IT UNIVERSITY OF COPENHAGEN

Algorithmic Fairness in Automatization of the Public Sector:

A Literature Review on Automating High-Level Discretion and Enhancing Fairness

Johanne Engel Aaen (jaae@itu.dk)

Katrine Boberg Knudsen (kknu@itu.dk)

Simone Etwil-Meyland (siet@itu.dk)

Research Project, Autumn 2021

STADS code: KIREPRO1PE

Software Design

Supervisors: Christian Østergaard Madsen & Leon Derczynski

Character Count: 55788

Table of Content

Abstract

Automated decision making (ADM) is quickly finding its way into public sectors around the world, where machines replace caseworkers. A lot of these caseworkers deal with cases containing a high level of discretion and they cooperate with their colleagues in making decisions. In automating decisions, a lot of ethical questions arise in relation to the algorithmic fairness (AF) of these models and how fairness can be enhanced.

We here present a literature review investigating the intersections between the two concepts discretion and AF. The investigation is divided into three sections. One explaining and identifying our literature. One analyzing the literature to define the terms discretion and AF and lastly a section discussing the overlaps and differences between discretion and AF in order to identify knowledge gaps for future research.

The literature review finds two perceptions of discretion - individual and collaborative. And investigates AF in the pre-process methods - data gathering and problem formulation from two overall approaches - relationality and rationality. There is found to be several overlaps between the two concepts in both processual and ethical concerns.

1. Introduction

For years Denmark has had a leading role in digitizing the public sector and is today one of the most digitized countries in the world (Digitaliseringsstyrelsen, 2020). This is due to digitization being a political priority, as well as an increase in the use of artificial intelligence (AI) and other decision-making models. Through political initiatives such as *Digitizable Legislation* (Digitaliseringsstyrelsen, n.d.) and *Municipal Signature Projects on Artificial Intelligence* (Olsen, n.d.) the government keep expanding digitization of public activities to new frontiers where cases within all fields of public government are being managed digitally. Especially the ambition to automate public services in municipalities and regions can affect citizens directly when casework and decision-making is transferred fully or partially from human caseworkers and public servants to algorithmic models.

This literature review investigates digitization of public services that affect citizens on an individual level, and which involve a high level of human discretion. Through the lens of state-of-the-art literature we will explore *how discretion in the public sector is practiced and perceived*. Furthermore, we will analyze *how recent literature on automation addresses algorithmic fairness and how fairness can be enhanced through different approaches, processes and considerations.* This is in order to locate *what the challenges are in automated decision-making in the public sector according to the literature.*

Based on our findings we will identify some of the knowledge gaps that future research on digitalization of casework in local and regional governments could benefit from further investigating. We will condense what other researchers have already concluded regarding

automation of discretion and AF in the public sector and discuss the most relevant trade-offs and challenges involved in this process. Furthermore, we will pinpoint what is not yet known or covered by the literature. The literature review is intended to serve as a basis for future research within this field, by identifying current knowledge gaps that exist in the intersection between discretion and algorithmic fairness.

The literature review is structured as follows: Section 2 is an account of the methods applied in the review. Section 3 is an analysis of our two main concepts, discretion and algorithmic fairness. Considering this analysis, Section 4 discusses some of the most important overlaps in the intersection between AF & discretion as well as identifying the discovered knowledge gaps that future research could focus on. Lastly, Section 5 concludes on the findings of the literature review.

2. Methods

The following section describes how we conducted our literature search to identify and select articles that constitute the literature analyzed in this literature review.

2.1 Searching for Literature

We conducted our primary search for relevant literature in September and October 2021. We began by establishing and identifying mutual search criteria's and implemented a 3-step iterative process structure as proposed by Webster and Watson (2002, 16).

The suggested process consists of the following steps:

- Search for relevant articles in scientific databases
- An iterative backward search
- A forward search based on results from previous steps

We chose this approach to ensure consistency and to collect a pool of papers that cover the key concepts. The following elaborates on the process of determining our final pool of papers.

2.2 Process

To establish our search criteria's we used the approach of a *search profile* as described in "The Good Paper" by Rienecker and Jørgensen (2020), for which we identified relevant search words to focus the scope of our search for literature (151).

First, we identified the core concepts (152): AF, discretion and Public AI. These words were chosen after investigating our research area and concluding that these words cover the overall concepts that we wanted to investigate. Within each core concept we furthermore identified sub-concepts to specify the scope of our search criteria even more (154). This enabled us to narrow down the

search criteria's as our core concepts alone would give a very large search result. Table 1 shows our final search profile.

Table 1: Search Profile

Core concepts	***Algorithmic Fairness***	***Discretion***	***Public AI***
Sub concepts	Algorithmic bias	Street-level Bureaucracy	Automated Decision Making
	Algorithmic Transparency	System-level Bureaucracy	Public AI
	Trustworthy Algorithms	Screen-level bureaucracy	Machine Learning in public services
	Algorithmic Fairness in AI	Algorithmic discretion	

With the search profile in place, we started searching for relevant articles in 5 databases: *Association for Computing Machinery (ACM), Digital Government Reference Library, ResearchGate, Google Scholar and ACM Conference on Fairness, Accountability*, and T*ransparency* (ACM FAccT). We chose ACM, ACM FAccT and Digital Government Reference Library because they provide access to case-specific peer reviewed research that is considered highly relevant. Additionally, we investigated Google Scholar and ResearchGate to ensure a wider search to capture all potentially relevant articles. Prior to the process we established the following criteria that all papers should meet to be deemed relevant:

- The papers should be peer reviewed or moderated prior to publishing
- They should be written in a time span from 2000-2021[1]
- The articles should be written in English[2]

We used these criteria on the research databases and selected papers based on keywords as well as abstracts. This initial search yielded 17 relevant articles. Finally, we did a backward and forward search as recommended by Webster and Watson (2002, 4). Our backward search consisted of investigating references in the articles, while the forward search was done by locating articles referencing these same articles. For this purpose we used Web of Science and Google Scholar. The efforts of both backward and forward search resulted in 13 articles making the total number of articles in the initial pool of papers 30.

