bias frameworks become, the more rigid they become. Metaphorically speaking, by shining a the available light on certain parts of an algorithmic system, they can keep other parts in a deeper shadow. With increased granularity, it is easier to fall into the trap of simply following the framework and using it as a kind of checklist.

Coarse frameworks encourage the researcher to adapt the framework and resist simple 'application.' In other words, they put more responsibility on the side of the researcher scrutinizing an algorithm, whereas more granular frameworks put more responsibility on the framework and its designers. The coarseness can also aid in communication, as extensive bias lists and too much detail may be overwhelming and hard to grasp in public outreach. One way to productively engage with these tensions around granularity could be frameworks with multiple levels of granularity that go into more depth but also have big categories. We still think, in many cases, it probably would be best to devise biases and explanations for different audiences and, while doing that, also reflect on the granularity of bias categories.

4.3.3 Technological Focus of Bias. There is an inherent limitation to bias frameworks: They focus on biases. While this statement may seem trivial, it prompts the question: What else is there to investigate besides biases? The framing of bias comes with the risk of focusing the analysis mostly on the ADM as a technology. For example, one framework concerned with measurement and algorithms based on social media data [53] framed all bias categories in terms of validity while acknowledging at the end that ethics are also

important. This is a problematic stance as there is no one objective way to describe the social and all its contingencies, thereby, ethical and political considerations need to be also part of bias analyses. For our case, it would suggest that there is a "correct" or "optimal" way to sort people into categories and assign them support measures. Yet, social sorting based on risk itself is an inherently political endeavor where no absolute and "correct" solutions exist.

Biases can distract from big-picture questions, such as other sources of injustice or operational and organizational aspects, or whether an algorithm should or can be built at all. A prime example is the algorithmic system's objective. In our case study, the objective is to distribute resources among job seekers, but in a way that is substantially different to most similar systems outside of Austria [3], namely by focusing on job seekers that promise the best return on investment of resources, not those who are most vulnerable. We identify biases and show how the system does not deliver on its promises of objectivity, thereby challenging the narrative of its adoption as inevitable technological progress. This kind of analysis does not necessarily question the fundamental, problematic premises of an objectification of austerity politics.

To conclude, merely identifying bias categories existing in a system ought to be only a starting point that can enable deeper investigations. For example, our analysis also highlights how injustices in the labor market factor into calculated risk scores. This insight challenges the intended use of the AMS algorithm as it shows how it, by definition, reproduces injustices if no mechanisms are put in place to counter these tendencies. Thus, the fundamental premise of the ADM is called into question, and the issue of how to deal with deep societal inequalities becomes a topic of debate, highlighting how structural critiques can start with or be supported by a bias analysis. Bias frameworks can aid in this process, but in themselves they are no guarantee that foundational questions will be raised.

5 CONCLUSION

In this paper, we presented a critical analysis of the AMS algorithm based on new information about its inner workings. We identified and constructed various biases that illustrate harms to job seekers and challenge promises used to justify the ADMs adoption. Based on this work, we unpacked the implications of using bias and frameworks for analysis to critique ADMs. We explained how articulated biases can be emancipatory and strategically used as boundary objects to build critiques without clear consensus of all parties involved. Despite the concerns we voiced on the use of bias frameworks, we think they can provide useful sensibilities for identifying and diagnosing problems using ADMs as a starting point when used appropriately. We believe these frameworks will likely remain relevant in some capacity in computing and therefore should receive more critical attention from scholars. Like any kind of classification scheme [10], they will also always be lacking in certain respects, in turn, we argue for a pragmatic stance towards them.

We discussed trade-offs between coarseness and specificity of categories in frameworks and how they may impact later analysis that builds on top of them. The explanatory capabilities of bias categories were discussed, highlighting how many categories focus only on the technology and often situate responsibility with the algorithm designers. This focus is concerning as it individualizes problems with ADMs instead of also centering organizational and other social causes, and interrogating premise and limitations of systems. We argued that biases should be seen as a starting point for further analysis, such as what are deeper historical and social explanations for their emergence, what underlying power dynamics stabilize them, and what can be done to counteract them. We also think interrogating a bias can enable fruitful conversations around what "unbiased" or natural states are assumed or desired, and if some of the biases should be upheld or abandoned. When bias frameworks are used, it is important for analysts to be cautious about what they make invisible and prioritize, and the context the framework was designed for, to ensure analysis is critical and reflexive. Similarly, it is important for analysts to be aware of their own positionalities and explore how the constructed categories fit their case so that they can be made actionable and communicable.

