UNDERSTANDING THE DIGITAL WELFARE STATE: THE FAILURE OF ALGORITHMIC PROFILING OF JOBSEEKERS IN AUSTRIA AND POLAND

By
Aiste Vaitkeviciute

Submitted to Central European University
School of Public Policy

In partial fulfilment of the requirements for the degree of MUNDUS MAPP in European Public Policy

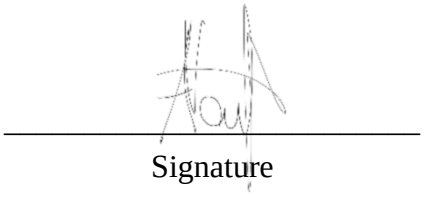
Supervisors: Professor Cameran Ashraf and Professor Jeremy Moulton

Budapest, Hungary
2021

Author's declaration

I, the undersigned, **Aiste Vaitkeviciute**, candidate for the MA degree in MUNDUS MAPP European Public Policy declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Budapest, 15 July 2021



Signature

Table of Contents

Glossary

ADMS – Automated Decision-Making Systems

AMS – Austrian Public Employment Agency

MLSP – Polish Ministry of Labour and Social Affairs

MSF – Multiple Streams Framework

PES – Public Employment Services

PUP – Polish Public Employment Services

Abstract

This thesis analyses the factors behind the termination of algorithmic profiling systems used in the public employment services in Austria and Poland. In both countries, the Ministries of Labour and Social Affairs adopted algorithmic profiling systems as innovative tools to modernise and optimise social protection services. Eventually, they had to terminate their use due to discriminatory impact and privacy violations. This thesis aims to shed light on the socio-technical factors that contributed to the failure of algorithmic profiling and thus contribute to the study of the risks related to the 'digital welfare state.' The public sector, especially the social protection and welfare domain, is often missing from automation and digitalisation discussions. This thesis uses J. Kingdon's Multiple Streams Framework (MSF) to analyse how the three independent streams, problem framing, policy development, and political conditions, shaped the fate of algorithmic profiling systems in Austria and Poland. By comparing two cases, this thesis aims to understand the processes, events, and developments that led to the termination of algorithmic profiling rather than improved governance, regulations, and supervision of these tools. The analysis revealed that the lack of policy-makers' attention to the problem largely contributed to the failure of these systems.

1. Introduction

This thesis aims to understand what factors led to the failure of algorithmic profiling systems used in public employment services in Austria and Poland. It seeks to better understand the risks associated with automated decision-making systems (ADMSs) in social protection services and the opportunities for mitigating those risks. In the last ten years, the unprecedented advances in data gathering, computing, and algorithmic systems enabled a rapid uptake of automated decision-making systems (ADMSs) in various fields, including the social protection services. The applications of ADMSs range from automating routine, repetitive tasks to managing and controlling the workforce to automating profiling, risk prediction, and eligibility assessments (AlgorithmWatch, 2019, 2020). The ADMSs refer to algorithmically controlled procedures that use data to infer correlations and derive information deemed useful for decision-making. Increasingly, decisions, and operations, previously left to humans, are delegated, fully or partially, to algorithms (Mittelstadt *et al.*, 2016). The public sector often applies ADMSs in policing and law enforcement (France, Switzerland), managing and delivering healthcare services (Italy), predicting risk of neglect or social exclusion (Denmark, Finland), predicting risk of welfare or tax fraud (the Netherlands, Denmark, France), determining eligibility for social welfare support (the Netherlands, Estonia, Austria, Poland) (AlgorithmWatch, 2019, 2020).

Many of these systems have been criticised for the lack of transparency, compatibility with privacy and anti-discrimination laws, as well as for their potential to jeopardise people's rights and well-being (AlgorithmWatch, 2019, 2020). In recent years, a pattern has emerged that ADMSs, rapidly adopted in secrecy with little or no due diligence, eventually are terminated due to high error rates, unfair and discriminatory outcomes, and detrimental social impacts. The most notable

examples include the 'Robodebt' scandal in Australia in 2016-2019, where an algorithmic fraud prediction tool falsely accused thousands of people of fraud, leading to unexplained and unjustified indebtedness (Mao, 2020). The SyRI algorithm, a welfare fraud prediction tool used in the Netherlands in 2016-2019, was ruled illegal after it was proven that it was only used selectively in the poor districts to police mostly immigrant communities and that the way it was collecting data was highly invasive and discriminatory (Vervloesem, 2020). In 2020, the Human Rights Watch reported that the Universal Credit system in the UK – currently still in use – exacerbates poverty by miscalculating benefit entitlements, leaving little or no room to seek explanation, justice, and compensation, especially for the most vulnerable that might not necessarily have digital skills necessary to navigate this system (Toh, 2020).

The ADMSs used for profiling, which are discussed in this thesis, determine how individuals or groups are categorised, classified, and managed based on certain characteristics and thresholds (Mittelstadt *et al.*, 2016). The use of ADMSs for profiling in social protection services makes these systems inextricably connected to broader institutional, political, socio-economic contexts. Therefore, the main concern about the ADMSs for profiling is that through categorisation, they may reinforce, exacerbate, and perpetuate the existing patterns of inequality, oppression, and marginalisation. The technology adds an additional layer to these risks by making algorithmic decisions more difficult to trace and challenge. The investigations of ADMSs are critical, as it is important to understand not only the risks associated with the ADMSs, such as violations of rights to equal treatment and privacy and detrimental socio-economic effects, but also why these risks are often overlooked and not adequately addressed. While this thesis looks at the ADMSs for profiling that are based on simple, algorithms, these concerns are even more relevant

in the context of the increasing use of more advanced AMDSs based on self-calibrating machine learning algorithms.

Many factors influence how the ADMSs is used, what impact it has on target groups, as well as whether and how potential risks are mitigated. This thesis aims to understand why algorithmic profiling tools used in public employment services in Austria and Poland were terminated, that is why the identified risks of discrimination and privacy violations had not been addressed and instead, the systems were cancelled. The findings will contribute to the broader understanding of whether the automation of social protection services, albeit possible, should be actively pursued to improve service delivery, access, and efficiency or any such attempts are meant to fail. To solve this puzzle and capture the multidimensionality of the ADMSs, this thesis will adopt Kingdon's Multiple Streams Framework (MSF) to analyse why the ADMSs in public services in Austria and Poland failed. The MSF posits that policy changes when three independent streams – problem, policy, politics – come together to open a 'window of opportunity for a change. This thesis shows that the window of opportunity did not open because the problem, policy, and politics streams did not merge due to the lack of policy entrepreneurship and political reluctance to deal with a problem of algorithmic bias and discrimination.

Following the introduction, this thesis contains five other chapters. The second chapter reviews the relevant literature on ADMSs, algorithmic bias and discrimination and the digital welfare state. The third chapter explains the research design – conceptual framework, case study selection, and data collection process. The fourth chapter presents case study analysis, where each stream – problem, policy, and politics – are discussed in three separate sub-sections, comparing the key actors, events, processes in Austria and Poland. The fifth chapter discusses the key findings of the comparative case study and identifies the main factors that led to the failure of the ADMSs

system. The last chapter concludes the thesis, summarises the main argument and points out areas for further research. The thesis report written in preparation for this thesis is attached as an annex at the end of this thesis.

2. Literature review

This section provides a literature review on automated decision-making systems (ADMSs), algorithmic profiling, bias and discrimination, and the digital welfare state. This section introduces the key problems and debates on these topics and demonstrates that the discussion of algorithmic bias in the digital welfare state has been largely overlooked. To understand the main issue related to the digital welfare state, it is critical to understand the concepts of automation, digitalisation, and datafication. In general, automation refers to the use of advanced technologies to minimise or completely remove human input. Digitalisation is a process of incorporating digital technologies into different areas of life by transforming processes to such that they could be read by computers (IBM, 2020). Datafication refers to a process through which social activities and interactions are transformed into quantified data, which enables large-scale data collection, tracking, and analysis (Cukier and Mayer-Schoenberger, 2013). Taken together, automation digitalisation and datafication have powered an increasing uptake of automated decision-making systems (ADMSs) in private and public sectors. The ADMSs refer to algorithmically controlled procedures that use data to infer correlations and derive information deemed useful for decision-making. Increasingly, decisions, operations previously left to humans, are delegated to algorithms, and are used to inform subsequent actions, decisions, and policies (Mittelstadt *et al.*, 2016). Anticipated high economic gains, improvements in efficiency often drive the rapid uptake of these systems. However, the critics of the ADMSs systems warn against a blind 'technosolutionism' that often drives the development and deployment of these technologies. Technosolutionsim refers to a belief that with the right code and technology, all complex human problems can be easily resolved (Morozov, 2013). Technosolutionism as a discourse frames all complex social situations as easily optimisable

problems with definite, computable solutions (Morozov, 2013; Broussard, 2019). Technosolutionism is also a driving force behind the proliferation of large amounts of poorly designed algorithmic systems that are being deployed without accompanying regulatory, governance, and technical safeguards against potential risks (Broussard, 2019).

2.1. Algorithmic bias

One of the key risks associated with ADMSs is algorithmic bias, which is mostly discussed in the context of the private sector. Algorithmic bias based on gender, race or other proected characteristic i is often discussed in the context of credit scoring and loan eligibility assessments (see Pasquale and Citron, 2014); algorithmic workforce management (Mateescu and Nguyen, 2019; Duggan *et al.*, 2020); Internet platforms and search engines (Sandvig *et al.*, 2014; Noble, 2018); targeted online advertising (see Lambrecht and Tucker, 2016). In the public sector, most attention to the use of algorithmic technologies has been paid in the areas of predictive policing and security (see Brayne, 2020), which often includes the use of advanced technologies, such as facial recognition, where racial and gender biases were proven ubiquitous (see Buolamwini and Gebru, 2018). The ADMSs in other aspects of the public sector have been largely overlooked. Because a government is often seen as a regulator of new technologies rather than a 'user' of these technologies, the focus in the literature is often on the governance and regulations of the new technologies used in the private sector or security sector rather than those used in other functions of the government, such as social welfare provision (Kuziemski and Misuraca, 2020).

2.2. Sources of Algorithmic bias

Friedman and Nissenbaum (1996) differentiate between different types of biases – pre-existing, technical, emergent, correlations, and feedback loops. Sandvig (2014) adds that the intended purpose with which the algorithm is designed can also be a source of bias (Sandvig *et al.*, 2014). Overall, these sources could be broadly divided into technical and socio-technical sources of bias. The technical sources of bias stem from algorithmic model specification, variable selection, or the data set (Friedman and Nissenbaum, 1996; Mittelstadt *et al.*, 2016).

The technical bias emerges from technical constraints or design limitations, such as algorithmic models, technical specifications, and data sets. Different model configurations, combining variable selection, parameter selection (e.g., coefficients and thresholds for categorisations) can lead to differing outcomes even if used in the same context (Barocas and Selbst, 2016; Lopez, 2019). Data used in the ADMSs can reproduce bias if the data available reiterates the historical patterns of marginalisation, exclusion, and discrimination. For example, bias in data can show up as missing data, overrepresentation of certain groups (Pasquale and Citron, 2014; Barocas and Selbst, 2016), or the lack of disaggregated data by gender, race or other characteristic (see Criado-Perez, 2019). With the increasing technical capacities of algorithmic technologies, the bias from correlations has become an increasingly more pressing problem. It refers to the problem when algorithms combine the data from a variety of sources and infer correlations that would not and often should not be known or used otherwise, e.g. protected characteristics (Andersen, 2018; Desiere and Struyven, 2020).