[1] Literature used for theoretical framing was exempted from this criterion

[2] Literature in Danish regarding implementation of ADM was exempted from this criterion

After finishing our search process we proceeded by reading all articles in the initial pool of papers. Each article was assigned to at least two group members to strengthen the reasoning and argumentation for choosing articles for the final pool of papers. Furthermore, we did short resumés to ensure that all group members had access to the main points, which enabled us to discuss the relevance of each article.

Having read the articles we proceeded to classify them. We chose to follow a concept-centric approach from Webster and Watson (2002) to get an overview of the relevant themes discussed in the texts (16-17). Our concept matrix (appendix 1) serves as the foundation for synthesizing our findings across the papers instead of analyzing the conclusions and knowledge gaps for each paper individually. As we read the articles we filled in the concept matrix to get an overview of the important concepts. First, we identified our main concepts (AF, discretion and public AI) and during our reading process we noted relevant sub-concepts.

To determine our final pool of papers we discussed the themes and conclusions within each article. 17 articles were discarded since they did not meet the scope of our research questions. This was due to different factors - some articles on AI and AF were simply outdated due to the wide time-range we had defined in our search process. Searching for articles published after 2000 made sense for articles on discretion, but we came to realize that many scientific articles on AI are not relevant to our scope if they do not reflect state-of-the-art within the field. Thus, we chose to prioritize newer articles. Other articles were irrelevant due to their main focus on policy and public finances since several of the articles we found investigated the policy-making around discretion and AI or the financial consequences it causes. This is out of scope in this literature review.

On the other hand, we chose to keep articles that generally discussed discretion and the practice of discretion and articles investigating methods and principles within AF and state-of-the-art approaches in this field, along with papers that discuss modern use of AI in the public sector. Our final number of articles is 12.

2.3 Displays

In order to describe and define the nuances of the research on discretion and AF we use Peter Dahler-Larsen's (2010) display of dimensions as a tool for defining concepts (206). This tool provides us with the ability to describe different dimensions within complex concepts and display a concrete overview in investigating what discretion and AF is (192).

Displays are used as an overview to clarify the most important aspects of a qualitative analysis and provide both the author and reader with a clarification of the most important aspects of the analysis by making a logical and systematic display of the findings (195). According to Dahler-Larsen a good display is based on three rules: the authenticity rule, the inclusion rule and the transparency rule. The authenticity rule is used to address the fact that duality exists within humans and the authors ability to challenge their pre-understanding of a subject if their analysis

finds something that challenges their belief (196). The inclusion rule implies that all data must appear in an analysis and that no data can be neglected if it does not fit within the rest. This rule is important to ensure that deviations and anomalies are detected and that the analysis covers all the material to avoid bias. The transparency rule implies that it should be clear how the display has been made by referring to how the data has been collected (198). In this paper the authenticity rule is applied by using quotes in the same context as they were intended. The inclusion rule is applied by discussing different views detected on both discretion and AF and the transparency rule is applied by describing how we have chosen our articles.

In this paper the use of displays gives us the opportunity to provide the reader with an overview of the nuances that exist within discretion and AF. This paper contains two displays that are visually different, since we believe that AF needs to be further divided into two sub-concepts. The distinction between the two and the specific use is further elaborated on in section 3.1.1 and 3.2.1. Our displays are followed by an analysis of the concepts and therefore the displays are broadly used as an overview of concepts that are further described and discussed later.

3. Analysis

The following section investigates discretion and algorithmic fairness as concepts and tries to encapsulate them to the scope of this paper. This is done in order to discuss the intersections between the two and identified knowledge gaps in section 4.

3.1 What is Discretion?

In this section we will investigate the concept of discretion. First, we introduce a display defining the five dimensions we believe define discretion as a concept. These are *street-level bureaucracy, system-level bureaucracy, collaborative discretion, individual discretion and formal/informal discretion*. The display is followed by an elaboration of each dimension and an analysis of our findings.

3.1.1 Display of Discretion

Table 2 shows our display containing the dimensions we find relevant to define the term discretion. Each dimension is explained by representative quotes from the investigated papers which indicate perspectives and definitions on each dimension.

Table 2: Display of Discretion

Concept	Dimensions	Representative Quotes
DISCRETION	Street-Level Bureaucracy	*"The decisions of street-level bureaucrats, the routines they establish, and the devices they invent to cope with uncertainties and work pressures, effectively become the public policies they carry out. ..[public policy] is actually made in the crowded offices and daily encounters of street- level workers"* (Bovens & Zouridis, 2002, 175) (Michael Lipsky, 1980)
		"In 1980, Michael Lipsky codified the role of the individual decision maker, the "street-level" bureaucrat. Such bureaucrats, including teachers, doctors, case workers, police officers, fire fighters, garbage collectors, and information service workers, provided government services, benefits, and punishments directly to the public" (Bullock, 2019, 751)
	System-Level Bureaucracy	"*No longer were bureaucrats out on the ground or sitting behind screens inputting data; instead information communication technology tools themselves augmented and automated more and more tasks, in some cases even replacing the role of expert judgement by human bureaucrats"* (Bullock, 2019, 752)
		"Routine cases are handled without human interference. Expert systems have replaced professional workers...The process of issuing decisions is carried out virtually from beginning to end by computer systems" (Bovens & Zouridis, 2002, 180)
	Collaborative Discretion	*"The social workers described how these decisions are often made individually. However, sometimes, if they feel 'stuck', they may engage in discussions with team members... Furthermore, their decision on what to inspect and examine is influenced by conversations with other stakeholders rather than relying exclusively on the judgement of the individual social worker"* (Petersen, Christensen & Hildebrandt, 2020, 91)
	Individual Discretion	*"Bureaucrats are well known to be small-minded pencil pushers who can reject or approve an application for no better reason than the fact that your existence has somehow annoyed them"* (Bovens & Zouridis, 2002, 74)
		"*The context-based discretion that is practiced by human case workers increases the risk of arbitrariness, injustice and inscrutableness"* (Justesen & Plesner, 2018, 17)
	Formal and Informal Discretion	*"...formal discretion is allowed within rules and informal discretion exists outside the body of rules...formal discretion occurs when policies and procedures are broad and vague and among other things can generate misinterpretations. Informal discretion, on the other hand, is the result of inadequate evaluation mechanisms of rules" (Petersen, Christensen & Hildebrandt, 2020, 4)*