The worrying neglect of categories and explanations grounded in social theories in the frameworks we reviewed above highlights a need for further critical research in this area. We argue that as such frameworks are made and reworked, they should make their perspective and epistemological commitments clearer, focus on more concrete phenomena instead of abstract terms like bias, foreground questions of power and justice, and draw from more relevant fields concerned with the 'social' such as Anthropology, Sociology, and Science and Technology Studies. This also entails not referring to all introduced categories through a combination of the name of a phenomenon and the suffix "bias," but instead also using more theoretically-informed descriptors that acknowledge the political nature of these categories. For example, analysts could draw on developed literature like agnotology and illustrate the co-production of different forms of ignorance through the use and evaluation of certain algorithms[29]. We argue frameworks should be constructed through more participatory approaches that seek to center the perspectives of marginalized communities [11, 47, 68].

There is a risk that these frameworks stabilize certain problematic notions of bias, for example, when they are enshrined in law. The stakes are high in matters of bias and harm, so it is important that possibilities for emancipatory debate remain open and that frameworks are circulated with this in mind. This entails constructing bias categories so they can act as useful boundary objects. There is also a risk of frameworks just becoming another piece of paper that is mostly ignored. We think these frameworks can also encourage debate and reflection among those involved in the creation of an algorithm. They can mark points in the development process where things could have been different, where important politics are played out, where creeping inequalities should have been addressed - and when an algorithm just is not a viable tool. Thus, we think future work should seek to both critique and improve them, and empirically study their use and impact through empirical methods such as participant observation.

ACKNOWLEDGMENTS

We are thankful to Christian Sandvig and the Infrastructure Lab at the University of Michigan for their comments on earlier drafts and to the reviewers and chairs for their useful feedback.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain, 2020-01-27) *(FAT* '20)*. Association for Computing Machinery, 252–260. https://doi.org/10.1145/3351095.3372871

[2] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. 3 (2020). https://doi.org/10.3389/fdata.2020.00005

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS Algorithmus-Eine Soziotechnische Analyse Des Arbeitsmarktchancen-Assistenz-Systems (AMAS). (2020). https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf

[4] Louise Amoore. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others.* Duke University Press, Durham.

[5] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and Don'ts of Machine Learning in Computer Security. In *Proc. of USENIX Security Symposium* (2022). 18.

[6] Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion.* Association for Computational Linguistics, Kyiv, 34–45.

[7] Agathe Balayn and Seda Gürses. 2021. *Beyond Debiasing: Regulating AI and Its Inequalities.* Technical Report. EDRi Report. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

[8] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]* (May 2020). arXiv:2005.14050 [cs]

[9] Glenn A Bowen. 2009. Document analysis as a qualitative research method. *Qualitative research journal* 9, 2 (2009), 27–40.

[10] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things out : Classification and Its Consequences.* MIT Press.

[11] Sharon Brisolara. 1998. The history of participatory evaluation and current debates in the field. *New directions for evaluation* 1998, 80 (1998), 25–41.

[12] Florian Cech. 2021. The agency of the forum: Mechanisms for algorithmic accountability through the lens of agency. *Journal of Responsible Technology* 7-8 (2021), 100015. https://doi.org/10.1016/j.jrt.2021.100015

[13] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 77–89. https://doi.org/10.1145/3442188.3445872

[14] Adele E. Clarke, Carrie Friese, and Rachel S. Washburn. 2017. *Situational Analysis: Grounded Theory after the Interpretive Turn.* Sage Publications.

[15] Sami Coll. 2014. Power, knowledge, and the subjects of privacy: understanding privacy as the ally of surveillance. *Information, Communication & Society* 17, 10 (2014), 1250–1263.