The socio-technical sources of algorithmic bias include the purpose of their design and application and how it is used to inform further social actions (Mittelstadt *et al.*, 2016; O'Neil,

2016). The socio-technical types of bias include pre-existing bias, bias by design and intention of use, emergent, unanticipated use bias, and the feedback loops (Friedman and Nissenbaum, 1996). The pre-existing bias refers to the bias embedded into technology due to existing broader power dynamics, socio-economic disparities, cultural norms, and specifically refer to explicit and implicit prejudices embedded in the system, e.g., assumptions about gender roles, identities (Friedman and Nissenbaum, 1996; Kraft-Buchman and Arian, 2019). The bias by design refers to the fact that algorithms are designed for a certain purpose and they are optimised to achieve certain goals, which reflect the interests and values of those who designed them. Therefore, the algorithms can lead to outcomes that meet the intended goal but are not necessarily fair, equitable or ethical (Sandvig *et al.*, 2014). Algorithms can also produce bias results when they are used in an unanticipated manner and when they are applied in new, and changing contexts and cannot capture the changing cultural norms, meanings, and knowledge. Lastyl, the previously discussed types bias in algorithmic systems can trigger feedback loops that further distort the results. Feedback loopsccur when data used in the algorithmic model leads to responses in the real world which are fed back into the algorithm, thus reinforcing and encouraging certain patterns of behaviour and are particularly dangerous when they are applied at large scale targeting specific groups of people (Friedman and Nissenbaum, 1996; O'Neil, 2016).

The algorithmic bias, whether technical or socio-technical, can have detrimental implications on peoples' lives. Scholars warn, however, that statistical error rate is inherent to statistical and thus algorithmic profiling (Desiere and Struyven, 2020). A study of different statistical profiling models across OECD countries showed that currently, the best performing profiling models have a 70 % accuracy rate. But with increasing improvements and adjustments in statistical models, development of new techniques, improving data collection methods, the

probability of errors is decreasing (Desiere, Langerbucher and Struyven, 2019). In the ADMSs for profiling, statistical error rates can result in inaccurate classifications of individuals into groups. However, scholars highlight that statistical error rates per se do not automatically have a detrimental impact on people (Desiere and Struyven, 2020, O'Neil, 2016). Even if the ADMSs were 100 % accurate, inequitable and unfair outcomes can still occur. The determining factor in whether or not algorithmic bias will have detrimental impact on targeted individuals or groups of people is how the ADMSs are used and applied, which includes the purpose of use, the context, the interests and positionality of those using the ADMSs, etc. (Friedman and Nissenbaum, 1996; Mittelstadt *et al.*, 2016; O'Neil, 2016). In algorithmic profiling, an unemployed person from a disadvantaged background e.g., with migration background, disability is more likely to be misclassified to the lower-category, based on the group characteristic (group disadvantage) rather than individual qualities (Desiere and Struyven, 2020). However, whether or not this classification creates a risk of discrimination and exclusion depends on how the algorithmic result informs the subsequent decisions, actions, and responses. For example, in Flanders (Belgium), the profiling model is used to prioritise jobseekers for appointments and support caseworkers, while in other contexts it is often used to allocate support. The management of job seekers' appointments versus the decision whether or not a job seeker receives support have different implications and raise different risks (Desiere and Struyven, 2020).

2.3. Algorithmic Opacity and potential solutions

The key challenge in identifying and addressing algorithmic bias and the related risks is algorithmic opacity. Algorithmic opacity refers to the lack of transparency in the way algorithms

are designed, operate, and impact people. Burell (2016) distinguishes among three main forms of opacity: 1) opacity as intentional business secrecy; 2) opacity as the lack of technical literacy; 3) opacity from the technical characteristics of algorithms (Burrell, 2016). The third source of bias, the technical, is mostly relevant for more advanced algorithmic systems that use self-calibrating machine learning algorithms, which are not discussed in this thesis. The first two sources of bias are particularly relevant for the ADMSs used for profiling in public employment services. First, because many AMDSs used in the public sector are often developed by private contractors, the models and other details are often protected by agreements between the private company and the public sector's agency involved. Second, the public sector agencies could decide to maintain the algorithmic models and relevant statistics on their impact secret in order to prevent manipulation. For example, if it is known what behaviours or features are marked as risky in the model for welfare fraud prediction, then people dealing with the system could adjust their behaviour to avoid being flagged as 'risky households.' Third, even if governments were to make relevant models, data, and other relevant information available to the public or relevant stakeholders, the public or the stakeholders have to have sufficient knowledge to understand and use the available information (Burrell, 2016).

The literature on potential solutions distinguishes among three sets of tools to address algorithmic bias – technical, transparency and accountability, and regulatory (Friedman and Nissenbaum, 1996). The tools to address technical issues and algorithmic transparency and accountability include algorithmic auditing (Osoba and Welser, 2017), impact assessment (Kraft-Buchman and Arian, 2019); ethical standards for ADMSs design (Floridi *et al.*, 2018, 2020; Morley *et al.*, 2019; Whittlestone *et al.*, 2019), guidelines for the public procurement of ADMSs systems (World Economic Forum, 2019). Among the most substantial regulatory solutions, the

anti-discrimination laws (Xenidis and Senden, 2020; Zuiderveen Borgesius, 2020). and data protection regulations (Malgieri, 2019; Choroszewicz and Mäihäniemi, 2020) were identified as the key components to remedy algorithmic bias and discrimination. The concept of indirect discrimination, acknowledged in the key EU anti-discrimination directives, is particularly useful for addressing algorithmic discrimination (Xenidis and Senden, 2020). The Equal Opportunity bodies were identified as important actors in monitoring and assessing ADMSs; however, appropriate training and in-house technical capacities of these institutions are critical(Crider, 2018). In the area of data protection, the GDPR is the key source of solutions in the EU. The Article 22 of the GDPR directly governs ADMSs andgrants the right to explanation and the rights to remedy when people deal with ADMSs. Although the GDPR provides general guidelines, the articles in the Directive are rather broad, leaving substantial room for national governments to operationalise the principles outlined in the GDPR. (Malgieri, 2019; Choroszewicz and Mäihäniemi, 2020).

2.4. Digital Welfare State

In 2019, the UN Special Rapporteur on Extreme Poverty and Human Rights, Philip Alston, coined the term 'Digital Welfare State' to describe the increasing use of data, predictive analytics and automated decision making in social policy (Alston, 2019). In the digital welfare state, ADMSs is often used to assess eligibility for social benefits; determine the type of support required, allocate resources and support services, and predict tax or welfare fraud. In the public employment services, the most used form of ADMSs is algorithmic profiling of the unemployed. Profiling, either algorithmic or not, refers to a process of differentiation of job seekers based on their likelihood to

find employment into different groups and allocate the resources accordingly (OECD, 2018). According to the OECD, profiling is a useful tool for more efficient service delivery, targeted interventions that are tailored to individual needs. Furthermore, profiling tools are a critical part of the 'activation' employment strategy and allows to focus constrained resources on those that are actively seeking for a job and have the highest chances to succeed (OECD, 2018; Desiere, Langerbucher and Struyven, 2019).

The practice of profiling is not new in the public employment services, thus algorithmic profiling is a technological extension of the established practice. There are three types of profiling – rule-based profiling, casework based profiling, and statistical profiling that are often used in combination. The rule-based profiling classifies job seekers in groups based on administrative eligibility criteria, such as age, educational level, or unemployment duration. The caseworker-based profiling is based on the caseworker's judgement to classify job seekers and is often supported by additional quantitative or qualitative tools. The statistical profiling uses a statistical model to predict labour market disadvantage for individuals, using combined data from administrative records, questionnaires or personal interviews (OECD, 2018; Desiere, Langerbucher and Struyven, 2019). Fully operational statistical profiling systems were adopted in the US and Australia already in the 1990s, and in the early 2000s, these systems have been increasingly discussed in the European Commission's Mutual Learning Programme for Public Employment Services (Weber, 2011; Desiere, Langerbucher and Struyven, 2019). Statistical profiling is increasingly more often used in OECD states, where statistical models range from logistic regressions (e.g., Australia, Austria, Italy, the Netherlands, Sweden, US), to advanced machine learning models (e.g., Denmark, Belgium, New Zealand) (Desiere, Langerbucher and Struyven, 2019). However, algorithmic profiling carries substantial risks.

The major risks associated with algorithmic profiling in the digital welfare state is the risk of systemic exclusion and marginalisation of the most vulnerable, reinforcing and exacerbating the existing patterns of inequality (Niklas, Sztandar and Szymielewicz, 2015; Eubanks, 2018; Alston, 2019). The digital welfare state per se transforms the relationship. between the citizens and the state, thus potentially threatening the social and economic rights of the citizens. The use of algorithmic profiling tools to determine who is eligible to receive support and who is not implies that the state has become less responsible for ensuring an adequate standard of living for all, and the citizens, in turn, have become responsible for proving that they deserve to get certain support or services. The digital welfare state no longer treats citizens as rights-holders but rather as service recipients. Therefore, the social protection services or public employment services, which act increasingly as service providers, are no longer obliged to support everyone, but only those who would provide the best return on investment (Alston, 2020). Such configuration provides sno incentives to provide support to the most marginalised, whose chances in the labour market are low. Such systems also create a feedback loop, whereby those who have the lowest chances in the labour market receive no support, and because they do not receive support, their chances remain low. Furthermore, the use of algorithmic profiling exacerbates the process by making it more difficult to navigate already highly bureaucratic social support services (Peña Gangadharan and Niklas, 2019) and further distancing the service providers from the citizens, as decisions can be made without personal meetings between case-workers and the citizens, potentially excluding large numbers of vulnerable groups from accessing any form of support (Eubanks, 2018; Alston, 2020).

Also, the ADMSs systems in the digital welfare state often place a ‘disproportionate focus on exposing deception and irregularities on the part of welfare applicants’ (Alston, 2020). The

systems are often designed with the goal to "automate, predict, identify, monitor, detect, target and punish" (Alston, 2019, 2020), which is often accompanied by reduced budget, restricted eligibility criteria, and reduced services, strict conditionality mechanisms and sanctions (Eubanks, 2018; Alston, 2020). These goals – policing, surveillance and budget cuts, shape how the AMDSs are designed, what goals are set, and how these systems are used in practice. Because the primary goal of such systems is to predict fraud or reduce spending on support, the default settings are likely to be set in a way that marks as many people as possible as 'risk flag' or groups them into categories that are not eligible for support. In her book "Automating Inequality", V. Eubanks (2018), extensively covers the detrimental impact of such systems, which jeopardised people's well-being by erroneously removing welfare benefits from legally eligible candidates. Eubanks shows that such systems, instead of alleviating poverty, reducing the need for welfare provisions, often are only used to 'manage poverty' and reinforces the existing inequalities. Most importantly, she shows the difficulties that people face when they are exposed to unfair algorithmic decisions, and the lack of practical and meaningful means of redress accessible to those affected (Eubanks, 2018).

3. Research Design

This section describes the research design and the methods used for data collection and analysis. A comparative case study design is used to identify the similarities and differences between the processes that led to the termination of the algorithmic profiling tools in Austria and Poland. Using J. Kingdon's Multiple Streams Framework (MSF) to guide the analysis, this thesis uses academic literature, reports, media articles, and policy documents to reconstruct the process that led to the termination of these systems. This section first introduces comparative case study design and selected case studies, then presents the MSF as an analytical framework, and lastly describes the pattern patching technique for data analysis.

3.1. Comparative Case Study Design

The comparative case study design allows to obtain an in-depth understanding of complex phenomena within its real-life context and understand how the context and the phenomena itself are interconnected (Yin, 2003; Bartlett and Vavrus, 2017). Two cases, algorithmic profiling systems used in public employment services in Austria and Poland, are compared. Each case will be analysed using the pattern matching technique, whereby the observed (empirical) patterns from selected case studies are compared to predicted (theoretical) patterns (Yin, 2003). To ensure the external validity of the findings, the comparison of the cases will be at the centre of the analysis. Building on the multi-axis model of comparative case study proposed by Bartlett and Vavrus (2013), which identifies horizontal, vertical, and transversal axes of comparison, this thesis focuses on the horizontal comparison. The horizontal axis not only compares and contrasts the outcomes of the cases in question but also traces social actors, documents, events, and processes across these

cases, allowing to better understand how policies or phenomena unfold in different conditions and also how they are shaped by them (Bartlett and Vavrus, 2017). This comparative case study seeks to contribute to the theory building that explains the process that led to the termination of automated profiling systems, highlighting the key risks and the associated challenges to address them. The pattern matching technique, combined with comparative case study design, allows to increase the external validity of the findings and develop the theory that could hold for similar automated profiling systems in other contexts.

