Auditing Risk Prediction of Long-Term Unemployment

CATHRINE SEIDELIN, University of Copenhagen, Denmark
THERESE MOREAU, Zetland, Denmark
IRINA SHKLOVSKI, University of Copenhagen, Denmark
NAJA HOLTEN MØLLER, University of Copenhagen, Denmark

As more and more governments adopt algorithms to support bureaucratic decision-making processes, it becomes urgent to address issues of responsible use and accountability. We examine a contested public service algorithm used in Danish job placement for assessing an individual's risk of long-term unemployment. The study takes inspiration from cooperative audits and was carried out in dialogue with the Danish unemployment services agency. Our audit investigated the practical implementation of algorithms. We find (1) a divergence between the formal documentation and the model tuning code, (2) that the algorithmic model relies on subjectivity, namely the variable which focus on the individual's self-assessment of how long it will take before they get a job, (3) that the algorithm uses the variable "origin" to determine its predictions, and (4) that the documentation neglects to consider the implications of using variables indicating personal characteristics when predicting employment outcomes. We discuss the benefits and limitations of cooperative audits in a public sector context. We specifically focus on the importance of collaboration across different public actors when investigating the use of algorithms in the algorithmic society.[1]

CCS Concepts: • **Human-Centered Computing → Collaborative and social computing**; *Empirical studies in collaborative and social computing*

KEYWORDS: Audit, Public Services, Algorithm, Job Placement, Accountability.

ACM Reference format:

Cathrine Seidelin, Therese Moreau, Irina Shklovski, and Naja Holten Møller. 2022. Auditing Long-Term Unemployment. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 6, GROUP, Article 08 (January 2022), 14 pages, https://doi.org/10.1145/3492827

1 INTRODUCTION

In 2019, a young Danish woman who had just graduated logged on to the web application Jobnet.dk[1], a website for jobseekers and employers in Denmark, where unemployed individuals can register to receive unemployment benefits. During her first encounter with Jobnet, the recent graduate was asked to fill out a questionnaire about her employment and education history as

This work is supported by the Innovation Fund Denmark, under grant 7050-00034A and the Independent Research Fund Denmark, under grant 8091-00025b.
Author's addresses: C. Seidelin, I., Shklovski and N. Holten Møller, Department of Computer Science, University of Copenhagen, Sigurdsgade 41, 2200 Copenhagen, Denmark; T. Moreau, Zetland, Njalsgade 19D, 1st floor, 2300 Copenhagen, Denmark.

[1] https://info.jobnet.dk/om-jobnet/jobnet-in-english

2573-0142/2022/1 - Art8
https://doi.org/10.1145/3492827

well as her expectations for how fast she would find a job. While completing the questionnaire, the woman noticed that some of the data had been pre-filled, identifying her as a descendent of parents of non-Western2 origin. As a Danish citizen, she wondered why her Kurdish origin was of any relevance to her application for unemployment benefits. The caseworker she met with did not know why or how the system used this information, and she also considered it irrelevant [20]. The case prompted public debate on the use of algorithms and Artificial Intelligence (AI) and the role of algorithmic profiling in the public sector, resulting in legal challenges and media attention. On the one hand, the Danish unemployment services agency argues the STAR algorithm should be regarded as a voluntary supplement to support the individual's interaction with their local jobcenter. Moreover, they argue it is not mandatory for an individual to be profiled to collect social benefits. On the other hand, critics question whether the unemployed individual understand their right to choose not to be profiled [9].

Governments increasingly use algorithms to support decision-making that affects human lives, including, for example, where to send the police [13, 21] and whether people qualify for social benefits [8] or are at risk of long-term unemployment [2]. Governments' excitement around algorithms as tools for decision-support relates, in part, to the promise of enhanced performance and efficiency of public services [6, 39]. However, the growing use of algorithms in this context has also introduced serious concerns related to unfair and biased outcomes [5]. The prevalence and growing impact of algorithms in the public sector, and society more broadly, calls for ways that enable individuals and organizations to examine and question them. 'Algorithmic audits' are approach to do so [22, 26, 29]. An algorithmic audit is the practice of collecting data about how an algorithm behaves in a particular context, and further analyzing that data to determine whether this behavior negatively impacts the people who are affected by that algorithm [7]. Alternative approaches to audits include examinations of documentation and code reviews [4].