Street-Level Bureaucracy: The fundamental dimension of discretion, which was first defined by Michael Lipsky in 1980 as a term to cover all public employees who practice discretion (Lipsky, 1980, 3). According to Lipsky, everything from police officers and judges to social workers and teachers are street-level bureaucrats. Bureaucrats hold a discretionary power when making decisions in cases that are too complex to be solved based on narrow definitions found in the rules defined by law. Instead, they have mandate to interpret and modify laws to solve specific problems (72). Lipsky argues that law-making is not only conducted by politicians but also by street-level bureaucrats who interpret policies on a daily basis to cope with uncertainties and complexity in their work which is not directly covered by law (Bullock, 2019, 751).

System-Level Bureaucracy: Defined in 2002 by Bovens and Zouridis and a continuation of Lipsky's street-level bureaucracy. As a result of the increasing digitization of the public sector, they found street-level bureaucrats to have a new role. Instead of being on the street practicing discretion founded in individual reflections, they were being replaced with so-called expert systems in non-complex cases where the role of the bureaucrat is to improve the automation, rather than manually evaluate each case (180).

The shift from street-level bureaucracy to system-level bureaucracy has shifted the practice of decision making as well as that of discretion. It is no longer the sole responsibility of street-level bureaucrats, but is to a greater extent something the developer and system-designer practices during the development of a model (177). Discretion is not removed, but practiced by technology rather than caseworkers.

Collaborative Discretion: Discretion can be identified as a group activity which is carefully practiced by capable professionals who reason based on experience and expertise and who reach decisions in collaboration with their peers (Petersen, Christensen & Hildebrandt, 2020, 91). Therefore we define the term collaborative discretion. Petersen et al. (2020) refers to discretion as a community of practice and describes it as being an ongoing reasoning among co-workers and stakeholders (98).

Individual Discretion: In addition to identifying collaborative discretion, discretion can also be defined as a decision made subjectively by the caseworker - we define this as individual discretion. Justesen and Plesner (2018) finds that even though caseworkers are educated to base their decisions on this kind of discretion, it is often a source of injustice and arbitrariness (17). Another aspect of individual discretion is defined by Bovens and Zouridis (2002) who conclude that human caseworkers might be likely to deliberately act based on their own emotions and personal opinions, resulting in similar cases having different outcomes (174).

Formal and informal discretion: Within discretion we can distinguish between two types - formal and informal (Petersen et al., 2020, 4). Formal discretion is seen as the *intentional* discretion where laws are defined such that there is room for a caseworker to practice some degree of discretionary power within the rules of the law. Petersen et al. concludes that in cases containing

formal discretion the laws and rules define what social workers must do, but not necessarily how it should be done (14). On the other hand informal discretion is not intended, but is a result of acting in a context where rules or evaluation mechanisms are insufficient in order to provide guidelines for the caseworkers (4).

3.1.2 The Practice of Discretion in the Public Sector

The display of discretion paints a picture of a concept that can be broadly defined and contains different dimensions. We now analyze the state of discretion in the public sector by investigating how discretion is practiced and how a shift from street-level to system-level bureaucracy has affected the practice and views on discretion.

As described above, Lipsky (1980) defined street-level bureaucrats as public servants who interacted directly with individual citizens using discretion in their decision making in a given case. Throughout the articles we identified two main comprehensions of how discretion is practiced by caseworkers in the public sector. Roughly said, one group of papers represented by Alkhatib and Bernstein (2019), Bullock (2019), Petersen et al. (2020) and Petersen, Christensen, Harper and Hildebrandt (2021) consider discretion as being collaborative, where Bovens and Zouridis (2002) and Justesen and Plesner (2018)) identify that caseworkers practice individual discretion based on their own reflections and opinions.

Bovens and Zouridis (2002) describe discretion as a highly individual task. They consider the practice of discretion to be conducted based on personal opinions of a caseworker and describe bureaucrats as "*small-minded pencil pushers who can reject or approve an application for no better reason than the fact that your existence has somehow annoyed them*" (174). They state as a fact that the decision outcome is based on emotions and personal opinions. Justesen and Plesner (2018) argue that caseworkers practice discretion based on context and their own impressions which (can) create injustice and arbitrariness (17). Justesen and Plesner's research is an interview-study with managers in the public sector in Denmark, which reflect an overall opinion among the interviewed managers that decisions made by two different bureaucrats will likely be very different and random whereas machines conduct more consistent decisions (16). We derive from both articles that they share a concern of bias in the decisions made by street-level bureaucrats.

Petersen et al. (2020) however find that very little effort and empirical research has been done on how social caseworkers really exercise their use of discretion in practice (2). In the paper *Discretion in the Age of Automation* they conduct an ethnographic field study in a Danish municipality, investigating how discretion is handled in a child protection department. Contrary to the description of street-level bureaucrats who base their decisions on personal beliefs and at a whim, as seen in Bovens and Zouridis (2002) and Plesner and Justesen (2018), Pedersen et. al. (2020) find that the practice of discretion is a much more collaborative effort (12).

The collaborative discretion-approach to casework is also described by Petersen et al. (2021) in *We Would Never Write that Down: Classifications of the Unemployed and Data Challenges for AI*. This is an empirical field study in a Danish job center, where Petersen et al. argue that caseworkers do not only practice discretion in collaboration with their colleagues, they also need to make situated judgments (23). They observe that caseworkers are expected to fit unemployed citizens into narrow boxes of either *ready to work* or *not ready to work.* These categories are defined by law and are the only official options in the decision making, but they rarely match the real-world scenarios they meet among citizens. Therefore caseworkers make situated judgments by adding informal categories to case files. These informal categories are often based on sensory impressions and interactions with the citizens (14). However, situated judgments are rarely written into the formal case files, as they often have a harsh tone and are quite personal. The situated judgment is merely used as a tool in the collaborative discretion, so caseworkers can obtain a more detailed picture of a specific case, in contrast to the rigid formal definitions provided by law (18). As a result Petersen et al. (2021) conclude that automation without these informal categorizations might cause issues, since algorithms will not have access to the entire foundation that decisions are made on (23).