This thesis focuses on the algorithmic profiling systems adopted in public employment services in Austria (AMAS Algorithm) and Poland (Syriusz). Given a recurrent trend of failures and critiques of various ADMSs systems, the cases were selected from a broader pool of systems that were identified in the literature as failed, which was conceptualised as 1) termination/ suspension; or 2) proven perverse effects on target populations (Vedung, 2012). Using typical cases, this thesis aims to explain the processes that led to the termination of these systems. Because the cases are used to explain the process, not the outcome, the cases were selected based on the following criteria: similarity in outcome, technological design, purpose of use, and institutional and regulatory context to control for confounding factors. First, both cases had similar outcomes – the termination of the automated profiling systems. Second, to control possible differences stemming from technological design, both cases were based on simple algorithmic models, not advanced machine learning models. Third, since the way how technology is applied shapes its impact, the AMDSs used for profiling in public employment services were selected. Lastly, to ensure similar institutional and regulatory context, cases were selected from the members of the Organisation for Economic Cooperation and Development (OECD) and the European Union (EU),

where the General Data Protection Regulation (GDPR), EU's non-discrimination directives, and OECD's best practices for public sector are applied (Weber, 2011; OECD, 2018; Desiere, Langerbucher and Struyven, 2019).

3.2. Conceptual Framework

Kingdon's Multiple Streams Framework (MSF) as a conceptual framework allows to analyse technical, political, social, and economic factors that contributed to the termination of automated profiling systems used in public employment services in Austria (AMAS Algorithm) and Poland (Syriusz). Within the MSF framework, three independent streams – problem, policy, and politics streams – must converge for the 'window of opportunity to open, which leads to policy change. The key actor in this framework is a 'policy entrepreneur' who aims to couple the streams and put the problems, policy solutions onto the political agenda and thus bring about change (Kingdon, 1984).

The *problem stream* focuses on what conditions the actors involved consider requiring attention and action. Not all conditions become policy problems (Kingdon, 1984). What conditions become problems are highly influenced by 'focusing events' and the proactive strategies employed by various actors in framing issues as problems (Kingdon, 1984; Ackrill, Kay and Zahariadis, 2013).

The *policy stream* focuses on the development of policy alternatives in response to a specified problem. J. Kingdon described the development of policies as a 'policy primaeval soup' where various ideas go through 'softening' (Kingdon, 1984). Only those policy alternatives that are feasible and acceptable to the broader policy communities can be coupled with other streams. The

key actors in the policy stream that propose and promote policy alternatives include epistemic communities, such as researchers, consultants, civil society actors, bureaucrats, interests groups (Kingdon, 1984). Some policies may be developed in anticipation of specific problems, while others can take time to develop to adequately respond to an existing problem. The sources of policy alternatives can be endogenous, such as domestic actors, or exogenous, such as international standards, regulations, obligations. (Ackrill, Kay and Zahariadis, 2013).

The *politics stream* focuses on the broader context within which policy is formulated, including interest group activities, party ideologies, and public opinion. These factors influence how problems and policies are framed and the ability and willingness of governments to act. In practice, policymakers operate under the conditions of uncertainty, as well as significant time constraints, and therefore cannot attend to all issues and have to focus on those warranting most attention (Kingdon, 1984; Ackrill, Kay and Zahariadis, 2013).

The MSF approach posits that 'policy entrepreneurs' are the key actors in framing problems and articulating policy solutions, thus coupling the streams together (Kingdon, 1984). However, to better fit the MSF framework to the specific context of the ADMSs, this thesis makes a conceptual distinction between 'problem brokers' and 'policy entrepreneurs.' According to the extensions of the MSF proposed by Knaggard (2015), a 'problem broker' is an actor that 'frames conditions as public problems.' Problem brokers define problems (Knaggård, 2015). The problem broker is different from policy entrepreneurs because they frame conditions as problems without necessarily developing policy alternatives. The policy entrepreneur, in turn, works to present a ready package of problems and solutions to policymakers at the right moment (Kingdon, 1984; Knaggård, 2015). One actor can be active both as a problem broker in the problem stream and

policy entrepreneur in the policy and politics stream. But policy entrepreneurs and problem brokers can also be two different actors (Knaggård, 2015).

The conceptual distinction between 'problem broker' and 'policy entrepreneur' is particularly relevant in the context of the ADMSs and other emerging technologies because of the high epistemic uncertainty and varying public perception of the risks and potential policy solutions (Goyal, Howlett and Taeihagh, 2021). High epistemic uncertainty and public perception impact problem framing, development of policy alternatives, and the degree of politicisation of the issue. For example, the new technologies challenge the existing economic, social, or environmental conditions. Therefore, they can be seen as problems once attention is drawn to their adverse effect on society (Goyal, Howlett and Taeihagh, 2021). For example, algorithmic biases embedded in these systems can have discriminatory outcomes. Second, these systems are likely to transform the power relationship between the government and the citizens, between the social service provider and the recipient, creating potentially imbalanced, untransparent and unaccountable processes (Kuziemski and Misuraca, 2020). Furthermore, epistemic uncertainty influences the pace of development of institutional regulatory frameworks that govern the deployment and use of new technologies. In the case of the ADMSs and other emerging technologies, the knowledge of how to address potential risks develops slower than technological uptake; therefore, there is often a lack of feasible policy solutions to potential problems (Goyal and Howlett, 2019). The public perception often influences how receptive different actors can be to certain problem framing or certain policy alternatives and the politics surrounding the use of certain technologies. In the political stream, the main challenge is to accommodate different interest groups involved in innovation and technology policy (Goyal and Howlett, 2019; Goyal, Howlett and Taeihagh, 2021).

The ADMSs, however, are often discussed as part of discourse and policies on economic innovation, not as part of policies of equality, social justice, or human rights, making the political dimension less receptive to the issues, such as algorithmic bias (Mittelstadt *et al.*, 2016).

This thesis will look at problems, policy, and politics streams to understand how and why ADMSs for profiling failed in Austria and Poland. Although initially the algorithmic profiling in selected cases was touted as an innovative, efficiency-enhancing, error-reducing tool inextricable from the modernisation of the public sector, eventually they had become objects of scrutiny, critique, investigations, which ultimately led to their termination in Austria and Poland. The MSF framework allows us to analyse the multitude of factors that could have contributed to the failure of these programs. In this thesis, both AMAS and Syriusz are analysed not merely as automated decision-making technology but also as systems involving technological, social, economic, and political aspects. Thus, using MSF allows capturing the multiple dimensions that impact the use of ADMSs in public employment services.

3.3. Data Collection and Analysis

This thesis uses the pattern matching technique and comparative case study design to compare and contrast the patterns identified in the selected cases (empirical findings) with the expected patterns (theoretical patterns), using the MSF as a guiding analytical tool. This thesis utilises various primary and secondary sources to obtain relevant information. Primary sources used in this thesis include media articles, blog posts, policy documents, and reports. Secondary sources include academic literature and reports on ADMSs in general and specifically on the selected cases of automated profiling in Austria and Poland. Some sources serve both as primary

and secondary sources. For example, academic articles written by Austrian researchers were used to obtain synthesised empirical data about the AMAS algorithm, but also provided first-hand insights into whether and how researchers frame the issue of algorithmic bias.

The data collection and analysis process can be divided into four steps. The first step is to overview the key literature on the digital welfare state, ADMSs for profiling, and algorithmic bias and develop a hypothesis about the process that led to the termination of the algorithmic profiling tools, using the MSF framework. The second step involves the collection of empirical data on the selected case studies using primary and secondary sources and to trace the process that led to the termination of the programs. The third step is the thematic analysis of the data collected at the second step using Nvivo software to classify the collected information into problems, policy, and politics streams for each case. The fourth step involves comparing the findings of each case by problem, policy, and politics streams of the MSF. Lastly, the findings for each stream and the identified relations between the streams are compared to the expected process identified at the first step.

The literature review and the MSF conceptual framework inform the development of the expected patterns of the process that led to the termination of the algorithmic profiling tools. The review of literature on digital welfare state and the associated risks (Eubanks, 2018; Alston, 2019) showed that no single variable can be distinguished as a single factor shaping the impact and workings of ADMSs in the public social protection services. Therefore, to capture the importance of multiple factors and how their interactions shape the fate of automated profiling systems, the key factors – such as models, data, regulations, transparency and accountability methods – will be grouped into three categories based on the MSF framework as factors relating to problem, policy,

and politics. Therefore, the expected patterns derived from the literature review are framed as the interaction between the problem, policy, and politics streams.

Building on the literature review and the MSF framework, the following theoretical mechanism explaining the failure of algorithmic profiling is derived. The mechanism establishes that the systems failed because the window of opportunity did not open. The coupling of streams would have resulted in modification of the current regulatory framework in data protection and equal treatment, adoption of additional risk mitigation measures, such as independent auditing, which would have enabled better governance, regulation, and use of the algorithmic tools. Thus, the problem of algorithmic bias would have been addressed by policy tools, leading to new kinds of institutional and regulatory framework. In contrast, the termination of algorithmic profiling tools represents the *status quo*, the lack of change in broader socio-political, institutional, and regulatory domains.

4. Case Studies

This section introduces the case studies, provides background overview and briefly describes the technical design and practical application of the algorithmic profiling tools used in public employment services in Austria (2016-2020) and Poland (2014-2019). The flaws in technical design and the questionable application of these tools raised concerns over algorithmic bias and discrimination, leading to their termination.

4.1. Context

In Austria and Poland, the authorities introduced algorithmic profiling as means to reduce public spending, increase the efficiency of resource allocation, and provide more targeted services based on individual needs (Niklas, Sztandar and Szymielewicz, 2015; Allhutter, Cech, *et al.*, 2020). In 2012, the Polish Ministry of Labour and Social Policy (MLSP) initiated a reform of labour offices in Poland (PUP - Powiatowe Urzędy Privacy), which also included the introduction of algorithmic profiling. The MLSP contracted a private company Sygnity to develop an automated profiling system (Syriusz), which was rolled out in 2014. The goal of these reforms was to reduce public spending and modernise the social protection services in line with recommendations and best international practices, such as activation strategy to increase employment rate (Niklas, Sztandar and Szymielewicz, 2015). In Austria, algorithmic profiling was introduced at the initiative of the Austrian Employment Agency (Arbeitsmarktamt– AMS). Unlike in Poland, the introduction of algorithmic profiling was not part of broader reforms, as the AMS has been transformed into a semi-autonomous service provider, actively utilising activation strategy since the mid 1990s (Allhutter, Cech, *et al.*, 2020). In 2016, the AMS started a program

to assess the chances of different groups of people in the labour market. In 2019, the AMS announced the launch of the algorithmic profiling tool, dubbed AMAS algorithm. The tool was developed by a private contractor Synthesis Forschung and was implemented by the AMS (Kayser-Bril, 2019).

Both algorithmic profiling tools were framed by the respective public employment services agencies, AMS in Austria and PUP in Poland, as advisory tools designed to support caseworkers. They highlighted the right of the caseworkers to override the algorithmic decision as a means to ensure accountability of algorithmic profiling; however, the later investigations into the systems revealed little evidence that the algorithmic classifications were ever questioned, challenged or overridden (Niklas, Sztandar and Szymielewicz, 2015; Kayser-Bril, 2019; Allhutter, Cech, *et al.*, 2020; Szigetvari, 2020). Therefore, both the Syriusz and AMAS algorithms received substantial critiques internally and within the broader society due to concerns over privacy issues, and the risks of discrimination and further exclusion of the most vulnerable groups (Niklas, Sztandar and Szymielewicz, 2015; Allhutter, Cech, *et al.*, 2020). In 2019, the Human Rights Commissioner in the Polish government filed a complaint to the Constitutional Court, which ruled the system illegal. The profiling tool was disbanded the same year due to the lack of legal base for its operations (Tarkowski, 2019). In August 2020, the Austrian Data Protection Authority ruled that the AMS algorithm cannot be used due to data protection violations (Szigetvari, 2020).