We conduct a cooperative audit [41] of what is often referred to as "the STAR algorithm" (from the abbreviation of the Danish Agency for Labor Market and Recruitment – in Danish: Styrelsen for Arbejdsmarked og Rekruttering, hereafter 'STAR'). The STAR algorithm is the core of a profiling tool that Denmark implemented in job centers nationwide in 2015. This profiling tool was formally presented to support the unemployed in their preparation for the job-seeking process and as a 'starting point' for the dialogue between the unemployed individual and their caseworker at the local job center [36]. In practice, the profiling tool uses a decision tree algorithm to create "data profiles" of unemployed citizens based on personal data. The stated purpose of the tool – and the underlying STAR algorithm – is to predict whether an unemployed citizen is at risk of long-term unemployment. If the algorithm designates an individual at risk of long-term unemployment, this impacts what the unemployment office expects of them (e.g. more meetings with a caseworker) [15]. Following the intense public debate, the Danish Institute for Human Rights filed a complaint against the Danish Agency for Labor Market and Recruitment (hereafter, "the agency") to the Danish Council for Equality in 2020. The complaint contained two main allegations: First, it argues the STAR algorithm's use of the variable "origin" for profiling unemployed individuals is in violation of the law prohibiting discrimination in the labor market,

[2] The category 'non-Western' is produced and used by Statistic Denmark, the central authority on Danish statistics. Non-Western countries include some European countries (e.g., Albania, Belarus, Turkey, and Ukraine). All countries in Africa, South and Central America and Asia. All countries in Oceania (except Australia and New Zealand) as well as stateless [17].

and second, it questions the assumption that individuals understand it is voluntary to be profiled [9]. It is now a legal trial which will be settled in 2022 and act as a case example for how public services make use of data-driven technologies for decision-making.

To understand and explain the STAR algorithm, we obtained the model tuning code and the documentation and conducted an in-depth analysis of both. According to the documentation, the algorithm is intended to predict long-term unemployment based on six variables: (1) the individual job seeker's self-identified potential for employment, (2) origin, (3) age, (4) employment rate the prior 36 months, (5) income level the past year, and (6) educational background (although the 5th and 6th variable do not influence the algorithm's predictions under certain circumstances, as we show in section 4). Our audit demonstrates an inconsistency between the documentation and the source code. Moreover, our audit demonstrates the cooperative challenges involved in accounting for the algorithm's inner working when stakeholders have contested assumptions of, for example, what counts as documentation for how variables were selected for the algorithmic model. The STAR algorithm is a case that allows us to develop a better understanding of what kinds of documentation and forms of collaboration are necessary to perform a cooperative audit in a public service context.

The main contributions of this note are: First, an empirical account of how a cooperative audit may be conducted of a public service algorithm. Second, our account demonstrates that established practices for documentation of algorithms (e.g., compliance requirements, conformity assessments, and certification approaches) do not preclude the necessity of developing approaches to auditing that ensure better, more transparent, and more accountable algorithmic systems in governmental operations.

2 RELATED WORK: AUDITING PUBLIC SECTOR ALGORITHMS

The context for this study is unemployment policy in Denmark and the use of AI and algorithms in the public sector more broadly. From the unemployed individual's perspective, Danish job placement is centered around the web application Jobnet. To receive unemployment benefits, an individual must register their unemployed status in Jobnet and participate in regular meetings with a caseworker. It is the responsibility of the individual to book meetings with the caseworker using the Jobnet portal and to update their 'job log,' providing the caseworker with an overview of the individual's job search activities.

Unemployment constitutes a significant cost to the Danish welfare state (and governments in general) and it is therefore in the interest of the state to ensure that newly unemployed get back to work [34]. The Danish agency relies on OECD's definition of long-term unemployment, which refers to people who have been out of work for 12 months or more [24, 34]. The development of the STAR algorithm was initiated in 2014 as a result of the adopted Employment Reform, which (amongst other things) aimed to reduce the unemployment rate by decreasing the time it takes for a person to obtain permanent employment [36]. To address this issue, the Danish Ministry of Employment called for a solution that could predict an individual's risk of long-term unemployment [36]. This political initiative reflected tendencies in many other countries, for example, Australia, USA, France, Sweden, and Germany, who also had been developing profiling tools for long-term unemployment [14]. Such profiling tools vary from country to country (e.g., whether the prediction is strictly data-based or partly data-based and includes case worker

evaluations). Common to these tools for assessing the risk of long-term unemployment is the underlying governmental desire for efficiency and cost-cutting [1, 14].

Recent studies of AI and algorithm use in the public sector question whether it merely represents a continuation of e-Government, or poses fundamentally different challenges to public services and administration [38, 40]. Veale and Brass [38] argue for an awareness of the different levels of governance on which such technologies are implemented. They argue it is important to be aware of because the level of decision-making (the macro, the meso, or the street-level), impacts how the skills, capacities, processes, and practices of public sector agencies will work, and this further has consequences for how AI and algorithms are implemented [38].

The most recent advances in AI and algorithms for public services are designed to perform specific tasks [25]. However, the assumption that caseworkers will eventually be replaced by AI prevails amongst stakeholders in the public sector [27]. Møller et al. [16] note to better understand the underlying assumptions of AI in a public sector context requires access to case materials and dialogue amongst relevant stakeholders. This is especially necessary if the goal is to make AI and algorithms available for public debate, policy change or the ability to opt out [16].