Adapting grey-areas of the real world to defined rules is also described by Alkhatib and Bernstein (2019) who find that street-level bureaucrats "fill the gaps" of laws to handle specific and unique cases, which we believe to be an example of situated judgment as well. They argue that filling gaps is not something the caseworkers do to work their way around legal frameworks. It is a practice used to refine their understanding of the policies and people they work with (530).

Another important concept in Alkhatib and Bernstein (2019) is reflexivity. They consider reflexivity as the discretionary part of decision making. Reflexivity is the ability to reflect and adapt to new and unknown situations. They argue that reflexivity is one of the reasons that there is more faith in decisions made by street-level bureaucrats than those made by street-level algorithms (the automated version of street-level bureaucrats). This is based on the fact that street-level algorithms only manage to reflect and adjust their findings after a, potentially wrong, decision has been made. In contrast, street-level bureaucrats are able to reflect upon long and short term consequences of different scenarios up until a decision has to be made (4).

This is in direct contrast to Bovens and Zouridis (2002), who express a strong faith in algorithms over humans, such that "*the personal preferences or biases of the official [street-level bureaucrat] handling the file can no longer play a role*" (181). However, they acknowledge that this moves part of the discretionary power to the designers and IT experts, who implement the algorithms by saying that there are *"...persons whose choices can affect the practical implementation of a policy*" and that not all types of cases can be conducted without human interaction (182).

A common understanding across the articles is that there exists scenarios where no or very little human discretion is needed - simply because the cases are simple and generic. We call this aspect

of discretion highway discretion[3], which is a definition borrowed from Justesen and Plesner (2018, 17). This understanding was already described by example in 2002 by Bovens and Zouridis who explain how speeding tickets are no longer issued by the police officers, but by a computer (179). Justesen and Plesner (2018) describe that simple cases without complexity can easily be digitized such that the public should not spend resources on caseworkers handling these cases (17). They argue that in cases that differ from the highway it should be possible to take the side road[4] having a caseworker to handle the case using their professional, human discretion.

It is interesting to highlight that several articles claim that the human caseworkers are easily replaced with technology in highway cases. This is supported by Bullock (2019), who claims that cases with low uncertainty and complexity are likely to be automated (757). Young, Bullock and Lecy (2019) supports the perception that highway cases can easily be automated to save money and to save human caseworkers from dull and repetitive tasks (304). However they also argue that there are several challenges in automating so-called high-level discretionary tasks, since they:

"...are defined by poor data, uncertainty about the factors that lead to success, or tightly coupled systems that are difficult to model. Decision-support tools and predictive analytics are inappropriate in these scenarios because the problems are not well-enough defined or there is not enough data to model them effectively" (304).

This view goes together with the conclusions of Petersen et al. (2021) who find that caseworkers do not document everything that underlies their decisions as discussed above.

After investigating the dimensions of discretion and dividing them into different perceptions of how discretion is practiced, we can conclude that discretion is a complex practice of decision making. Throughout the articles we see a common understanding of the fact that discretion can be practiced on many levels - some cases require a high level of human discretion, where others require very little. We have also identified that the understanding of how discretion is practiced varies a lot from article to article. We see that Bovens and Zouridis (2002) and Justesten and Plesner (2018) state that discretion is individual and they present human bias as a potential issue that needs to be handled and consider automation as a potential solution. On the other hand, the articles by Alkhatib and Bernstein (2019), Bullock (2019), Young et al. (2019) Petersen et al. (2020) and Petersen et al. (2021) believe discretion to be a collaborative task or a situated judgment and are concerned that the level of complexity will be difficult or impossible to automate.

Automation-sceptic or not, there is a general belief across the literature that automation of discretion needs to be fair. Some writers take this as a given more than others, which leads us to

[3] In Danish "*motorvejen*"
[4] In Danish "bivej"

our investigation of AF. In the following section, we will investigate different pre-processing methods and approaches to enhance fairness in automation of discretion.

3.2 What is Algorithmic Fairness?

In the following section we dive deeper into the concept of AF. However, a few comments are in place before we move on with our analysis. While reviewing the literature, we discovered that AF is not a concept that is easily encapsulated. The question of whether a model is fair or not is highly contextual and ensuring fairness depends on which definition of fairness that is applied. Therefore we have chosen to focus on two overall approaches: rationality and relationality, introduced by Birhane (2021). These approaches are intended to give a condensed overview of the potential scope of AF.

3.2.1 Display of Algorithmic Fairness

In describing AF we create a display containing two general dimensions - *pre-process methods* and *approach*. Unlike the display of discretion, we also identified several sub-dimensions to further elaborate on the concept. We focus our investigation on AF in the pre-process methods of a development process since we believe that these aspects are interesting to evaluate in relation to discretion. We acknowledge that there exist more sub-dimensions such as in-process, post-process and evaluation of AF, but these are outside the scope of this paper.