4.2. Design and Application

Both AMAS algorithm and Syriusz categorised the unemployed into three groups based on their predicted chances to find employment, using a wide range of data available to public

employment services. In Austria, AMAS algorithm used administrative data from different sources, while in Poland, questionnaire information complemented the administrative data. The algorithmic classifications determined what resources and support was available to each group. The types of support ranged from job placement, training, apprenticeship, activation allowance, or no support at all (Lopez, 2019; Kuziemski and Misuraca, 2020). People with the highest chances to find a job were classified into the first group (Group 1 in Syriusz and Group A in AMAS). Most of the people within this group were educated, skilled, actively looking for work or having sufficient professional qualifications, thus they're entitled to some support from the PES. The second group (Group II in Syriusz and Group B in AMAS), which was predicted to have medium chances to find a job in the medium term, was identified as the priority group and most of the resources of PES in Austria and Poland were dedicated to this group. In contrast, the third group (Group III and Group C) was predicted to have the lowest chances in the labour market, thus almost no support was offered by the public employment services to this group. In Austria, this group was directed to alternative support, but in Poland, these people were deprived of any state assistance (Lopez, 2019; Kuziemski and Misuraca, 2020). The Ministries of Labour and Social Affairs in these countries justified such profiling system using the 'activation paradigm' which posits that to increase the effectiveness of services, they should focus on those who could benefit the most form that assistance, instead of spending resources on those 'who are not interested or seeking employment (Lopez, 2019, Kuziemski and Misuraca, 2020, JRC, 2015).

Table 1: *Categorisation of Job-seekers and Available Support*

	Group 1 / Group A	Group 2 / Group B	Group 3 / Group C
Prospects in the labour market	High chances	Medium chances	Low chances
Support available	Some	All	None (Syriusz) / Alternatives (AMS)

Source 1*: LOPEZ, 2019 AND KUZIEMSKI ET AL, 2020*

The algorithmic categorisation of the unemployed and how it informed the allocation of resources has emerged as the key concern in the debates surrounding the AMAS algorithm and Syriusz. The systems were designed to categorise people based on their chances, thus those with the least opportunities (Group C and Group III) were further marginalised as the support resources were diverted from them, reinforcing structural and intersectional inequalities (Lopez, 2019). Such use of algorithmic profiling is concerning due to the following reasons. First, there are chances of misclassification of individuals due to algorithmic bias or error, and the systems are not transparent about the decision-making process nor provide mechanisms for redress (Pasquale and Citron, 2014; Barocas and Selbst, 2016). Second, depending on the variables included, certain groups of people, especially those already marginalised, would be systemically categorised into group three with no chance to change their status and improve their situation (Eubanks, 2018; Peña Gangadharan and Niklas, 2019). Thus, diverting resources could lead to further marginalisation of already socioeconomically disadvantaged groups (Lopez, 2019; Kuziemski and Misuraca, 2020). Although the systems were introduced as an attempt to modernise public employment service, increase the efficiency of labour market activation policies and reduce costs, the issues of algorithmic bias and discrimination have become increasingly salient in the discussions

surrounding these systems. The following sections look at problem, policy, and politics streams to identify key actors, events, and developments that led to the increasing awareness of the problem and the subsequent termination of these algorithmic profiling systems.

5. Analysis and Discussion

This section analyses the selected case studies using the MSF framework. The first three subsections of this chapter are thematically divided into the three streams – problem, policy, and politics stream – identified in the MSF framework. Each stream discusses and compares relevant actors, events, and processes related to the AMAS algorithm in Austria and Syriusz algorithm in Poland. The problem section shows how different actors contributed to the emergence, definition, and framing of the issue of 'algorithmic bias' as a problem warranting attention and requiring policy solutions. The policy stream shows the key actors and ideas that shaped the discussion on the possible solutions to the identified problem. The politics stream discusses the lack of attention paid to the issue from policymakers, as well as resistance to the criticism demonstrated by public employment services. The last subsection of this chapter discusses how the empirical findings relate to the expected patterns identified through literature review.

5.1. Problem Stream

This sub-section discusses the role of problem brokers in identifying and framing a problem of algorithmic bias and discrimination as a policy problem. It shows that civil society (Austria, Poland) and academic research institutions (Austria) played a critical role in drawing attention to the risks of algorithmic profiling and framing the topic as a policy problem.

5.1.1. Civil society as problem brokers

Civil society organisations, concerned with privacy and human rights, emerged as the key problem brokers that strategically engaged in the identification and definition of the problem by obtaining evidence, conducting investigations, and publicising the issue. In Poland, the

Panoptykon Foundation, an advocacy and research organisation, emerged as a main problem broker. The Foundation obtained significant evidence by requesting to make the models and data used in the Syriusz algorithm available and accessible to the researchers and the public. For example, in 2014, the Foundation requested the Polish Ministry of Labour and Social Affairs (MLSP) to make available the questionnaire used in the course of profiling. Despite the initial reluctance by MLSP to reveal the information, the questionnaire was made available to the Panoptykon Foundation and shared on their website (Niklas & Szumańska, 2014). Furthermore, the questionnaire, which became public as a result of the Panoptykon Foundation's request, was published in an online periodical *Dziennik Internautów* (Niklas et al 2015). Thus, the foundation contributed to the formulation of the problem by generating evidence and sources for further analysis that were not public ex ante.

The digital rights and privacy NGO epicenter.works played an important role in generating further information about the algorithm and raising awareness about the issue in Austria. Although some parts of the AMAS model were made public from the beginning, the information was limited. The NGO epicenter.works submitted several requests to the AMS and the Synthesis Forschung to release additional technical information about the algorithms used (epicenter.works, 2019). Their request resulted in a detailed paper published by Synthesis Forschung in May 2019, revealing technical aspects of the models and discussing social compatibility of the algorithms (Allhutter et al., 2020; Holl et al., 2019). The epicenter.works, along with other civil society organisations, initiated the petition against the AMS algorithm, requiring it to rescind it (*Stoppt den AMS-Algorithmus!*, no date). Furthermore, it has published a number of articles on AMS-algorithm, data protection, and human rights, thus contributing to the raising salience and 'de-technocratisation' of the issue (epicenter.works, 2019). They also compared the AMAS algorithm with the Syriusz

in Poland, using it as a 'cautionary tale' example to substantiate their advocacy efforts (epicenter.works, 2019).

5.1.2. Epistemic communities as problem brokers in Austria

Academic research institutions, such as Austrian Academy of Sciences, Technical University Vienna, University of Vienna, as well as a non-governmental organisation called *epicenter.works* emerged as the key problem brokers who brought the issue of algorithmic bias, privacy, and lack of transparency in AMAS as problems requiring action. The research produced in the academic institutions provided scientific, expert evidence to substantiate the framing of the problem (Allhutter, 2019; Lopez, 2019; Allhutter, Cech, *et al.*, 2020; Allhutter, Mager, *et al.*, 2020). Although the development of the AMAS-Algorithm started in 2016, the topic entered the public discourse and gained more attention in 2018, after the publication of the methods paper by Synthesis Forschung, which published one of the models used in the AMAS-Algorithm (Holl, 2018). The publication provided opportunities for researchers and scientists in the leading research institutions in Austria to further analyse the algorithm and identify potential risks (Lopez, 2019; Allhutter, Mager, *et al.*, 2020, 2020). For example, a report on how AMS algorithm reinforces intersectional inequality, published by Paola Lopez at the University of Vienna, used the publicly available models to provide empirical evidence for how the AMS model disadvantage women and other discriminated groups (Lopez, 2019). Extensive research has been also carried out byjoined research group involving researchers from the Austrian Academy of Sciences, Technical University Vienna, University of Vienna (Allhutter, Cech, *et al.*, 2020; Allhutter, Mager, *et al.*, 2020). The publication of the models also led to increased attention to the AMS-algorithm in the media: since October 2018, when the some models were released, 33 articles on 'future zone.at'

and 75 on ‘der Standard’ were published, including interviews with AMS staff, as well as with leading researchers investigating the algorithm (futurezone.at, 2020; derstandard.at, 2020). No similar research papers, published between 2014-2019, were found in Poland The issue of algorithmic profiling with Syriusz and other ADMSs applications in Poland have been discussed briefly in foreign-based publications after the termination of Syriusz. In Poland, the issue also received little attention from the media.

5.1.3. Framing the problem

The problem brokers in Austria and Poland played a key role in changing the narrative surrounding algorithmic profiling. In the official discourse, the responsible institutions framed these tools as innovative, objective instruments to increase efficiency and improve service delivery (Niklas, Sztandar and Szymielewicz, 2015; Kuziemski and Misuraca, 2020). The problem brokers framed the problem by showing that these systems are linked to broader questions of equality, inclusion, human rights, and privacy. The problem brokers played a role in showing that algorithmic profiling can threaten these principles and rights. The obtained evidence and additional investigations conducted by the Panoptykon Foundation led to the publication of the seminal report on the inner workings of Syriusz in 2015, titled “Profiling the unemployed in Poland: social and political implications of algorithmic decision making” (Niklas, Sztandar and Szymielewicz, 2015). The report provided a detailed overview of the inner workings of the Syriusz system, using empirical evidence obtained from data and models from the public employment offices and interviews with caseworkers to show that the system has detrimental effects on those who are classified in the third group due to their low chances to succeed and receive no support. The report highlighted that automated profiling perpetuates the stigmatisation of poverty, making it more difficult for already vulnerable and marginalised groups to navigate the already complicated

welfare processes, thus violating social human rights (Peña Gangadharan and Niklas, 2019). Furthermore, in addition to the risks of systemic exclusion, the systems also lacked transparency and was marked by substantial risks to privacy and non-discrimination principles. (Niklas, Sztandar and Szymielewicz, 2015).

In Austria, the AMS algorithm was primarily criticised for being biased and discriminatory, perpetuating systemic and intersectional inequalities (Lopez, 2019), lacking transparency (Allhutter, Cech, *et al.*, 2020), raising concerns over privacy (Epicenter.works, 2019). One of the main issues with the AMS algorithm, highlighted by researchers, was discriminatory and exclusionary risks of the model. The problem was the explicit inclusion of variables such as gender, age, disability, citizenship, and care responsibilities (the latter only for women) into the logistic regression model used to categorise job-seekers (Holl, 2018). These variables largely overlap with characteristics that are considered protected under the discrimination law due to risks of discrimination on the basis of sex, gender, age, disability, etc. Researchers analysing the AMS model, published in the methods paper in 2018, found that the model systematically reduced labour market prospects for women, as well as people with intersecting vulnerabilities, e.g. old age, non-EU citizenship, disability, (Lopez, 2019; Allhutter, Cech, *et al.*, 2020).

Therefore, the Panoptykon Foundation in Poland and epicenter.works and research institutions in Austria played a critical role as problem brokers in identifying and framing the problem. They shifted a discourse of algorithmic profiling from a depoliticised, technocratic issue to the problem that is closely related to human rights, social justice, and privacy.

5.2. Policy Stream

Policy stream analyses whether and what policy solutions have emerged in response to the identified problem. This stream looks at the role of policy entrepreneurs in developing policy solutions. In selected cases, multiple actors, from civil society to researchers to data protection authorities, equal treatment bodies, ministries, to public employment service representatives, engaged in debates surrounding the issue; however, no policy entrepreneur emerged that would be able to put the circulating ideas and critiques into concrete policy proposals.

5.2.1. Policy ideas but not solutions

Most of the research on policy solutions to the algorithmic bias has been done in the US, much less knowledge about possible solutions is available in Austria and Poland. Generally, commonly researched policy solutions can be divided into the following categories: technical, transparency, accountability and monitoring, right to remedy and regulations (Mittelstadt *et al.*, 2016). The Panoptykon Foundation in Poland, Academy of Sciences and affiliated researchers in Austria were particularly active in discussing potential policy ideas. Since 2018, the GDPR has become an important endogenous source of regulatory solutions; however, despite the discussions between various institutions, neither regulatory nor alternative solutions were translated into practice.