Together these studies emphasize how the increasing use of AI and algorithms impacts public sector processes across stakeholders, organizational boundaries, and levels of governance. Our case, the STAR algorithm, was developed on the macro national level, and it is not yet clear how or whether street-level caseworkers use it. To better understand the STAR algorithm, its predictions, and its implications, we turn to auditing.

2.1 Auditing as an approach to meet the need for explaining AI and algorithms

The need to explain and assess algorithms to a greater extent has induced new approaches to conduct algorithmic auditing. The emerging debate addresses both ethical and methodological concerns related to audits [28]. Drawing on insights from traditional audit studies, recent studies consider pros and cons related to internal and external audits when investigating algorithms [29]. For example, Raji et al. [28] point out that external audits, in contrast to internal ones, are a useful way to avoid influence from organizational considerations and internal interests. However, whereas external audits often need to base their work on model outputs, internal audits tend to provide direct access to systems, intermediate models, or training data which enables different kinds of insights. To harness the benefits of both internal and external algorithm audits, Wilson et al. [41] present "the cooperative audit". This form of algorithmic auditing constitutes "a framework for external algorithm auditors to audit the systems of willing private companies". Conducting a cooperative audit involves three steps: First, the auditors should clarify what they are and are not examining. Second, the auditors need to establish "the baseline requirement for conducting the audit". For Wilson et al. [41] this baseline includes considering transparency, renumeration, access, materials, and possibilities for independent testing. Finally, the auditors should clarify how they manage the relationship with the company. This means that rather than positioning an external audit as an adversarial activity conducted against company wishes, a cooperative audit provides a means to conduct independent assessments that could potentially result in constructive changes. Thus, the main distinction in a cooperative audit from the kinds of audits described by Sandvig and colleagues [29] is that the organization using the algorithmic system is aware of the audit and participates in a dialogue with the auditors.

3 A COOPERATIVE AUDIT OF THE STAR ALGORITHM

Our investigation of the STAR algorithm resembles a cooperative audit [41]. Our initial interest was to show how the STAR algorithm produces predictions about whether a person is at risk of long-term unemployment. Moreover, our own prior research of job placement had demonstrated the potential harm of even a simple algorithm deployed in social services [2]. We therefore focused on looking for evidence of direct discrimination based on the original complaint filed by the Danish Human Rights Institute against the agency [9]. The complaint revolved around the agency's use of the variable, origin, to predict risk of long-term unemployment. This discovery originated from an investigation by the Danish media [3, 10, 31]. Based on these earlier insights, we defined our starting question as: does the training model source code use demographic data directly as input for the risk prediction of long-term unemployment? To answer this question, we looked at the algorithm documentation and the model logic evident within it, as well as the model tuning code originally provided by the agency. We did not evaluate how and why particular values and thresholds were selected because we did not have access to the original logic of the design.

Members of the research team have been collaborating with the agency since 2019 [2]. In this case, however, we gained access to the materials by collaborating with the Danish news media company 'Zetland'. Our collaboration with the media came about because the second author had collaborated with Zetland in the context of another study of a different public sector algorithm [18, 30]. The research team and media company exchanged insights and acquired the code and documentation through the media company's subject access. The code and documentation are owned by the agency, which is the governmental agency responsible for implementing and assessing employment policy in Denmark. Access to the records was not restricted by trade secrecy protections and laws, as it may be in the context of private companies [41] or in cases where the public service algorithms are "licensed products" provided by private developers. In contrast to Wilson et al. [41] we did not establish a 'baseline' with the agency to gain access and initiate our audit activities. However, we made the agency aware of our audit activities that took place between January and April 2021. We also maintained an open line of communication with the agency after the audit activities. The ongoing dialogue allowed us to ask for necessary clarifications as we worked on the audit as well as during the review process of this paper. The agency was genuinely interested in our analysis and wanted to find ways to understand and improve the utility of the STAR algorithm. Following Wilson et al. [41], we ensured that all of our activities were as transparent as possible – the audit team had the full rights to make public all aspects of our investigation. We remained independent from the agency and our audit was part of our collaboration with Zetland and their data scientists' investigation into the STAR algorithm [23].

After gaining access to the materials, the second author "translated" the available source code from the original SAS format to Python. Then, we manually examined the source code and identified the algorithm as a decision tree. The available source code included model-tuning, showing the model was trained on a dataset consisting of more than 90 initial variables and 152,000 observations (data from unemployed individuals). We reviewed the available source code analysis, which showed that 84 variables had been excluded in the final algorithmic model, including gender and personal relationships. The implemented STAR algorithm includes six actionable variables: (1) the individual job seeker's self-identified possibility for employment, (2)

origin, (3) age, (4) employment rate, (5) income level the past year, and (6) educational background. The documentation describes the mathematical foundation for selecting these variables using an algorithm that identifies which of the variables best correlates with the target variable: long-term unemployment. However, the described variable selection methodology relies on statistical correlation, and the documentation presented no evidence of causality between the selected variables and long-term unemployment. The algorithm is essentially reduced to a small decision tree, meaning that the predictions produced by STAR are based on a small number of variables that gain significant influence on the produced outcomes. The STAR algorithm's reported accuracy is 69.9%, although it is not clear to us how this measure was obtained [33].