Table 3: Table of Algorithmic Fairness

Concept	Dimensions	Sub-dimensions	Representative Quotes
ALGORITMIC FARINESS	Pre-process Methods		"*The foundation of a predictive model is the data on which it is trained. The data available for any consequential prediction task, especially data measuring and categorizing people, can induce undesirable properties when used as a basis for decision-making*" (Mitchell, Potasch, Barocas, D'Amour & Lum, 2021, 144)
		Data Gathering	*"One major challenge stems from biases inherent in the dataset. Such biases may arise for example, when the labeling process was performed in an already unfair manner, or if there are under-represented populations in the dataset, or in the case of systematic lack of data and in particular labels"* (Pessach & Shmueli, 2020, 20)
			"*Often those undertaking data collection are separate from those responsible for the remaining model lifecycle; e.g. the collectors might be from a different institution or organizational unit. Information on the decisions made as part of data collection is important for understanding the potential risks, limitations, and implications*" (Cobbe, Lee, Singh, 2021, 604)
		Problem Formulation	"*Records relating to the aims of and rationale for the algorithmic system, giving insight into the values and norms behind its commissioning, development, and operation, are therefore relevant to various ac-countability relationships. Documentation and other records of the system's aim, scope, and justification—what it will do, why it is required, and the role it will play are worth considering*" (Cobbe, Lee, Singh, 2021, 603)
			"If we wish to take seriously the work of unpacking the normative implications of data science systems and of intervening in their development to ensure greater fairness, we need to find ways to identify, address, and accommodate the iterative and less visible work of formulating data science problems" (Passi & Barocas, 2019, 9)
	Approach	Rationality	*"Anything that can be doubted is eliminated"* (Birhane, 2021, 3)
			"Because the rationalist's focus is to uncover what remain constant regardless of context, culture, and time, the rationalist view embraces abstraction, generalization, and universal principles at the expense of concrete, particular, and contextual understanding - that is, knowledge grounded in active, concrete and reciprocal relationships" (Birhane, 2021, 6)
		Relationality	*"Knowledge claims are not worked out in isolation from others but are developed in dialog with the community. It is taken for granted that there exists an inherent connection between what one does and how one thinks"* (Birhane, 2021, 4)
			"Thinking in relational terms about ethics begins with reconceptualizing data science and machine learning as practices that create, sustain, and alter the social world" (Birhane, 2021, 7)
			We draw on an understanding of ADM [automated decision making] as a socio-technical process involving both human and technical elements, beginning before a decision is made and extending beyond the decision itself" (Cobbe, Lee, Singh, 2021, 598)

Pre-process methods: One of the concepts identified from the pool of papers concern decisions made pre-process when developing a model. The papers by Cobbe et al. (2021), Passi and Barocas (2019) and Pessach and Shmueli (2020) argue that by considering problem formulation and data gathering as important first steps of the process, fairness can potentially be enhanced. We identified these as *pre-process methods* since this is the foundation the remaining process relies on.

> ***Problem Formulation:*** Mitchell et al. (2021) argue that the formulation of a problem is one of the first places where AF can be challenged, and that it is important that the problem formulation is considered carefully even before a development process begins (158). The article concludes that biased or insufficient problem formulations can increase the risk of unfairness in the results that a model provides in the end (146). Their point is therefore not merely to use a technical mechanism, but a consideration of fairness from the first discussion of why and how a given problem should be solved using a model. According to Cobbe et al. (2021) this consideration cannot stand alone. They emphasize the importance of documenting the process of defining a problem formulation, because it will provide traceability of values and norms leading to decisions (603). This should be supported by a clear documentation of the decision-process throughout the development of the model (604).

> ***Data gathering:*** According to the identified articles the state of data sets used for training and testing a model can affect the level of fairness in the final product, since the foundation of a model is the available data (Mitchell et al., 2021, 144). Mitchell et al. highlight the challenges regarding fairness that can occur if data is not carefully selected and investigated for bias. They argue that there exist two types of biased data: statistical biased data, which consists of non-representative data and measurement errors and societal biased data that reproduces retrospective injustices (145). Pessach and Shumeli (2020) also identify a number of issues that can be caused by datasets. Bias can be implicitly included in the collected data (20) and missing values can affect the representativeness in the dataset (2). The state of data is not the only thing to consider. In the paper by Cobbe et al. (2021) it is stressed that data collection can also affect fairness. It might be different people who collect the data and develop the model, so it is important that information regarding this process is documented such that decisions, limitations and risks regarding data collection are clearly stated (604).

Approach: In the articles we identified two approaches on how a higher level of AF can be affected within a model - rationality and relationality. These approaches are taken from Birhane (2021). We

consider approaches as something similar to an ideology in which models originate as a product of the data and context they are part of.

> ***The rational approach:*** According to Birhane (2021) rationalists focus on uncovering "*what remain constant regardless of context, culture, and time*" (6). Rationality handles knowledge as something that is *"..rooted in the ideal rational, static, self-contained, and self-sufficient subject that contemplates the external world from afar in a ''purely cognitive'' manner as a disembodied and disinterested observer*" and data is handled as a neutral representation of the world where everything that is odd is not tolerated (3). Birhane argues that in the rational approach values regarding ethics and morals are separated from scientific work and something that scientists should not contaminate their work with. It should not influence the outcome of their scientific work as they believe that abstract and intellectual thinking is the most reliable groundwork and that emotion, culture and ethnicity should be left out. Within this belief the scientists are invisible and their choices and motivation regarding data is never questioned (4).
>
> ***The relational approach:*** Birhane (2021) suggests relationality as a contrast to rationality (1). Relationality is rooted in the belief that everything co-exists in a web of relations (3). Instead of thinking exclusively in datasets as in the rational approach, scientists should consider the data setting in order to gain a real perspective of the context and the world they are part of. In the relational approach scientists are a part of the development process where a team can cooperate on discussing their choices and motivations in developing datasets and this can help make models more fair, transparent and less biased. In this approach technical solutions to problems are partially open, e.g. a dynamic, reiterative and revised practice, instead of a static solution to ever changing problems (as in a rational approach) (6).

3.2.2 Challenges when Improving Algorithmic Fairness

The display above encapsulates our focus and main points on the concept AF. The most relevant findings identified in the literature are two general concepts that we find crucial for the enhancement of AF - the pre-process methods of handling data and problem formulation and the approaches used prior to and during a development process.

We analyze potential enhancement of AF evolving around the approaches - rationality and relationality. This is done in order to discuss pros and cons as well as possible overlaps and disagreements regarding the improvement of AF between the two approaches later on.

Several of the analyzed articles point to a tendency toward thinking primarily in technical terms and/or model building when discussing AF, thereby giving lower priority to human aspects and the

contexts of which the technology is part of (Birhane (2021), Cobbe et. al. (2021), Passi & Barocas (2019)).