In Poland, the Panoptykon Foundation outlined an extensive list of wide-ranging policy solutions. For example, the Foundation highlighted the importance of regulating algorithmic profiling through legal acts rather than through internal guidelines or internal documents. It also reiterated the importance of transparency and accountability by 1) ensuring the right to obtain detailed information about all aspects of the process by affected parties and relevant stakeholders;

2) obliging the public institutions using algorithmic profiling to publish statistical data on the structure of generated categories of people and the distribution of resources among these groups. It also suggested impact assessments and independent auditing of algorithmic systems, taking into account human rights-related risks. It encouraged to keep humans in the loop of ADMSs and provide adequate resources and training for caseworkers. To ensure the right to remedy, the governments were suggested to shift the burden of proof from individuals to the users of algorithmic profiling tools (Niklas, Sztandar and Szymielewicz, 2015).

In Austria, in the policy brief published by the Austrian Academy of Sciences (2019), D. Allhutter outlined very similar proposals. Allhutter emphasised the need for 'comprehensive transparency', which requires not only making public the inner workings of the algorithm and statistics on its implications, but also ensuring that the affected individuals and important stakeholders can meaningfully interpret the information, use it and participate in the process when needed. Furthermore, legal and regulatory boundaries should define under what conditions and with what justification the automated classification could be changed, thus strengthening the legal protections against discrimination. Specific training and new skills are needed among the caseworkers and general public employment service workers to understand and assess the limitations of the algorithmic methods and use their discretion to make judgements. Allhutter also emphasised that independent monitoring, auditing and evaluation is necessary, which could be conducted by external actors not affiliated with the institution using the tool. Furthermore, non-technical solutions should not be discarded, e.g., increasing the caseworker-jobseeker ratio could improve decision-making and reduce the reliance on algorithmic systems as primary and sole decision-makers (Allhutter, 2019).

In the area of regulations as solutions to algorithmic bias, the General Data Protection Regulation (GDPR) is an important element. The GDPR provides an exogenous source of potential policy solutions to the cases in question, for it directly includes provisions on automated decision making under the Article 22, and specific provisions specifically on algorithmic bias in recital 71, which grants the right to explanation (Edwards and Veale, 2017). However, while Austria and Poland are obliged to comply with the GDOR, the regulation only sets broad, general standards and leaves a lot of room for member state discretion on how these provisions are transposed into national law and vary by country. Furthermore, the recitals, including recital 71 are non-binding thus are likely not to be implemented (Edwards and Veale, 2017). Therefore, to effectively use the GDPR as a source for regulatory solution to the issue of algorithmic bias, the policy entrepreneurship is critical. Policy entrepreneurs could revisit the national implementation of the GDPR, as well as check the compatibility of the algorithmic tools with this regulation.

5.2.2. Lack of policy entrepreneurship

The lack of attention to the policy ideas identified by the experts can be explained by the lack of policy entrepreneurship in the policy stream. Despite the discussions, critiques, and investigations conducted by the Data Protection Authorities (DPAs) and Equal Opportunities Bodies in each country, the policy ideas have not been translated to policy solutions.

In Poland, the Inspector General for Personal Data Protection (IGPDP) exchanged 14 official letters with the Polish Ministry of Labour and Social Affairs (MLSP), expressing reservations over the extent to which Syriusz comply with data protection regulations, especially the right to privacy and personal data protection. The key concern was the lack of regulated procedure on challenging the assigned profile. The Polish Equal Opportunities Office also submitted two statements on profiling to the MLSP. The statements expressed concerns over the

transparency of the profiling processes as well as the lack of clarity in legal provisions permitting algorithmic profiling (Niklas, Sztandar and Szymielewicz, 2015). In Austria, the DPA and equal opportunities body also expressed concerns and launched their own investigations into the AMAS algorithm. The most concrete, yet of limited impact, policy development was the establishment of a 'Sounding Board' by the Austrian Ombudsperson for Equal Opportunities, with a specific mandate to evaluate the discriminatory potential of ADMSs systems, including AMAS algorithm, which conducted an investigation into the algorithm in March 2019. (Allhutter, Cech, *et al.*, 2020; Bergmann and Pretterhofer, 2020). The second investigation conducted by the Austrian Ombudsperson for Equal Opportunities reveals that the way data was collected, and used in the model, was largely discriminatory. However, no concrete proposals were proposed (Allhutter, Cech, *et al.*, 2020).

The key discussions in Austria were led by the Equal Opportunities Board, focusing on the disparate impact of the algorithm. In Poland, the discussion among these different institutions was led by the Data Protection Authority. However, the discussions in Poland took place in 2014-2015, before the GDPR came into force, but became relevant for Poland after 2018. Nonetheless, the policy ideas on how to address algorithmic bias, endogenous or exogenous, have been 'circulating' among different constituencies and stakeholders. The endogenous solutions were discussed in policy briefs, reports among experts. The key source of exogenous solution was the legal provisions of the GDPR that could be widely interpreted and adapted to national contexts. Neither policy ideas or tools were developed into concrete policy actions due to the lack of policy entrepreneurship, or an actor able to couple the problem and policy streams and place them onto the political agenda.

5.3. Politics Stream

This stream focuses on the broader political context within which problems emerge and policy solutions develop. The politics stream captures a variety of conditions, events, such as political climate, dominant ideologies, public opinion, interest group activity which influence the extent to which policymakers are receptive to problems and policies and willing and able to take action. The 'policy entrepreneurs' in politics stream are the key actors that couple problems and policy solutions together and place it on the legislative agenda. This section shows that the issue of algorithmic discrimination has not been addressed due to lack of policy entrepreneurship at the political level. The lack of policy entrepreneurship can be attributed to the lack of attention to the issue among the policymakers and the reluctance to acknowledge the problem by the instrumental institutions, such as public employment service agencies and relevant ministries.

5.3.1 Lack of Attention

Algorithmic profiling was introduced as part of a broader political agenda and therefore received little attention from policymakers. In Poland, algorithmic profiling was introduced as part of larger reforms of the public sector in 2014, that sought to modernise, increase efficiency and effectiveness of various social protection services, in line with international standards. The reforms sought to make the social protection system more service-oriented, individualised, and based on activation strategy in order to allocate the resources more efficiently (Weber, 2011; Niklas, Sztandar and Szymielewicz, 2015; OECD, 2018). In Austria, the launch of the AMAS algorithm coincided with the adoption of the policy strategy called the ‘Labour Market Policy Targets’, under the leadership of the Minister for Labour and Social Affairs, Hartinger-Klein, in 2019 (Allhutter, Cech, *et al.*, 2020). The strategy directly mentioned algorithmic tools as a means to increase

efficiency and improve social protection services; however, it also reduced the funding allocated the public employment services, increasing the resource constraints of the AMS, and thus reiterating the role of algorithmic profiling as a cost-reduction tool (Wimmer, 2019).

Because algorithmic profiling was a very small aspect of the broader political agendas, it received relatively little attention among policy makers. In Poland, following the investigation by the Panoptykon Foundation and critical enquiries into the system by the Data Protection Authority and ten parliamentary interpellations over the lack of possibilities to change the classification between 2014-2015, there was little political debate on the topic until its termination in 2019 (Niklas, Sztandar and Szymielewicz, 2015). In Austria, a parliamentary inquiry into the AMAS algorithm was launched by the Austrian Parliamentarian Heinisch-Hosek (Social Democratic Party) in 2018, followed by two investigations launched by the Equal Opportunites Ombudsoffice in 2019. The system was terminated shortly after the investigations by the Data Protection Authority, with little or no attention from the policymakers to the issue (Allhutter, Cech, *et al.*, 2020).

5.3.2. Denying the Problem

Introducing these systems as part of a broader political agenda has two implications. First, it makes it difficult to separate these tools from other developments; therefore, challenging the algorithm might mean challenging the entire policy agenda, e.g, activation strategy, austerity politics, modernisation and innovation plans, etc. Second, if the broader political agenda was widely supported – as it was in Austria and Poland, where the parties in power were largely in favour of reforms, since they passed – the political climate, by extension, is likely to be less receptive and even reluctant to acknowledge the problem, such as algorithmic bias. The reluctance

to acknowledge the problem by instrumental institutions were substantial political obstacles for addressing the issues of algorithmic bias and discrimination.

The instrumental institutions, such as the Public Employment Services and the Ministries of Labour and Social Affairs, whose mandates, expertise, and input would be critical for developing and implementing various technical and institutional solutions, such as algorithmic auditing. In Poland, the Ministry of Labour and Social Affairs (MLSP) responded to investigations into the Syriusz Algorithm as well as the requests for transparency of the models as irrelevant and stated that the actors questioning the system 'lack competence on the matter' (Niklas et al, 2015). In Austria, the public employment agency AMS also refused to recognise the problem. For example, the head of the Austrian public employment services (AMS), J. Kopf, has released multiple statements for the media (Szigetvari, 2018), and his personal blog (Kopf, 2019) expressing support for the AMAS algorithm, justifying the inclusion of protected characteristics, such as gender, ethnic background, and 'care burden' as variables, by saying that such design of the model 'reflects the reality of the labour market' and refusing to seriously consider the problems associated with AMAS algorithm (Szigetvari, 2018). These actors could have emerged as policy entrepreneurs due to their institutional positioning and expertise on the issue, but their reluctance, evident form dismissive comments and responses to the demands of transparency, and the lack of proactive reaction to the problem, meant that some of the potential solutions at the institution-level, such as independent algorithmic auditing, training for caseworkers and other employees of the public employment services, were lost.

5.4 Coupling the streams

This subsection explains why the problem, policy, and politics streams did not couple to open the window of opportunity for the change in the governance of automated profiling. According to the theoretical mechanism developed based on the literature review, the expected process of policy failure is as follows. The coupling of streams would have resulted in modification of the current regulatory framework in data protection and equal treatment; in the adoption of additional risk mitigation measures, such as independent auditing, impact assessments, and mechanisms for transparency and accountability. These solutions would have created a new institutional framework for better governance, regulation and use of the ADMSs that is able to capture and address the potential risks, such as algorithmic bias. In contrast, the termination of algorithmic profiling tools represents the *status quo.* The analysis of the selected cases in Austria and Poland fits this explanatory pattern. Due to the lack of political attention to the problem and the lack of policy entrepreneurship, the problem and potential policy solutions were never placed on the political agenda. Therefore, instead of leading to a new regulatory and institutional framework for the governance of ADMSs for profiling, the systems were cancelled in 2019. Several factors explain the lack of policy entrepreneurship and lack of political attention.

First, the lack of attention to and often an overt reluctance to recognise algorithmic bias among the policymakers and key institutions were the key factors that contributed to the lack of policy entrepreneurship at the policy and political levels. Despite some critical inquires and debates, no single or a group of politicians undertook an initiative to bring the problem of algorithmic bias and discrimination onto the political agenda and initiate any legislative action. Because algorithmic profiling system fit well with the dominant ideological paradigms dominant

among the policymakers, such as austerity politics, activation paradigm, and technosolutionist logic, the problem of algorithmic bias and discrimination was not considered important. For example, the algorithmic profiling tools used to allocate support based on certain categories to reduce costs reflect the austerity politics, embedded in the public sector reforms of 2014 in Poland and Labour Market Agenda of 2018 in Austria. Also, the specific focus to allocate the most resources to those that are most likely to succeed reflect the dominant 'activation' paradigm in contemporary social welfare management that has been discussed among the OECD countries as the best practice since the 1990s. The reluctance to recognise potential biases in the systems, expressed by AMAs head J. Kopf in Austria and the MLSP in Poland, reflect the technosolutionist logic that technology is an objective, neutral and quick fix to all problems.