We began our code analysis by focusing on four concrete combinations of variables (paths through the decision-tree model) that, according the agency, lead to segregation of the risk group [36]. The combinations include just four of the six variables: the individual job seeker's self-identified possibility for employment, origin, age, and employment rate (see table 1). For each value combination we ran the code and compared the output. To document this part of the process, the second author created a .CSV file with combinations of values (the materials are available via Github[3]).

Table 1. Replication of the agency's four concrete combination of variables

Variable 1 *How long do you think it will take before you get a job?*	**Variable 2** *Origin*	**Variable 3** *Age*	**Variable 4** *Employment rate the past 36 months*	**Risk** *(%)*
Combination 1 - More than 6 months - I will go on maternity leave soon - I will retire soon				83,1 %
Combination 2 - Within 6 months - I don't know	- Western immigrants, - Western descendants - non-Western descendants			67,7 %
Combination 3 - Within 6 months - I don't know	- Danish origin, - non-Western immigrant - Unknown origin	< 56 years		64,3 %
Combination 4 - Within 6 months - I don't know	- Danish origin, - non-Western immigrant - Unknown origin	> 56 years, - Unknown age	> 0.08	65,1 %

The initial analysis emphasized the algorithm's use of the data category "origin, which includes the values "Danish origin", "Western immigrant", "Non-Western immigrant", "Western descendant", "Non-Western descendant", and "Unknown origin". Through our ongoing dialogue with the agency, we clarified that this terminology relies on categories produced and used by the national agency for statistics (Statistics Denmark). According to this terminology, "Danish origin" refers to a person who has at least one parent who is both a Danish citizen and who was born in Denmark. The categories "immigrants" or "descendants" refer to people who do not have a parent

[3] https://github.com/theresemoreau/star

who is both a Danish citizen and born in Denmark. These categories differ in that "immigrants" were born abroad, while descendants were born in Denmark [17]. The distinction between whether a category is characterized as "Western" or "Non-Western" has been used since 2002 and groups EU countries, EEA countries, Switzerland, Australia, New Zealand, the USA, and Canada as "Western countries". All other countries are grouped as "Non-Western countries". There is no documentation for how the definitions of Western and Non-Western countries were developed. The requirements for the definition were that it should consist of two or at most three groups, as it would otherwise become too complex [32].

In the process of going through the combinations, the second author used the documentation to better understand the design of the algorithm. She noticed a difference in the cut-off value disclosed in the source code (0.5) and the cut-off value in the documentation (0.6). For diagnostic purposes, the cut-off values are defined as the dividing points between two or more categories that fall on a continuous scale. Here, the scale is the probability space, which is bounded between and including 0 and 1, i.e., $p \in [0, 1]$. As such, the cut-off value of 0.6 tunes the relation between the algorithm's precision and its contribution ratio. In practice, this means that a data observation needs a score of 0.6 or higher to be classified as high risk for long-term unemployment ($p >= 0.6$). In contrast, a data observation that scores less than 0.6 is classified as low risk for long-term unemployment ($p < 0.6$). The agency had, according to the documentation, "assessed that "the cost" of a false positive classification is significantly higher than a false negative" [33]. We therefore decided to systematically compare the source code and the documentation both to validate the difference and to identify any other differences between the code and the documentation. Our comparison did not show additional discrepancies.

4 AUDIT FINDINGS

Working through the variable combinations (Table 1) allowed us to understand where and how the actionable variables were activated in the model and what substantive difference was caused by the change in the cut-off value. The algorithm heavily relies on the individual's self-evaluation of their probability for remaining unemployed to predict long-term unemployment risk (see figure 1). Self-evaluations can be surprisingly robust and accurate in many areas of life [19, 37] and believing that one can get a job can be an important component of determining whether an individual is willing to actively engage in job seeking activities, such as attending workshops and writing job applications. Thus, the STAR algorithm primarily pertains to people who already believe that they will have trouble getting a job in a reasonable timeframe. For those that indicated they expect to struggle finding a job, the algorithm relied on other variables to produce a prediction for likelihood of long-term unemployment.