As described above, Birhane (2021) finds that rational approaches in research on AF has focused too much on technical solutions and thereby forgotten the historical and ethical context that automation is part of. Passi and Barocas (2019) and Cobbe et. al (2021) also take their analytical starting point from this perspective. As well as Mitchell et. al. (2021) who describe how research on fairness often takes socially relevant aspects of the pre-process and initial state as given in a development process (143). Although these articles do not use the term rationality when describing the tendency to focus on technological aspects or taking socially relevant aspects for granted it is the same point as described by Birhane in what she calls a rational approach to technological solutions.

It should be stated that although the articles in some way criticize the primary focus on technological aspects, they do consider model building an integral focus point for enhancing fairness. However, they emphasize other parts of a development process, such as pre-process methods, as being equally important.

As a push back to what she calls the dominant orthodoxy of rationality, Birhane suggests an ethical framework rooted in relationality, where complex issues such as bias, fairness and justice in algorithmic models are grounded in the social and historical contexts they are a part of (2).

The relational approach is in close relation to the concept reviewability introduced by Cobbe et al. (2021) which is a framework introduced to improve accountability in model development processes (598). According to Cobbe et. al. accountability is not something that exclusively exists in or applies to the implementation part of a process. Rather it is part of every aspect of the process from problem description to auditing and each step of the development process should therefore be carefully considered and logged (602). This is a (cyclic) way of ensuring that information regarding data gathering and structuring is accounted for - a lack of which Birhane points to as problematic in the rational approach. Cobbe et. al. furthermore emphasize that what should be accounted for and to whom there should be accounted differ depending on which area of development one is involved and is therefore highly contextual (599). We identify this contextuality as an implicit acknowledgement of a discretionary presence, which is also described as situated (601).

By accounting for various assumptions and choices in a model-development process, as well as the reasons behind these choices, different aspects regarding whether a model is fair will to a greater extent become reviewable and revisable (605). This focus acknowledges that datasets are never without context or human decisions and therefore never by default unbiased. Passi & Barocas (2019) also point out that choices such as algorithmic tasks, business requirements and choices of labelling data are always dependent on the *"... practitioners' normative commitments and beliefs*" (Passi & Barocas, 10).

None of the analyzed articles are in complete compliance with the aim of a relational approach as described by Birhane, in which the "ultimate" goal is understanding above prediction and which rejects the notion that a dataset can ever be rid of biases. But they point to a shift in the handling of AF enhancement where the human actors involved in shaping technical solutions are acknowledged as being active and contextual contributors.

Other recurring subjects are issues regarding fairness and bias as something to be approached processually, e.g. interchangeable in each step of a development process - this is found in the articles by Birhane (2021), Passi and Barocas (2019) and Cobbe et. al. (2021). Furthermore, all articles point to the entire process as being cyclic when dealing with complex issues such as fairness.

In Passi and Barocas (2019) the focus is on the pre-process activity of problem formulation, and how problem formulation plays an integral part if we wish to ensure fairness when developing models that handle high-level abstract and complex goals (Passi & Barocas, 10). They argue that the problem formulation is an iterative process that involves continuous involvement of human actors - such as the development team and the stakeholders - who have differentiating goals that influence the scope of the problem to be solved (8). Cobbe et al. also mention problem formulation and argue both that it is a part of the entire process and that norms in the problem definition should be documented (604). The argumentations in both articles can be said to be in alignment with Birhanes definition of relationality, since they are founded in the belief that the way a problem is formulated and the thoughts and decisions that influence these formulations are essential for the ability to prevent bias and unfairness.

As mentioned, relationality is suggested by Birhane (2021) as an approach for the development processes, since a strictly rationalist approach can cause biases and unfair results for historically marginalized groups (2). Data gathering is a crucial aspect of this, since data scientists decide what is worth measuring in the first place and important information regarding data collection and structure is at risk of being removed (4). In a relational approach this process is considered carefully and is also here in line with Cobbe et al.'s (2021) definition of reviewability, since documenting the decisions made improves accountability and transparency.

The issue of bias caused by data gathering also appears in the articles by Alkhatib and Bernstein (2019), Cobbe et al. (2021), Mitchell et al. (2021) and Pessach and Shmueli (2020). Based on these articles we identified two main discussions regarding bias: One concerns bias found in data, the other in the ethics of the approach when working with algorithmic models and how this can cause bias. As shown in the displays' sub-dimension data gathering (Table 3), Mitchell et al. differentiate between statistical and societal bias as two aspects to keep in mind when handling data. Alkhatib and Bernstein (2019) underline the same points as Mitchell et al. (2021) when it comes to reproducing societal bias from datasets and state that a dataset can never be large enough to

represent the diversity of the real world because "*increased training data is insufficient: for important cases at the margin, there may be no prior cases*" (7).

Pessach and Shmueli (2020) reach similar conclusions when discussing that labeling of data can result in bias, even if the data itself is not necessarily biased (17). However, they also stress that increased fairness can cause loss of accuracy and that it is a question of finding a balance where neither loss of fairness nor accuracy is significant (7).

The notion of societal bias as seen in Mitchell et al. (2021) can be linked to Birhanes observation that discussions regarding algorithmic fairness and development processes tend to overlook marginalized groups. Mitchell et al. also write that "*incorrectly assuming that a sample is representative can lead to biased estimation of conditional probabilities*" (145), which is closely related to Birhanes interpretation of a rational approach, in which the representativeness of the data is not considered (4).

All papers point to challenges in automated decision making, but two papers also state that automation cannot blindly be replaced by humans. Mitchell et al. (2020) underline that humans are not perfect either: They are just as capable of making mistakes as algorithms (158). Alkhatib and Bernstein (2019) claim that caseworkers do not guarantee a sufficient level of fairness, accountability and transparency either as "*street–level bureaucrats have historically been agents of immense prejudice and discrimination*." (530).

Through our analysis we have examined pre-process methods and approaches that can either hinder or enhance fairness. This leads us to the preliminary conclusion that the state-of-the-art literature on this field considers AF to be highly contextual and dependent on the processes. A high level of AF is only achieved if the possibility of biased data, problem formulation and ethics is considered throughout the entire development process - preferably from the moment the idea to make a model occurs. However, the articles show slight differences in what an ideal process should look like and tend to highlight different aspects to be important, which we consider underlining the complexity of AF as well as the problems in ensuring a high level of AF. All articles consider the responsibility to be human, since enhancing fairness lies in the development process where everyone involved needs to carefully consider the people they make decisions for.