Second, given the context, the framing of the problem along the lines of human rights, social justice, inclusion, non-discrimination, and privacy was not effective in garnering the attention of policy makers. Although the problem brokers, such as the Panoptykon Foundation in Poland and epicenter.works and research organisations in Austria, were very active in identifying the problem, conducting research, and obtaining evidence to support their claims, their evidence was not sufficient to convince the policymakers of the importance of the problem. Furthermore, by requesting for more transparency, these policy brokers not obtained the evidence to substantiate their claims, but also opened opportunities to create more transparency and accountability in these systems. However, these opportunities were not utilised outside these expert and activist circles identified as policy brokers. This shows that despite the existence of the problem and substantial evidence to prove it, the problem has not gained attention from the policy makers because the framing did not match the dominant ideologies, interests and views of the key institutions and policy-makers.

Third, although it is commonly believed that AMDSs often fail due to the lack of knowledge about and the policy solutions to the associated risks and the lack of policy alternatives. This analysis revealed that the knowledge about the risks associated with algorithmic profiling and ideas about how it could be solved from technical, governance, and regulatory perspectives were circulated among the expert actors. First, the problem brokers, such as civil society organisations and research organisations, have discussed and publicised information about potential policy responses to the problem. Second, since 2018, the GDPR, which Austria and Poland are obliged to comply to, include binding and non-binding provisions on automated decision-making (Article 22) and specifically on algorithmic profiling (recital 71) which offered a source of possible regulatory solutions to the issue of algorithmic discrimination. However, these ideas have not been translated into practical policy proposals that were placed on the legislative agenda. Therefore, the key obstacle for adequately addressing the algorithmic bias and discrimination was not the lack of potential solutions, but the lack of policy and political entrepreneurs who would take an initiative to put the problem and the potential policy solutions on the legislative agenda. Furthermore, policy entrepreneurship could have played a role at two levels – at the level of public employment service agencies and at the higher, national-level decision-making level. At the agency level, policy entrepreneurship would have meant implementing technical and administrative resolutions, such as technical changes in the model, training of caseworkers, etc. At the national decision-making level, the regulatory solutions could have been developed, which in this case could have been revisions of the national adaptations of the GDPR regulation. However, no policy entrepreneurship emerged at either level, which can also be explained by the dominant paradigms and ideologies held by those leading the instrumental institutions and the key players in the political process.

6. Conclusion

The issue of algorithmic bias, identified by civil society and academia, was not sufficient to garners more attention to the risks of automated profiling in the public employment services and transform the institutional and regulatory framework on how such systems are governed, managed, and used across various domains. The comparative analysis of the processes that led to the termination of algorithmic profiling in Austria and Poland reveal that the lack of political attention to the issue and the lack of policy entrepreneurship were the main factors that led to the termination of these systems.

The lack of political attention to the problem can be explained by the fact that algorithmic tools were adopted as part of broader, largely supported political agendas. The algorithmic tools also fit the goals and logic of these political agendas, based on austerity measures, activation paradigm, and technosolutionist logic. The problem of algorithmic bias and discrimination, identified by the civil society and researchers, was framed as issues of human rights, equality, and social justice, which in a way contradicted the dominant views espoused by the key institutions, such as the public employment services, and key political actors. Therefore, despite the existence of possible technical, regulatory, and governance solutions, the mismatch between how the problem was framed and the priority areas of policy makers, prevented the emergence of policy entrepreneurship that could have placed the problems and policies onto legislative agenda.

This comparative analysis focuses only on two cases and more research could be done to test the mechanisms through which ADMSs systems succeed or fail. Further research could focus on the risks of ADMSs and the possible technological and policy solutions in ADMSs based on advanced models, such as machine learning, which are increasingly adopted even in the private

sector. The role of data protection regulations, especially the GDPR or equivalent legislation in other regions, its national applications, and its relevance for ADMSs and machine learning could be also further investigated. Lastly, while the issue of risks associated with the ADMSs have been mostly focused on the US, and there is a growing interest in the topic in OECD or European countries, the issues associated with the ADMSs and the digital welfare state in emerging economies have not been studied extensively.

7. Bibliography

Ackrill, R., Kay, A. and Zahariadis, N. (2013) 'Ambiguity, Multiple Streams, and EU Policy'.

AlgorithmWatch (2019) *Automating Society: Taking Stock of Automated Decision-Making in the EU*. Berlin. Available at: https://algorithmwatch.org/en/automating-society-2019/.

AlgorithmWatch (2020) *Automating Society 2020*. Berlin: AlgorithmWatch gGmbH and Bertelsmann Stiftung. Available at: https://automatingsociety.algorithmwatch.org.

Allhutter, D. (2019) 'AMS Algorithm on Trial. ITA Dosier No.43'. Austrian Academy of Sciences.

Allhutter, D., Cech, F., *et al.* (2020) 'Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective', *Frontiers in Big Data*, 3, p. 5. DOI: 10.3389/fdata.2020.00005.

Allhutter, D., Mager, A., *et al.* (2020) *DER AMS-ALGORITHMUS Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*.

Alston, P. (2019) *Report of the Special Rapporteur on Extreme Poverty and Human Rights*. A/74/48037.

Alston, P. (2020) *What the "digital welfare state" really means for human rights*, *OpenGlobalRights*. Available at: https://www.openglobalrights.org/digital-welfare-state-and-what-it-means-for-human-rights/?lang=English (Accessed: 15 July 2021).

Andersen, L. (2018) *Human Rights In The Age Of Artificial Intelligence*. Access Now.

Barocas, S. and Selbst, A. D. (2016) 'Big Data's Disparate Impact', *California Law Review*, 104(3), p. 671. DOI: 10.15779/Z38BG31.

Bartlett, L. and Vavrus, F. (2017) 'Comparative Case Studies: An Innovative Approach', *Nordic Journal of Comparative and International Education (NJCIE)*, 1(1). DOI: 10.7577/njcie.1929.

Bergmann, N. and Pretterhofer, N. (2020) 'Artificial Intelligence and Genderbiases in Recruitment and Selection Processes – a topic not much discussed yet in Austria'.

Brayne, S. (2020) *Predict and Surveil : Data, Discretion, and the Future of Policing*. New York: Oxford University Press. DOI: 10.1093/oso/9780190684099.001.0001.

Broussard, M. (2019) *ARTIFICIAL UNINTELLIGENCE: How Computers Misunderstand the World*. The MIT Press.

Buolamwini, J. and Gebru, T. (2018) 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'.

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

Choroszewicz, M. and Mäihäniemi, B. (2020) 'Developing a Digital Welfare State: Data Protection and the Use of Automated Decision-Making in the Public Sector across Six EU Countries', *Global Perspectives*, 1(1), p. 12910. DOI: 10.1525/gp.2020.12910.

Crider, C. (2018) *Mapping RegulatoryProposals for Artificial Intelligence in Europe*. Access Now.

Cukier, K. N. and Mayer-Schoenberger, V. (2013) 'The Rise of Big Data How It's Changing the Way We Think About the World', *Foreign Affairs*.

Desiere, S., Langerbucher, K. and Struyven, L. (2019) *Statistical profiling in public employment services: An international comparison*. OECD Social, Employment and Migration Working Papers 224. DOI: 10.1787/b5e5f16e-en.

Desiere, S. and Struyven, L. (2020) 'Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off', *Journal of Social Policy*, pp. 1–19. DOI: 10.1017/S0047279420000203.

Duggan, J. *et al.* (2020) 'Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM', *Human Resource Management Journal*, 30(1), pp. 114–132. DOI: 10.1111/1748-8583.12258.

Edwards, L. and Veale, M. (2017) 'Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for', *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2972855.

epicenter.works (2019) *epicenter.works veröffentlicht Details zum AMS-Algorithmus*. Available at: https://epicenter.works/content/epicenterworks-veroeffentlicht-details-zum-ams-algorithmus (Accessed: 6 July 2021).

epicenter.wroks (2019) *Das Problem mit dem AMS-Algorithmus*. Available at: https://epicenter.works/content/das-problem-mit-dem-ams-algorithmus (Accessed: 6 July 2021).

Eubanks, V. (2018) *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

Floridi, L. *et al.* (2018) 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines*, 28(4), pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

Floridi, L. *et al.* (2020) 'How to Design AI for Social Good: Seven Essential Factors', *Science and Engineering Ethics*, 26(3), pp. 1771–1796. DOI: 10.1007/s11948-020-00213-5.

Friedman, B. and Nissenbaum, H. (1996) 'Bias in computer systems, *ACM Transactions on Information Systems*, 14(3), pp. 330–347. DOI: 10.1145/230538.230561.

Goyal, N. and Howlett, M. (2019) 'Framework or metaphor? Analysing the status of policy learning in the policy sciences', *Journal of Asian Public Policy*, 12(3), pp. 257–273. DOI: 10.1080/17516234.2018.1493768.

Goyal, N., Howlett, M. and Taeihagh, A. (2021) 'Why and how does the regulation of emerging technologies occur? Explaining the adoption of the EU General Data Protection Regulation using the multiple streams framework', *Regulation & Governance*, p. rego.12387. DOI: 10.1111/rego.12387.

Holl, J. (2018) *Das AMS-Arbeitsmarkt Chancen-Modell*.

IBM (2020) 'What is automation'. Available at: https://www.ibm.com/topics/automation (Accessed: 5 April 2021).

Kayser-Bril, N. (2019) 'Austria's employment agency rolls out discriminatory algorithm, sees no problem', *AlgorithmWatch*. Available at: https://algorithmwatch.org/en/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm (Accessed: 13 July 2021).

Kingdon, J. (1984) *Agendas, alternatives and public policies*,. Boston: Little, Brown and Company.

Knaggård, Å. (2015) 'The Multiple Streams Framework and the problem broker: The Multiple Streams Framework and the problem broker', *European Journal of Political Research*, 54(3), pp. 450–465. DOI: 10.1111/1475-6765.12097.

Kopf, J. (2019) 'Offener Brief von Johannes Kopf an Fr. Prof. Sarah Spiekermann › Blog von Johannes Kopf'. Available at: https://www.johanneskopf.at/2019/09/24/offener-brief-fr-prof/ (Accessed: 8 July 2021).

Kraft-Buchman, C. and Arian, R. (2019) *A AFFIRMATIVE ACTION FOR ALGORITHMS - Artificial Intelligence, Automated Decision-Making & Gender Position paper.pdf*.

Kuziemski, M. and Misuraca, G. (2020) 'AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings', *Telecommunications Policy*, 44(6), p. 101976. DOI: 10.1016/j.telpol.2020.101976.

Lambrecht, A. and Tucker, C. E. (2016) 'Algorithmic bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads', *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2852260.

Lopez, P. (2019) 'Reinforcing Intersectional Inequality via the AMS Algorithm in Austria'.

Malgieri, G. (2019) 'Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations', *Computer Law & Security Review*, 35(5), p. 105327. DOI: 10.1016/j.clsr.2019.05.002.

Mao, F. (2020) *The human cost of Australia's illegal 'robo' hunt for welfare cheats - BBC News*. Available at: https://www.bbc.com/news/world-australia-54970253 (Accessed: 12 July 2021).

Mateescu, A. and Nguyen, A. (2019) *Algorithmic Management in the Workplace*. Data and Society.

Mittelstadt, B. D. *et al.* (2016) 'The ethics of algorithms: Mapping the debate', *Big Data & Society*, 3(2), p. 205395171667967. DOI: 10.1177/2053951716679679.

Morley, J. *et al.* (2019) 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices', *Science and Engineering Ethics*. DOI: 10.1007/s11948-019-00165-5.

Morozov, E. (2013) *To save everything, click here : the folly of technological solutionism*. New York : PublicAffairs.

Niklas, J., Sztandar, K. and Szymielewicz, K. (2015) *PROFILING THE UNEMPLOYED IN POLAND: SOCIAL AND POLITICAL IMPLICATIONS OF ALGORITHMIC DECISION MAKING*.

Noble, S. (2018) *Algorithms of Oppression. How Search Engines Reinforce Racism.* New York, NY: NYU Press.