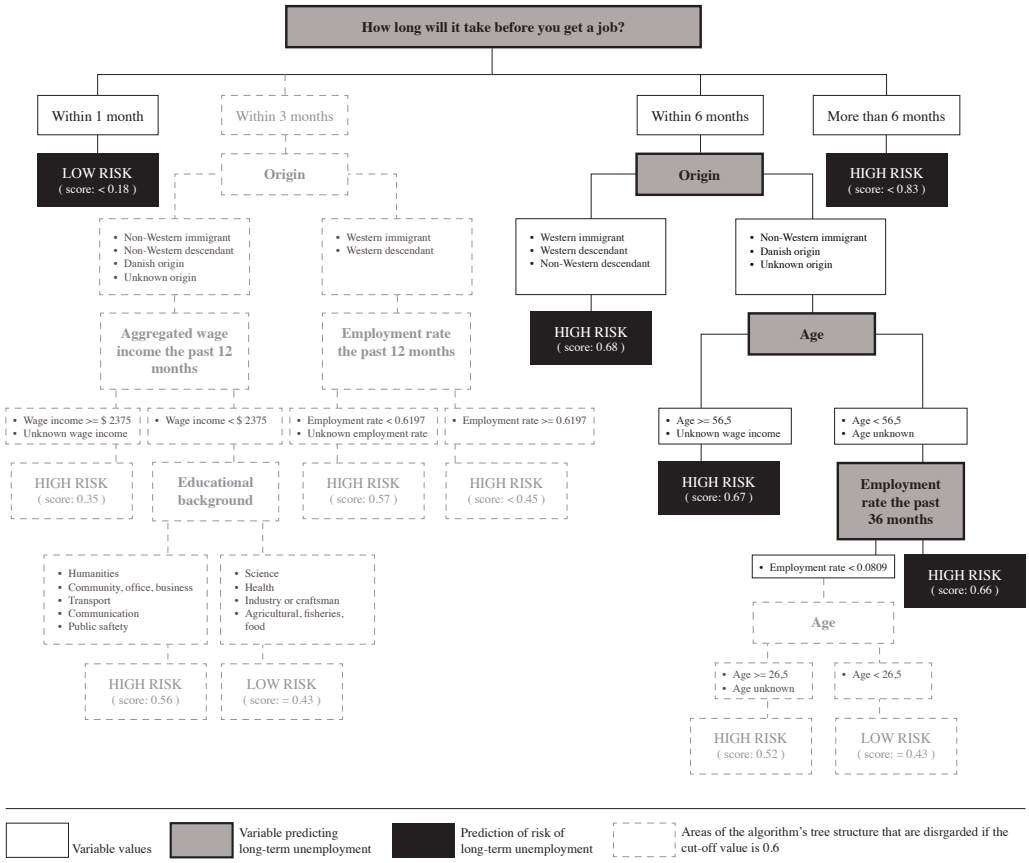

Figure 1. Visualization of the STAR algorithm's tree structure based on the source code. The grey dotted area of the tree structure does not influence prediction outcomes if the cut-off is 0.6.

4.1 Consistent Prediction of Risk of Long-term Unemployment

Under both cut-off points (0.5 and 0.6), jobseekers with any of the following answers will be determined to be at high risk of long-term unemployment with a risk score of 0.83, equivalent to ~83%:

- "It will be more than 6 months"
- "I expect to retire soon"
- "I expect to go on maternity leave soon"

In situations where the individual jobseeker's self-identified possibility for employment is presumed to be between 3-6 months ("Within six months") or they cannot assess when they will get a job ("I do not know"), the jobseeker moves down the decision three to the white box on the right side of figure 1 representing the variable origin. If the individual is a Western immigrant, a Western descendant, or a Non-Western descendant, they will be categorized as at risk of long-term unemployment. The algorithm categorizes differently if the unemployed is of Danish origin, a Non-Western immigrant, or if their origin is unknown. In this case, the jobseeker moves further down the decision three. The algorithm now determines risk of long-term unemployment based

of the person's age. According to the structure of the STAR algorithm, a person 56 years older or older is at risk of long-term unemployment. If the person is <56 years old, the prediction is determined based on their employment rate during the past 36 months. At this point we see a divergence in the decision tree between the cut-off points 0.5 and 0.6. At 0.5, if the individual's employment rate is less than 0.0809 during the past 36 months, the risk of long-term unemployment is yet again based on age; if the person is not yet 26.5 years old, they are categorized at being at risk of long-term unemployment. At 0.6, the final age-check is skipped.

4.2 Problematic Predictions of Risk of Long-term Unemployment

If the cut-off is at 0.6, the following answers will lead to a prediction of low risk of long-term unemployment.

- "I have a new job but haven't started yet"
- "Within one month"
- "Within three months"

If the cut-off is at 0.5 however, only the top two answers "I have a new job but haven't started yet" and "Within one month" will lead to a prediction of low risk with a risk score of 0.18, equivalent to ~18%. If the jobseeker indicates that they will obtain a new job "within three months", the algorithm moves down the decision tree and weights the person's origin to predict risk of long-term unemployment. If the person is a Western immigrant or a Western descendant, the algorithm takes the employment rate into account to predict risk of long-term unemployment. This differs for people who are categorized as Non-Western immigrants, non-Western descendants, of Danish origin, or if the origin is unknown. In these cases, the algorithm predicts risk of long-term unemployment based on the individual's aggregated wage income within the past 12 months. If the aggregated wage income is equal to or below 14.449 DKK (~ $ 2,375) the algorithm moves further down the decision tree, checking educational background. Here those who indicate an educational background in, for example, humanities, are predicted to be at risk of long-term unemployment.