[16] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: An Algorithmic Bias Effort in Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow Scotland Uk, 1–8. https://doi.org/10/gnw5xw

[17] David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

[18] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. 2019. *Statistical profiling in public employment services: An international comparison.* OECD Social, Employment and Migration Working Papers 224. https://doi.org/10.1787/b5e5f16e-en Series: OECD Social, Employment and Migration Working Papers Volume: 224.

[19] Sam Desiere and Ludo Struyven. 2021. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy* 50, 2 (April 2021), 367–385. https://doi.org/10.1017/S0047279420000203

[20] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press.

[21] William Eadie. 2009. *21st Century Communication: A Reference Handbook.* SAGE Publications, Inc. https://doi.org/10.4135/9781412964005

[22] epicenter.works Plattform Grundrechtspolitik. 2022. Stoppt den AMS Algorithmus. https://amsalgorithmus.at/

[23] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press. 1–221 pages.

[24] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. 2022. A survey on bias in visual datasets. *Computer Vision and Image Understanding* 223 (2022), 103552.

[25] Ed Finn. 2017. *What Algorithms Want.* MIT Press.

[26] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. 14, 3 (1996), 330–347.

[27] Gabriel Grill and Nazanin Andalibi. 2022. Attitudes and Folk Theories of Data Subjects on Transparency and Accuracy in Emotion Recognition. 6 (2022), 78:1–78:35. Issue CSCW1. https://doi.org/10.1145/3512925

[28] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[29] Gabriel Grill. 2022. Constructing certainty in machine learning: On the performativity of testing and its hold on the future. *OSF Preprints* (2022). https://doi.org/10.31219/osf.io/zekqv

[30] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. 22, 7 (2019), 900–915. https://doi.org/10/gf2m8v

[31] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. *Das AMS-Arbeitsmarkt chancen-Modell.* Technical Report. http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

[32] Pat Irving and Dr Isabelle Deganis. 2015. Dr Sally-Anne Barnes and Sally Wright Institute for Employment Research University of Warwick Coventry CV4 7AL Sally-Anne.Barnes@warwick.ac.uk. *Final report* (2015), 117.

[33] Sheila Jasanoff. 2004. Ordering knowledge, ordering society. Routledge.

[34] Florian Jaton. 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society* 8, 1 (Jan. 2021), 1–15. https://doi.org/10.1177/20539517211013569

[35] Pratyusha Kalluri. 2020. Don't Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power. 583, 7815 (2020), 169–169. Issue 7815. https://doi.org/10.1038/d41586-020-02003-2

[36] Christoph Kern, Ruben L. Bach, Hannah Mautner, and Frauke Kreuter. 2021. Fairness in Algorithmic Profiling: A German Case Study. *arXiv:2108.04134 [cs, stat]* (Aug. 2021). http://arxiv.org/abs/2108.04134 arXiv: 2108.04134.

[37] Rob Kitchin. 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. (Dec 2014), 1–285.

[38] Rob Kitchin. 2016. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (02 2016), 14 – 29. https://doi.org/10.1080/1369118x.2016.1154087

[39] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (Jan. 2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[40] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? *arXiv:2105.01441 [cs, stat]* (May 2021). http://arxiv.org/abs/2105.01441 arXiv: 2105.01441.

[41] Issie Lapowsky. 2021. *Facebook Told Workers to Avoid the Words 'discrimination' and 'Bias'.* Protocol. https://www.protocol.com/policy/facebook-papers-fairness

[42] Iwona Laub. 2020. Warum der polnische AMS-Algorithmus gescheitert ist. https://epicenter.works/content/warum-der-polnische-ams-algorithmus-gescheitert-ist

[43] Susan Leigh Star. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, technology, & human values* 35, 5 (2010), 601–617.

[44] Paola Lopez. 2019. Reinforcing intersectional inequality via the AMS algorithm in Austria. In *Critical Issues in Science, Technology and Society Studies. Conference Proceedings of the th STS Conference (Graz: Verlag der Technischen Universität).* 1–19.

[45] Paola Lopez. 2021. Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review* 10, 4 (Dec. 2021). https://doi.org/10.14763/2021.4.1598

[46] Artan Loxha and Matteo Morgandi. 2014. Profiling the Unemployed: a review of OECD experiences and implications for emerging economies. (2014).