OECD (2018) *'Profiling tools for early identification of job seekers who need extra support, Policy Brief on Activation Policies*. Paris: OECD Publishing.

O'Neil, K. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

Osoba, O. and Welser, W. (2017) *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Santa Monica, Calif: RAND Corporation.

Pasquale, F. and Citron, D. (2014) 'The Scored Society: Due Process for Automated Predictions', *Washington Law Review*, 89.

Peña Gangadharan, S. and Niklas, J. (2019) 'Decentering technology in discourse on discrimination', *Information, Communication & Society*, 22(7), pp. 882–899. DOI: 10.1080/1369118X.2019.1593484.

Sandvig, C. *et al.* (2014) 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms'.

Stoppt den AMS-Algorithmus! (no date). Available at: https://amsalgorithmus.at/ (Accessed: 6 July 2021).

Szigetvari, A. (2018) *AMS-Vorstand Kopf: 'Was die EDV gar nicht abbilden kann, ist die Motivation', DER STANDARD*. Available at: https://www.derstandard.at/story/2000089096795/ams-vorstand-kopf-menschliche-komponente-wird-entscheidend-bleiben (Accessed: 21 February 2021).

Szigetvari, A. (2020) *Datenschutzbehörde kippt umstrittenen AMS-Algorithmus, DER STANDARD*. Available at: https://www.derstandard.at/story/2000119486931/datenschutzbehoerde-kippt-umstrittenen-ams-algorithmus (Accessed: 21 February 2021).

Tarkowski, A. (2019) 'POLAND', *AlgorithmWatch*. Available at: https://algorithmwatch.org/en/automating-society-2019/poland (Accessed: 8 May 2021).

Toh, A. (2020) *Automated Hardship: How the Tech-Driven overhaul of the UK's Social Security System Worsens Poverty, Human Rights Watch*. Available at: https://www.hrw.org/report/2020/09/29/automated-hardship/how-tech-driven-overhaul-uks-social-security-system-worsens (Accessed: 12 July 2021).

Vedung, E. (2012) 'Six models of evaluation', in *Routledge Handbook of Public Policy*. Routledge. DOI: 10.4324/9780203097571.ch29.

Vervloesem, K. (2020) *How Dutch activists got an invasive fraud detection algorithm banned, AlgorithmWatch*. Available at: https://algorithmwatch.org/en/story/syri-netherlands-algorithm/ (Accessed: 9 December 2020).

Weber, T. (2011) *PROFILING SYSTEMS FOR EFFECTIVE LABOUR MARKET INTEGRATION*. DG Employment, Social Affairs and Inclusion, p. 28.

Whittlestone, J. *et al.* (2019) *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. London: Nuffield Foundation.

Wimmer, B. (2019) *Was der neue AMS-Algorithmus für Frauen wirklich bedeutet*. Available at: https://futurezone.at/netzpolitik/was-der-neue-ams-algorithmus-fuer-frauen-wirklich-bedeutet/400617302 (Accessed: 8 July 2021).

World Economic Forum (2019) *AI Procurement in a Box: AI Government Procurement Guidelines*.

Xenidis, R. and Senden, L. (2020) 'EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination, in Ulf Bernitz et al (ed.) *General Principles of EU law and the EU Digital Order*. Kluwer Law International.

Yin, R. K. (2003) *Case study research: design and methods*. Thousand Oaks, California: Sage Publications.

Zuiderveen Borgesius, F. J. (2020) 'Strengthening legal protection against discrimination by algorithms and artificial intelligence, *The International Journal of Human Rights*, pp. 1–22. DOI: 10.1080/13642987.2020.1743976.

Appendix 1: Thesis Report

Artificial Intelligence in Social Protection Services:

Mitigating the Risk of Gender Discrimination and Social Exclusion

Thesis Report

Word Count: 6026 (excluding bibliography and footnotes)

Table of Contents

Introduction

This thesis report serves as a theoretical basis for the master's thesis to be completed throughout the academic year 2020-21. The thesis aims to analyze whether and how governments can mitigate the risks of gender discrimination and social exclusion posed by the use of AI in social protection services. The thesis will focus on selected EU member states to assess the extent to which national governments are ready to mitigate these risks, to identify existing limitations, and to analyze how EU-level tools could address the shortcomings existing at the national level. The problem is that the uptake of AI across sectors has significantly exceeded corresponding legal, policy, and institutional adjustments that are needed to mitigate potential risks and ensure the accountability of AI systems. The AI systems are being implemented without having developed proper monitoring, accountability, liability, and regulatory tools. The use of AI in social protection services caries a heightened risk of exclusion and discrimination. Social protection services involve working directly with more vulnerable groups, which tend to lack necessary technical skills and are likely to generate less data that is used to make decisions in AI-based identification and benefit/job allocation systems. The uptake of AI in the public sector is happening along with similar trends in the private industries, which has raised concerns about systemic unemployment caused by accelerating job automation. Therefore, the role of social protection services is becoming increasingly important in ensuring a more equitable transition through society-wide automation. Automating social protection services without mitigating the risks of discrimination can contribute to and exacerbate the existing systemic marginalization of particular groups of people, thus perpetuating inequality, infringing on fundamental rights, and increasing overall societal vulnerability.

The use of AI in the public sector and the associated risks, so far, have received significantly less attention in the academic literature than the use of AI in the private sector. Among those works that focus on governments and AI, the governments are primarily seen as 'regulators' of AI, not as 'users' of these technologies. The goal of this report and the subsequent thesis is to contribute to the knowledge about the governance of AI when AI solutions are used in the public sector, focusing specifically on risk mitigations from gender and social justice perspective. This report first briefly overviews the EU context and relevant terminology in AI. Second, it introduces the issues of algorithmic opacity and algorithmic discrimination. Third, it provides a literature review focused on two complementary frameworks in AI governance – ethics framework and regulatory framework. Then it outlines the "3 Layer Governance' analytical framework that will be used in the future analysis of selected case studies. Lastly, it provides an overview of research design and methods, accompanied by a Gant Chart showing the thesis completion schedule.

Background and Key Terminology

Context

In 2018, the EU released the White Paper announcing its AI strategy and commitment to ethical and human-centered artificial intelligence, accompanied by several new institutions (AI Watch, AI4EU Forum, High Expert Group on AI, etc.), projects, investment and funding schemes. The commitment to ethical and human-centered AI distinguishes the EU from the US and China, two major players in the global AI development market. However, this approach also requires additional legal, policy, and regulatory measures to accompany AI development and deployment. In 2018, all 28 member states signed the Declaration of Cooperation on AI, supporting the Commission's strategy and the Coordinated Plan on AI. [1] By February 2020, sixteen EU member states had their national strategies on AI published, six member states were at the final stages of the strategy drafting process, and the other six started developing their own.[2]

The degree to which EU member states have committed to 'ethical' and 'human-centered' AI in their strategies varies as well as the types of commitments they foresee.[3] They further vary in their ranking on the AI Readiness Index[4] as well as in the priorities set out in their national strategies. Critics of the EU-level strategy highlight the lack of attention paid to the use of AI in the public

[1] Charlotte Stix, "A Survey of the European Union's Artificial Intelligence Ecosystem," Element AI (Cambridge: Leverhulme Centre for the Future of Intelligence, University of Cambridge, March 2019).

[2] Van Roy, V., AI Watch - National strategies on Artificial Intelligence: A European perspective in 2019, EUR 30102 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-16409-8 (online), doi:10.2760/602843 (online), JRC119974.

[3] Cori Crider, "Mapping RegulatoryProposals for Artificial Intelligence in Europe" (Access Now, 2018).

[4] "Government AI Readiness Index 2019 — Oxford Insights," Oxford Insights, accessed August 13, 2020, https://www.oxfordinsights.com/ai-readiness2019.

sector, where it is deemed a 'low-risk' area, despite potential systematic harms. [5] National strategies, in turn, despite showing the intention to increase the uptake of AI in the public sector, also lack specific regulatory proposals.[6]

Artificial Intelligence (AI)

Artificial Intelligence (AI) does not have a single agreed-upon definition. The definition proposed by John McCarthy, one of the founders of the field of AI, is that "AI is the science and engineering of making intelligent machines," whereby intelligence refers to an ability to achieve goals.[7]Another commonly used definition of AI, developed by Stuart Russel and Peter Norvig, is "AI is a rational agent/system that acts to achieve the best outcomes. There are four categories of 'intelligence': 1) Systems that think like humans; 2) Systems that act like humans;3) Systems that think rationally; 4) Systems that act rationally.[8] Rather than being a specific technology, AI is a field, which entailing several subfields. These subfields include machine learning, robotics, neural networks, vision, natural language processing, and speech processing.[9] Scholars distinguish between Narrow AI and General AI (AGI). Narrow AI is currently widely used and refers to AI systems that "demonstrate some properties associated with intelligence but only in a specific task domain."[10] These tasks include image recognition, language processing, media content curation, medical diagnostics, autonomous vehicles. General AI(AGI) does not yet exist, but it would be

[5] Javier Espinoza and Madhumita Murgia, "The Four Problems with Europe's Vision of AI," *Financial Times*, February 26, 2020, https://www.ft.com/content/6759046a-57bf-11ea-a528-dd0f971febbc.
[6] Crider, "Mapping RegulatoryProposals for Artificial Intelligence in Europe."
[7] "Basic Questions," accessed August 14, 2020, http://www-formal.stanford.edu/jmc/whatisai/node1.html.
[8] Stuart J. Russell, Peter Norvig, and Ernest Davis, *Artificial Intelligence: A Modern Approach*, 3rd ed, Prentice Hall Series in Artificial Intelligence (Upper Saddle River: Prentice Hall, 2010).
[9] Lindsey Andersen, "HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE" (Access Now, 2018).
[10] Iason Gabriel, "Artificial Intelligence, Values and Alignment," *ArXiv:2001.09768 [Cs]*, January 13, 2020, http://arxiv.org/abs/2001.09768.

systems that can reason across different domains and would be able to complete a wide range of cognitive tasks.[11] For this thesis, the focus is exclusively on narrow AI. I use the operationalized definition of AI as tangible technology proposed by M. Scherer "AI is a machine/software capable of performing tasks that, if performed by a human, would be said to require intelligence."[12]

Another essential concept in the contemporary discourse surrounding AI is 'aligned AI,' which various policy and strategy documents refer to as 'ethical and human-centered AI.' [13] Value alignment in AI means ensuring that the objectives of machines align with human values. Failure to ensure value alignment may produce harmful outcomes in the short and medium-term, and in the long run, with the potential progression towards AGI, may pose a serious risk of machines getting out of control.[14] Ensuring value alignment in AI consist of two components: 1) technical part that focuses on how "to encode values or principles in AI systems so that they reliably do what they should do"; 2) normative part that focuses on what values and principles should be encoded in AI.[15] There has been an increasing interest among researchers, practitioners, and regulators in the technical aspect of value-alignment that resulted in expanding technical research on fairness, accountability, and transparency in machine learning (FATMLT). This research aims to address the potentially discriminatory impact of machine learning and develop adequate technical and computing solutions.[16] The emerging field of AI governance, in turn, helps address

[11] Gabriel.
[12] Matthew U. Scherer, "REGULATING ARTIFICIAL INTELLIGENCE SYSTEMS: RISKS, CHALLENGES, COMPETENCIES, AND STRATEGIES.," *Harvard Journal of Law & Technology* 29, no. 2 (2016): 353–400.
[13] "ETHICALLY ALIGNED DESIGN A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems" (IEEE, n.d.).
[14] Stuart Russell, "Q&A: The Future of Artificial Intelligence"," 2016, http://people.eecs.berkeley.edu/~russell/temp/q-and-a.html.
[15] Gabriel, "Artificial Intelligence, Values and Alignment."
[16] "Fairness, Accountability, and Transparency in Machine Learning," accessed August 14, 2020, https://www.fatml.org/.

the normative component of value-alignment. AI governance is a rapidly developing academic field that focuses on how people can best navigate the transition to advanced AI systems,[17] focusing on the political, economic, military, governance, and ethical dimensions. It is primarily concerned with the institutions and contexts in which AI is built and used and "ensuring that the development and use of AI have the goals, incentives, worldview, time, training, resources, support and organizational settings to do so for the benefit of humanity."[18]

Machine Learning

In recent years, the AI field has expanded and acquired a more transformative and disruptive potential due to rapidly advancing techniques, such as Machine Learning (ML). Machine learning is a sub-field of AI, whereby algorithms can improve in performance over time by 'learning' from the data it uses. The 'learning' refers to a statistical process that starts with real-world historical data and tries to derive a rule to explain that data or predict future data, based on the input data.[19] AI and Machine learning are increasingly more commonly adopted in algorithmic decision-making (ADM) processes in the private and public sectors. Algorithmic decision-making refers to a decision-making process based on statistical data analysis that yields scores or categories to aid the decision-making process. The algorithmic decision-making process may involve traditional preprogrammed statistical models as well as Machine Learning. Both simple statistical models and ML algorithms are prone to similar problems, such as poorly sampled data,

[17] In this context, "Advanced AI" refers to systems that are substantially more capable than existing systems, and does not necessarily refer to advanced capabilities that would be demonstrated by Artificial General Intelligence (AGI), which currently does not exist.