The most obvious difference in the use of the higher cut-off value of 0.6 is that this eliminates the majority of the 'left side' of the decision-tree and the problematic predictions for the "Within three months" answer to the top-level question. It also becomes clear that the variable origin plays an important role in the decision-tree for both cut-off points, but the interpretation of its implications changes depending on which cut-off point is examined. The emphasis on origin as the important determinant raises concerns about potentially biased outcomes, which may be in violation of the law prohibiting discrimination in the labor market [9].

4.3 Cooperative auditing – agency interaction

Having conducted our analysis, we contacted the agency to validate whether there was indeed a difference between the SAS source code we received (which had implemented a cut-off value of 0.5), the documentation (which stated the STAR algorithm implemented a cut-off value of 0.6) and the actual implementation of the algorithm as part of the agency systems and web application. The agency confirmed that the current source code was implemented with 0.6 as the cut-off value. Through our ongoing dialogue with the agency, they explained the seeming difference in the cut-off value disclosed in the source code and documentation as follows:

"I believe the reason for the confusion is that we cannot implement SAS source code directly on jobnet.dk [web application], which is now an old lady. Instead, the paths are programmed in SQL. As for the SAS source code and the calibration of the model you refer to: After the model was calibrated, we introduced an exogenous intersection point of 0.6. It is the remaining paths that are implemented on jobnet.dk.". (E-mail from contact at the agency. May 2021).

This correspondence demonstrates the challenges that arise when public sector legacy systems (web application) create implementation challenges. In this case, it means that additional SQL paths were necessary for the algorithmic model to be implemented with the correct cut-off value. Moreover, this example of our ongoing dialogue with the agency also demonstrates the benefit of our cooperative audit approach. We would not have been able to make sense of the identified discrepancy and its effects in practice on our own. The cooperative approach enabled us as auditors to "connect" the provided source code, the documentation, and the practical implementation of the STAR algorithm. The simplicity of the algorithm may have made it easier for us to do so. We suspect more complicated and more far-reaching algorithms may be much harder to audit collaboratively. Our audit also emphasized the need for further discussion to trace the reasoning behind the decision-making that underpins the algorithmic model. Specifically, this manifested as the need to better understand the introduction of the exogenous intersection point. Finally, our cooperative approach emphasized that the documentation provided to the media is not the whole story and in fact may unintentionally mislead external stakeholders' interpretation of the case at hand.

5 DISCUSSION

Algorithms can have complicated stories. The STAR algorithm did not start out as a profiling tool, but was initially developed as a tool to help the government to statistically predict overall potential numbers of unemployed [35]. As the political situation of the Danish unemployment context changed, the purpose of the algorithm changed as well. Over the five years since it was launched, the STAR algorithm has evolved from an overall unemployment prediction tool to a tool to support the dialogue between the caseworker and the unemployed person by providing pre-assessment. What started out as an advanced statistical tool was eventually turned into an algorithmic profiling tool with apparently uncertain outcomes.

As with other types of software development, algorithmic systems are distributed systems that emerge out of the sum of decisions often made over long periods of time and with a rotating cast of characters including software developers, project managers, politicians, end users, and others [12]. Especially in a public sector context, decision-makers ought to recognize the possibility that the original source code, its attendant documentation, and its potential journeys through re-implementation can diverge as algorithmic systems come to be integrated with existing legacy systems, as is the case with the web-application Jobnet.

An algorithmic audit like the one we conducted cannot provide in-depth explanations for why the algorithmic structure is organized the way it is, but it does provide insight into how it works. The differences in the cut-off value between the source code and the documentation made available to the Danish news media resulted in the reduction of the active variables used for predicting long-term unemployment. Of the six original variables used for the algorithm,

educational background and aggregate wage income over the prior 12 months are the two that the individual most likely has the agency to alter.

Individuals of course, cannot alter their place of origin and although all of us change in age, it is predictably in the same inexorable direction. Thus, the current implementation of the algorithm will not alter its designation of potential for long-term unemployment no matter how much additional education an individual might attain. This does not bode well for the ideal of life-long learning.

Despite the public controversy around the profiling produced by the STAR algorithm, its real-world impact has limited influence on the individual job seeker. Current research with unemployment caseworkers in Denmark has demonstrated that the binary designation of whether or not someone may be at risk of long-term unemployment has limited effect on the individual's interaction with the job center and thus their job search process [2, 16, 27]. Yet this does not reduce the relevance of our audit, because the STAR algorithm has been referred to as an example of algorithmic over-reach and potential discrimination by politicians and civil society representatives. Therefore, it is important to understand the inner workings of the algorithm if we want to discuss what constitutes algorithmic over-reach and how discrimination may be manifested in the algorithmic model in this case.