In the next section we will discuss some of these challenges and compare them to the challenges we identified in our analysis of discretion to identify potentials and pitfalls in automating discretion in the public sector.

4. Discussion of Intersections and Contrasts

Through our research, we have identified a number of overlaps between our findings on discretion and AF. First and foremost, we identify an important correlation between the two perceptions of discretion: *collaborative* and *individual,* and the two approaches to AF: *rationality* and *relationality*.

The articles by Bovens and Zouridis (2002) and Justesen and Plesner (2019) include arguments that automated decision making ensures greater objectivity than decisions made by humans, since discretionary decisions rely on individual caseworkers' moods and opinions. These articles can be said to include aspects that view technology in a traditional rationalistic way since these aspects consider the objectivity of technology given and the subjectivity of caseworkers to be irrelevant.

In contrast, the articles by Passi and Borocas (2019) and Cobbe et. al. (2021) highlight technology as always influenced by human opinions and choices, and tend to view the development of technology as an iterative and collaborative effort, involving many active actors, where decisions are revisited throughout development. This more relational approach correlates to the definition of collaborative discretion where human aspects and decisions are considered and revised continuously as seen in Bullock (2019), Petersen et al. (2020) and Petersen et al. (2021). We consider this duality (rationality/individual discretion and relationality/collaborative discretion) as crucial as it highlights two dominant ways of approaching the process when automating the public sector.

The aspect of process is important to highlight, since this is another interesting overlap between the articles that consider it essential for enhancing AF (Cobbe et al. (2021), Mitchell et al. (2020), Passi and Barocas (2019) and Pessach and Shmueli (2020)) and the articles on collaborative discretion (Bullock (2019), Petersen et al. (2020), Petersen et al. (2021) and Young et al. (2019)). Both emphasize either a need for cyclical processes to enhance fairness or discretionary casework as being cyclic/iterative to ensure a more valid foundation for decisions. In AF we consider this to be grounded in relationality by the reviewability framework (Cobbe et. al,) where the process is considered more important than the decision itself and where the discretionary practice lies (602). Furthermore, both the articles on AF and those on discretion acknowledge that human discretion cannot be completely eliminated from any process as it will either be practiced by caseworkers or by developers.

On the other hand, Bovens and Zouridis and Justesen and Plesner do not discuss in detail how to handle this unavoidable discretionary practice in the development of models, and they do not specifically articulate the relationship between discretion and AF in the process of model development. This is in contrast to the articles mentioned above who believe that the discretionary power is moved from caseworkers to developers, where especially Justesen and Plesner believe that discretion is removed completely when automating decision-making.

Another overlapping area exists in the point that caseworkers are regularly forced to make *situated judgments* in order to elaborate on real-world situations that do not necessarily comply with goals defined by law (Petersen et al., 2021). This can be related to the concept of reflexivity described by Alkhatib and Bernstein (2019), since models are only able to reflect on a decision after it has been made. Additionally, we know from Petersen et al. (2021) that there are situations where caseworkers do not write everything down. This results in incomplete data regarding cases, which can potentially cause problems when training models to make these decisions. This is a paradox,

since automating decision-making based on incomplete data provided by caseworkers can undermine the collaborative aspects of discretion made by humans. Also, it results in models making wrongful decisions, since we also established that models are only as good as the data we provide them with.

In the case of collaborative discretion and AF it is not only the overall approach that overlaps. The literature also identifies a number of steps in the processes that are similar. Reflexivity is highlighted as important to both collaborative discretion and when enhancing AF. Alkhatib and Bernstein (2019) argue that reflection is necessary in the discretionary process, however it is almost impossible to conduct it technically without losing accuracy. Together with Mitchell et al. (2021) they argue that best practice of human reflexivity is more nuanced and able to consider long-term consequences, which algorithms cannot.

From our analysis we have also detected a number of relevant contrasts in the literature besides from the general contrast between collaborative discretion and relationality and individual discretion and rationality. One important contrast to highlight is between the way of thinking of discretion as *highway* and *side road* and the articles on AF. According to Justesen and Plesner simple cases can be fully digitized because they are unambiguous. However, none of the articles on AF discuss how to deal with these cases, nor do they acknowledge that some cases might not need a high level of human discretion.

4.1 Knowledge Gaps

Based on the sections above we identify discovered knowledge gaps.

We have found that the articles consider that there are pros and cons to both human discretion and artificial discretion when it comes to accuracy and bias, but the specific differences in bias between the two kinds of discretion are not investigated in detail. Future research could benefit from a more detailed investigation of pros and cons in human discretion vs. artificial discretion to see how bias occurs in cases involving a high level of discretion. This is also in relation to the question whether machines are actually more capable at making accurate and fair decisions than humans. The analyzed articles are mainly qualitative, meaning that the papers do not give a quantifiable perspective. By making a comparative study one could investigate how discretion is performed, practiced and how fairness is enhanced in human or automated decision making – and if the decisions differ at all.

Other areas which could benefit from further investigation is empirical research related to how different aspects of the process are handled in the intersection between AF and discretion. Research could be conducted on how multiple actors of a development process approach or consider decision making. This could be by investigating if following a framework such as reviewability has an effect on the result of a model or the presumption of fairness, and in which

ways. Another focus could be to investigate what would happen to the outcome of a model if caseworkers are considered more directly in data gathering and in the process of building a model.

Yet another aspect that could benefit from more detailed research is when automation of casework involving a high degree of human discretion is beneficia, in particular for the employees that these tools impact and for the citizens that are impacted by the decisions. In relation to this, it could be investigated which problems arise if we automate casework with a high degree of discretion and how public governments would handle these potential problems.