[47] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2020-04-21). ACM, 1–14. https://doi.org/10.1145/3313831.3376445

[48] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[49] Torin Monahan and David Murakami Wood (Eds.). 2018. *Surveillance Studies: A Reader.* Oxford University Press.

[50] Martin Müller and Carolin Schurr. 2016. Assemblage thinking and actor-network theory: conjunctions, disjunctions, cross-fertilisations. *Transactions of the Institute of British Geographers* 41, 3 (2016), 217–229. https://doi.org/10.1111/tran.12117

[51] Jędrzej Niklas, Karolina Sztandar, and Katarzyna Szymielewicz. 2015. *Profiling the unemployed in Poland: Social and political implications of algorithmic decision making.* Technical Report. Fundacja Panoptykon, Warsaw. 51 pages.

[52] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019), 13. https://doi.org/10.3389/fdata.2019.00013

[53] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. 2 (2019), 13. https://doi.org/10.3389/fdata.2019.00013

[54] Sebastian Pfotenhauer, Brice Laurent, Kyriaki Papageorgiou, and and Jack Stilgoe. 2022. The Politics of Scaling. 52, 1 (2022), 3–34. https://doi.org/10.1177/03063127211048945

[55] Nikolaus Poechhacker and Severin Kacianka. 2021. Algorithmic accountability in context. Socio-technical perspectives on structural causal models. *frontiers in Big Data* 3 (2021), 519957.

[56] Mirjam Pot, Nathalie Kieusseyan, and Barbara Prainsack. 2021. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights into imaging* 12, 1 (2021), 1–10.

[57] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[58] Lisa Reutter and Hendrik Storstein Spilker. 2020. The Quest for Workable Data Building Machine Learning Algorithms from Public Sector Archives. In *The Democratization of Artificial Intelligence. Net Politics in the Era of Learning Algorithms*. transcript Verlag, Bielefeld, 95–107.

[59] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[60] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022-06-20) *(FAccT '22)*. Association for Computing Machinery, 2138–2148. https://doi.org/10.1145/3531146.3534631

[61] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (Dec. 2017), 205395171773810. https://doi.org/10.1177/2053951717738104

[62] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598

[63] Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468 arXiv:1912.11078

[64] Eva Sierminska. 2015. Does It Pay to Be Beautiful? *IZA World of Labor* (2015).

[65] Selena Silva and Martin Kenney. 2018. Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon (1960-)* 55, 1 & 2 (2018), 9–37.

[66] David Silvermann. 2015. Interpreting Qualitative Data. In *Methods for Analyzing Talk Text and Interaction* (5th edn. ed.). Sage.

[67] Sergio Sismondo. 2010. *An Introduction to Science and Technology Studies*. Vol. 1. Wiley-Blackwell Chichester.

[68] Katta Spiel, Laura Malinverni, Judith Good, and Christopher Frauenberger. 2017. Participatory Evaluation with Autistic Children. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017). 5755–5766.

[69] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

[70] Norman Makoto Su, Amanda Lazar, and Lilly Irani. 2021. Critical affects: Tech work emotions amidst the techlash. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[71] Harini Suresh and John V. Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization* (Oct. 2021), 1–9. https://doi.org/10.1145/3465416.3483305 arXiv:1901.10002

[72] Wim Van Lancker. 2020. Automating the Welfare State: Consequences and Challenges for the Organisation of Solidarity. In *Shifting Solidarities*. Springer International Publishing, Cham, 153–173.

[73] Bert Van Landeghem, Sam Desiere, and Ludo Struyven. 2021. Statistical profiling of unemployed jobseekers. *IZA World of Labor* (2021). https://doi.org/10.15185/izawol.483

[74] Salomé Viljoen. 2021. The promise and limits of lawfulness: Inequality, law, and the techlash. *Journal of Social Computing* 2, 3 (2021), 284–296.

[75] Ben Wagner. 2018. Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In *BEING PROFILED: COGITAS ERGO SUM*. Amsterdam University Press, 84–89.

[76] Linan (Frank) Zhao. 2020. Data-Driven Approach for Predicting and Explaining the Risk of Long-Term Unemployment. *E3S Web of Conferences* 214 (2020), 01023. https://doi.org/10.1051/e3sconf/202021401023