In the STAR case, the algorithm is simple, and its efficacy was questioned by the agency in our dialogue. Other algorithms currently coming online in Denmark and other places in Europe that promise efficiency gains in government operations are more complex and will pose greater challenges for auditing efforts. Setting an example of cooperative auditing of STAR is one step towards convincing other public agencies to be open to these kinds of assessments. However, it has not been without its challenges. In contrast to Wilson et al. [41], who collaborated with one corporate stakeholder, we engaged with stakeholders in the public sector context with diverging assumptions of, for example, what counts as a thorough and useful explanation for how variables were selected for the STAR algorithm. This poses new challenges for us as researchers in terms of how to navigate varying (sometimes downright opposite) interests and timelines. In this case, we experienced how the news media works with significantly shorter publication cycles, and the agency was oriented towards the public debate. As a result, the slow pace of our research process became a challenge, as we asked the involved stakeholders to comment on the findings of the paper at times where they were either on to the next story or were in the process of a legal trial. In this case, the public debate to some extent formed the context of our audit through which stakeholders interpreted the STAR algorithm.

Tackling such challenges and the increased level of complexity it entails to cooperate with core stakeholders of the public debate, we join Irani and others in calling CSCW and HCI researchers to *act* together with media and other institutions central to democratic society [11] to further develop approaches to audits of public service algorithms. Similar to other contexts, researchers and media in Denmark are often prevented from investigating algorithmic systems due to trade secrecy protections and laws. Public services typically do not "own" the code and documentation underlying the algorithmic systems; rather, they purchase a license. In this study we relied on the Danish news media for accessing the code and documentation. We were also relying on the agency, which answered our questions and readily engaged with us. This kind of open

engagement and willingness to take seriously outcomes of a public audit should be a form of best practice for organizations that use algorithmic systems in the provision of public services.

6 CONCLUSION

In this note we set out to audit a highly contested public service algorithm for profiling risk of long-term unemployment. The audit was a result of a collaboration between the research team and the Danish media company, Zetland, that had initially acquired the source code and the documentation. We also worked together with the Danish Agency for Labour Market and Recruitment, whose algorithm we were auditing to resolve questions about important implementation discrepancies. Algorithmic audits are often perceived as adversarial actions against the organizations that utilize the algorithms in question. However, this does not necessarily need to be the case. Audits are a critical tool for finding ways to ensure that algorithmic systems are implemented in less harmful ways. Our study demonstrates the clear need to ensure that civil society and the public can conduct algorithmic audits, especially for systems that public organizations use to operate. We argue that cooperative audits conducted by independent actors and in open conversation with the public organizations in question can be more productive and can help us ensure that increasingly digital and algorithmic societies can remain open and democratic.

ACKNOWLEDGMENTS

We thank our collaborators from the Danish Agency for Labour Market and Recruitment, especially Carsten Søren Nielsen, and Zetland's Frederik Kulager besides Peter Maarbjerg Dønvang – as well as colleagues Asbjørn Ammitzbøll Flügge, Trine Rask Nielsen, and Thomas T. Hildebrandt for providing feedback. This research has been supported by the Innovation Fund Denmark (EcoKnow: award number 7050- 00034A) and the Independent Research Fund Denmark (PACTA: award number 8091-00025b).

REFERENCES

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

[2] Ammitzbøll Flügge, A., Hildebrandt, T. and Møller, N.H. 2021. Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective. *Proceedings of the ACM on Human-Computer Interaction*. 5, CSCW1 (Apr. 2021), 40:1-40:23. DOI:https://doi.org/10.1145/3449114.

[3] Andersen, T. 2019. Jurist: Dataprofilering af langtidsledige med etnicitet er ulovlig. *Version2*.

[4] Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*. 5, CSCW1 (Apr. 2021), 1–34. DOI:https://doi.org/10.1145/3449148.

[5] Benjamin, R. 2019. *Race after technology: abolitionist tools for the new Jim code*. Polity.

[6] Bright, J., Ganesh, B., Seidelin, C. and Vogl, T.M. 2019. *Data Science for Local Government*. Technical Report #ID 3370217. Social Science Research Network.

[7] Brown, S., Davidovic, J. and Hasan, A. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*. 8, 1 (Jan. 2021), 2053951720983865. DOI:https://doi.org/10.1177/2053951720983865.

[8] Crawford, K. 2021. *Atlas of Ai: power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[9] Danish Human Rights Institute 2020. Klage til Ligebehandlingsnævnet.

[10] Danske A-kasser 2019. Hvis sagsbehandleren stoler blindt på algoritmerne, risikerer han at stemple de forkerte | Danske A-kasser.

[11] De La Garza, A. 2019. Meet the Researchers Fighting to Make Sure Artificial Intelligence Is a Force for Good. *Time*.

[12] Dourish, P. 2017. *The stuff of bits: an essay on the materialities of information*. The MIT Press.