5. Conclusion

Through the literature we have been able to identify two perceptions of discretion: The collaborative and the individual. The perspectives on the practice of discretion arise from different views on what the practice of discretion contains and how it is conducted. Where those who think discretion as an individual practice consider automation to be a tool to avoid bias and prejudices, those who consider discretion to be collaborative see automation as a cause of the same.

The divisions are not as strict when it comes to AF which is considered to be highly contextual and defined by the processes in which the model is developed. All analyzed articles circle around approaches and pre-process methods as necessary tools to enhance algorithmic fairness.

The literature proposes no clear or unambiguous answer to how automated decision-making can be implemented in the public sector. All articles represent some possibilities and challenges that can arise when it comes to automating public government activities that involve discretion. Where the collaborative and/or relational approach argue that automation of discretion will cause fairness issues, they do not define an exact limit to which automation is possible. Those who have a more rational and/or individual approach to automation and discretion argue that bias can be limited by automation. But there is no strict answer on how to avoid bias and consider the unavoidable human interaction in the development of decision-making tools for cases that traditionally involve a high degree of human discretion.

Thus, both overall approaches contain known as well as unknown challenges and there exist several areas for further research.

6. Literature

Alkhatib, A. & Bernstein, M. (May 4-9, 2019), Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions, *CHI Conference on Human Factors in Computing Systems Proceedings* [Paper presentation], ACM, Glasgow, Scotland, United Kingdom, 1-10

Birhane, A. (2021), Algorithmic injustice: a relational ethics approach, *Patterns, Vol 2(2)*, 1-9.

Bovens, M. & Zouridis, S (2002), From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative

Discretion and Constitutional Control, *Public Administration Review, Vol 62 (2)*, 174-184

Bullock, J. (2019), Artificial Intelligence, Discretion, and Bureaucracy, *American Review of Public Administration, Vol 49 (7),* 751-759

Cobbe, J., Lee, M. S. A. & Singh, J. (March 3-10, 2021), Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems, *ACM Conference on Fairness, Accountability and Transparency (FAccT)* [Paper presentation]*,* ACM, Virtual Event, Canada, 598-609

Digitaliseringsstyrelsen (12-07, 2020) *Ny FN-måling: Danmark er fortsat verdensmestre i offentlig digitalisering,* link

Digitaliseringsstyrelsen (n.d.), *Hvad er digitaliseringsklar lovgivning?* link

Justesen, L. & Plesner, U (2018), Fra skøn til algoritme: Digitaliseringsklar lovgivning og automatisering af administrativ sagsbehandling, Tidsskrift for Arbejdsliv, Vol. 20 (3), 9-22

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage Foundation

Mitchell, S., Potasch, E., Barocas, S., D'Amour, A. & Lum, K. (2021), Algorithmic Fairness: Choices, *Assumptions, and Definitions, Annual Review of Statistics and Its Applications, Vol 8*, 141-163

Olsen, L. (n.d.) *Signaturprojekter med kunstig intelligens i kommuner og regioner,* Digitaliseringsstyrelsen, link

Passi, S. & Barocas, S. (January 29-31 2019) Problem Formulation and Fairness, *Proceedings of the ACM Conference* (FAT' 19), ACM, Atlanta, Georgia, USA, 1-10

Pessach, D. & Shmueli, E. (2020), Algorithmic Fairness, arXiv preprint arXiv:2001.09784, 1-31

Petersen, A. C. M., Christensen, R.L., Harper, R. & Hildebrandt, T.T. (2021), We Would Never Write That Down: Classifications of Unemployed and Data Challenges for AI, *PACM on Human-Computer Interaction Vol 5 Article 102,* 2-23

Petersen, A. C. M., Christensen, R.L., & Hildebrandt, T.T. (2020), The Role of Discretion in the Age of Automation, *Computer Supported Social Work (CSCW), 29,* 1-27

Rienecker, L. & Stray Jørgensen, P. (2020). *Den Gode Opgave: Håndbog i Opgaveskrivning på videregående uddannelser*. Copenhagen, Denmark: Samfundslitteratur

Webster, J. & Watson, R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly, Vol 26.(2)*, xiii-xxiii

Young, M. M., Bullock, J. & Lecy, J. D. (2019), Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration, *Perspective on Management and Governance, Vol 2 (4),* 301-313

Appendix 1 – Concept Matrix

		Concepts								
	Articles	**Algorithmic Fairness**			**Discretion**			**Artificial Intelligence**		
	Sub concepts:		Fairness definition & measurement	Fairness enhancement	System-level bureaucracy	Legality / democracy of discretion	Human intuition / actions	AI in public services	AI (general)	Limitations
1	Alkhatib, A. & Bernstein, M. (2019), *Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions*	x		x	x		x			
2	Birhane, A. (2021), *Algorithmic injustice: a relational ethics approach*	x								
3	Bovens, M. & Zouridis, S (2002), *From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control*				x	x	x	x		
4	Bullock, J. (2019), *Artificial Intelligence, Discretion, and Bureaucracy*	x			x	x	x	x	x	x
5	Cobbe, J., Lee, M. S. A. & Singh, J. (March 3-10, 2021), *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*	x							x	
6	Justesen, L. & Plesner, U (2018), *Fra skøn til algoritme: Digitaliseringsklar lovgivning og automatisering af administrativ sagsbehandling*						x	X	X	
7	Mitchell, S., Potasch, E., Barocas, S., D'Amour, A. & Lum, K. (2021), *Algorithmic Fairness: Choices, Assumptions, and Definitions*	x								
8	Passi, S. & Barocas, S. (2019), *Problem Formulation and Fairness*			x					x	x
9	Pessach, D. & Shmueli, E. (2020), *Algorithmic Fairness*		x	x						
10	Petersen, A. C. M., Christensen, R.L., & Hildebrandt, T.T. (2020), *The Role of Discretion in the Age of Automation*				x		x	x		
11	Petersen, A. C. M., Christensen, R.L., Harper, R. & Hildebrandt, T.T. (2021), *We Would Never Write That Down: Classifications of Unemployed and Data Challenges for AI*	x			x	x	x			x
12	Young, M. M., Bullock, J. & Lecy, J. D. (2019), *Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration*						x	x	x	