[13] Ferguson, A.G. 2017. Policing Predictive Policing. *Washington University Law Review*. 94, (2017), 82.

[14] Højbjerg Jacobsen, R. and Bendix Kleif, H. 2016. *Effekten af indsatser for langtidsledige og ledige i risiko for langtidsledighed: litteraturreview*. KORA.
[15] Holten Møller, N., Nielsen, T.R. and Le Dantec, C.A. 2021. Work of the Unemployed. An inquiry into individuals' experience of data usage in public services and possibilities for their agency. (2021).
[16] Holten Møller, N., Shklovski, I. and Hildebrandt, T.T. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn Estonia, Oct. 2020), 1–12.
[17] Immigrants and Descendants: *https://www.dst.dk/da/Statistik/emner/befolkning-og-valg/indvandrere-og-efterkommere/indvandrere-og-efterkommere*. Accessed: 2021-08-09.
[18] Kulager, F. 2021. Kan algoritmer se ind i et barns fremtid? I Hjørring og Silkeborg eksperimenterede man på udsatte børn. *Zetland*.
[19] Kum, H.-C., Duncan, D.F. and Stewart, C.J. 2009. Supporting self-evaluation in local government via Knowledge Discovery and Data mining. *Government Information Quarterly*. 26, 2 (Apr. 2009), 295–304. DOI:https://doi.org/10.1016/j.giq.2008.12.009.
[20] Marie, H. Blinde vinkler.
[21] Meijer, A. and Wessels, M. 2019. Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*. 42, 12 (2019), 1031–1039. DOI:https://doi.org/10.1080/01900692.2019.1575664.
[22] Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data*. 21.
[23] Moreau, T. and Kulager, F. 2021. Vi har skilt jobcentrenes algoritme ad. Og kan se, hvor anderledes indvandrere og efterkommere bliver behandlet. *Zetland*.
[24] OECD Glossary of Statistical Terms - Long-term unemployment Definition: 2002. *https://stats.oecd.org/glossary/detail.asp?ID=3586*. Accessed: 2021-08-06.
[25] Pasquale, F. 2020. *New laws of robotics: defending human expertise in the age of AI*. The Belknap Press of Harvard University Press.
[26] Pasquale, F. 2015. *The Black box society: the secret algorithms that control money and information*. Harvard University Press.
[27] Petersen, A.C.M., Christensen, L.R., Harper, R. and Hildebrandt, T. 2021. "We Would Never Write That Down": Classifications of Unemployed and Data Challenges for AI. *Proceedings of the ACM on Human-Computer Interaction*. 5, CSCW1 (Apr. 2021), 102:1-102:26. DOI:https://doi.org/10.1145/3449176.
[28] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *arXiv:2001.00973 [cs]*. (Jan. 2020).
[29] Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. (Seattle, WA, USA, 2014), 4349–4357.
[30] Siboni Lund, C. 2019. *Algoritmer i socialfaglige vurderinger - en undersøgelse af socialarbejderes opfattelse af at anvende algoritmer til vurdering af underretninger*. Technical Report #39. The Danish Social Workers' Association.
[31] Sorgenfri Kjær, J. and Danielsen, L. 2019. »Det er jo værre end sagen om Gladsaxe-modellen, det her«: Algoritmer skal udpege langtidsledige. *Politiken*.
[32] Statistics Denmark 2018. Bilag 4 - Notat om ny landegruppering.
[33] The Danish Agency for Labour Market and Recruitment 2019. 3.3 Modeldokumentation.
[34] The Danish Agency for Labour Market and Recruitment 2016. *Effekten af indsatser for langtidsledige og ledige i risiko for langtidsledighed: litteraturreview*. KORA.
[35] The Danish Agency for Labour Market and Recruitment 2019. Notat - Kronologisk oversigt over udviklingen af det digitale afklarings-og dialogværktøj på Jobnet.
[36] The Danish Agency for Labour Market and Recruitment, C.S. 2020. Beskrivelse af profilafklaringsværktøjet til dagpengemodtagere.
[37] Usher, C.L. 1995. Improving Evaluability Through Self-Evaluation. *American Journal of Evaluation*. 16, 1 (1995), 59–68.
[38] Veale, M. and Brass, I. 2019. *Administration by Algorithm? Public Management Meets Public Sector Machine Learning*. Technical Report #ID 3375391. Social Science Research Network.
[39] Vogl, T., Seidelin, C., Ganesh, B. and Bright, J. 2019. Algorithmic Bureaucracy. *20th Annual International Conference on Digital Government Research on - dg.o 2019* (Dubai, United Arab Emirates, 2019), 148–153.
[40] Vogl, T.M., Seidelin, C., Ganesh, B. and Bright, J. 2020. Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. *Public Administration Review*. 80, 6 (2020), 946–961. DOI:https://doi.org/10.1111/puar.13286.
[41] Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K. and Polli, F. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada, 2021), 666–677.

Received July 2021; revised September 2021; accepted October 2021.