[18] Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute: Oxford, UK, 2018. Available online: https://www.fhi.ox.ac.uk/govaiagenda/ (accessed on July 23 2020).

[19] Andersen, "HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE."

biased data, and measurement errors.[20] The impact of error rates in both statistical and machine learning algorithms can have significant implications for two reasons. First, even when the error rate is close to zero, if the algorithm is used to process information of millions of users, many could be negatively affected. Second, while ML systems are more accurate than humans, it is dangerous to assume that the accuracy necessarily leads to better outcomes. What makes ML algorithms different from traditional statistical models, is its capacity to calibrate themselves through feedback, which raises an issue of algorithmic opacity, or black box problem.

"Black Box" Problem – Algorithmic Opacity in ML

The 'learning' capacity of ML, coupled with the availability of Big Data, can capture a multitude of patterns that cannot be captured in a single equation of a traditional statistical model. This learning feature of ML makes it impossible for humans not only to understand how the decision was made but also to trace the logic of ML decisions or recommendations.[21] This lack of traceability is also known as a 'black box' or algorithmic opacity problem. Scholars distinguish between three types of algorithmic opacity that require different policy solutions: 1) intentional opacity due to corporate or state secrecy, 2) opacity stemming from technical illiteracy among employers, employees, policymakers, or the general public, and 3) opacity arising from the characteristics of machine learning.[22] The first two sources of opacity require institutional and policy solutions rather than technical ones. Research conducted by Rand Corporation proposes two possible solutions to the problem of opacity 1) ensuring more transparency; 2) adopting

[20] Andersen.
[21] Andersen.
[22] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (January 5, 2016): 205395171562251, https://doi.org/10.1177/2053951715622512.

algorithm audits.[23] Ensuring transparency entails three components. First, new regulations on patents, copyrights, requirements to share underlying code behind a system with responsible authorities would help address the opacity stemming from corporate or state secrecy. Second, increasing awareness about AI among policymakers, regulators, and the general public could help solve the opacity stemming from technical illiteracy. However, given the rapid advancements in machine learning and deep networks used in AI systems, it is often impossible to deconstruct the underlying algorithm, and even if it was, it is likely to be too complicated to analyze and derive meaningful insight. A commonly cited solution to the 'black box' problem is an algorithmic audit, focusing on the consequences of their outputs, not by the underlying code or inner workings. The output-based auditing is the most feasible for policymakers seeking to regulate AI.[24]

Algorithmic Bias

Consensus prevails that algorithms often exhibit bias. Research has shown that various AI systems, produced by different companies, have recurrently exhibited bias against people based on their race, sex, gender. For example, Natural Language Processing (NLP) systems tend to perpetuate gender stereotypes by associating sex with stereotypically male or female professions, e.g., these systems associate 'man' with a 'doctor,' and a 'woman' with 'nurse.' Face recognition technology tends to misrecognize women, transgender people, and people of color. A study of systems produced by Microsoft and IBM for identifying if people wear medical masks in public revealed that the underlying face recognition technology failed to recognize face masks on women,

[23] Osonde Osoba and William Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Santa Monica, Calif: RAND Corporation, 2017).
[24] Osoba and Welser.

mislabeling masks on women as 'duct tape' or 'makeup' with a reported 96% confidence level.[25] Furthermore, AI-based scoring and profiling systems in finance, employment, insurance, law enforcement often scored women and non-white people lower, yielding systematically less favorable outcomes to particular groups of people with no justifiable reason.[26]

Bias in AI systems can occur in two ways: 1) bias at the input level; 2) bias at the system level. Bias at the input level occurs due to:

- The use of biased historical data
- The input data not being representative of the target population (selection bias) and thus the outcomes favor certain groups over others;
- Low-quality data at the input level;
- Incomplete, incorrect, or outdated data[27]

The second mechanism through which bias transmits into the algorithmic system is at the system level, when developers, usually unintentionally, build their personal biases into the parameters of the model or the labels they define. Bias at the system level occurs in two ways:

- Developers allow the system to conflate correlation with causation,
- Developers choose to include parameters that are proxies for protected characteristics, e.g., income as a proxy for race.[28]

Biased data and parameters are a default rather than the exception in ML and tech. One of the broader societal issues that perpetuate this bias at the input level is the lack of data. Several

[25] "Researchers Discover Evidence of Gender Bias in Major Computer Vision APIs," *VentureBeat* (blog), August 6, 2020, https://venturebeat.com/2020/08/06/researchers-discover-evidence-of-gender-bias-in-major-computer-vision-apis/.
[26] Nicol Turner-Lee Barton Paul Resnick, and Genie, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (blog), May 22, 2019, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.
[27] Andersen, "HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE."
[28] Andersen.

mechanisms contribute to the lack of data and, thus, biased learning algorithms. First, large datasets that are needed to train ML are expensive to collect or purchase, thus excluding many companies, public, and civil society from the machine learning market. Second, classes of individuals that do not generate much data, such as more impoverished communities, communities in rural areas, or those not sharing their data, would be automatically excluded.[29] Third, there is a lack of sex-, age-, race-, desegregated, open big data across various fields, contributing to data gaps across multiple domains. In her book, "Invisible Women," Caroline Criado Perez highlights the problem of a gender data gap in healthcare and medical research, international development, and humanitarian aid. These fields are also said to be significantly impacted by the increasing use of AI.[30] Another important factor contributing to the algorithmic bias, especially at the system-level, is the lack of diversity in the big tech community.[31]

[29] World Economic Forum Global Future Council on Human Rights 2016-18., "How to Prevent Discriminatory Outcomes in Machine Learning" (World Economic Forum, March 2018).
[30] Caroline Criado-Perez, *Invisible Women: Data Bias in a World Designed for Men.* (New York: Abrams Press, 2019).
[31] "Five Years of Tech Diversity Reports—and Little Progress | WIRED," accessed August 14, 2020, https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/.

Literature Review

Societal implications and risks of the Black Box and Algorithmic Bias

Among the risks associated with widespread and unregulated use of AI technology is the disruption of labor markets, accountability issues, threats to privacy due to large-scale data collection, exacerbating inequality due to biased decision-making algorithms.[32] Algorithmic-bias, coupled with the black box problem brought by advanced Ai techniques, create a serious risk that algorithmic bias could translate to discrimination and exclusion of individuals and groups. Citron and Pasquale (2014) argue that concurrent use of AI in credit, criminal, and employment services to "score" a person as potential achievement/result based on the historical data carries a high risk of negatively affecting the access to equal opportunity. Without proper regulations, adequate institutional and legal adjustments, the use of such ADM systems are at high risk of violating an individual's right to fairness, accuracy, transparency, and existing venues for redress.[33] Other researchers highlight that AY systems can have a 'disparate impact' on entire subgroups of people, as opposed to just individuals, thus perpetuating and solidifying structural inequality and exclusion.

Furthermore, with the most recent developments in AI, the traditional efforts to conceal sensitive data from learning algorithms were rendered ineffective, since these algorithms can

[32] Gianluca Misuraca, "Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU," Science for Policy (Luxembourg: AI Watch, 2020).
[33] Frank Pasquale and Danielle Citron, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89 (2014): 1.

construct the protected characteristics, such as gender, race, from proxy variables.[34] Researchers studying how algorithmic bias in the times of AI translates in discrimination have identified two main mechanisms that (re)produce inequality. First, algorithmic discrimination can reinforce distributive inequality by maintaining discriminatory access to social goods, such as labor, health services, social benefits, the exercise of rights. Second, it can reinforce symbolic inequality by misrepresenting or rendering certain groups of population invisible. (e,g. search engines returning sexualized or stereotypical images of women, and those images are used to train AI). Therefore, algorithmic discrimination stems from and further entrenches stereotypes and structural, institutionalized patterns of inequality.[35] Social protection systems play a critical role in income and wealth redistribution in societies, thus failing to address the risks of algorithmic discrimination threatens to undermine the main function of this institution.

AI Governance for Risk Mitigation

AI has the potential to transform who people can become, what they can do and achieve, how they interact with each other and the world. If used responsibly, it can expand and create new opportunities.[36] However, consensus prevails among scholars that AI deployment cannot go unregulated, for AI technology carry too many risks if adopted liaises fair. Currently, technological developments in AI and its widespread implementation across various fields vastly exceeds the corresponding developments in standardization, regulatory, legal, and policy developments.

[34] Solon Barocas and Helen Nissenbaum, "Big Data's End Run around Procedural Privacy Protections," *Communications of the ACM* 57, no. 11 (October 27, 2014): 31–33, https://doi.org/10.1145/2668897.
[35] Raphaële Xenidis and Linda Senden, "EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination," in *General Principles of EU Law and the EU Digital Order*, ed. Ulf Bernitz et al (Kluwer Law International, 2020).
[36] Luciano Floridi, "Soft Ethics and the Governance of the Digital," *Philosophy & Technology* 31, no. 1 (March 2018): 1–8, https://doi.org/10.1007/s13347-018-0303-9.

Scholars agree that it is critical to shift from a reactionary approach to technological innovation to a pro-active approach that directs and leads innovation.[37] Despite the increasing deployment of AI systems, the more advanced versions of AI technologies are still in nascent stages, implying that any governance framework developed at this stage will determine the future impact of AI on economy, society, and politics. Policymakers now are at the critical juncture for ensuring ethical and human-centered use of AI technologies to promote democratic values and minimize potential risks. [38] The main discussion in AI governance revolves around what are the best methods and tools for ensuring ethical and human-centered AI. While the responsible use of AI can increase opportunities, its underuse can hinter those opportunities, and its misuse can result in regrettable or highly detrimental outcomes. The underuse of AI may result from policy failures or unintended effects of policy interventions, that manifests itself as misconceived regulation, underinvestment, or public backlash against AI. The misuse or excessive use of AI may result from misaligned incentives, geopolitics, profit-driven motivations, or neglect of risks-vs- benefits analyses.[39] Two major approaches have emerged within the AI governance literature - the ethics framework and the regulatory proposals.

[37] Luciano Floridi, "Soft Ethics and the Governance of the Digital," *Philosophy & Technology* 31, no. 1 (March 2018): 1–8, https://doi.org/10.1007/s13347-018-0303-9.

[38] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Accessed: September 6, 2016.

[39] Floridi, L, Cowls, J, Beltrametti, M, Chatila, R, Chazerand, P, Dignum, V, Luetge, C, Madelin, R, Pagallo, U, Rossi, F, Schafer, B, Valcke, P & Vayena, E 2018, 'AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', Minds and machines, vol. 28, no. 4, pp. 689-707. https://doi.org/10.1007/s11023-018-9482-5

The Ethical Framework

The ethics approach aims to develop a set of overarching ethical principles guiding development and the use of AI technologies. There are several sources of these principles. First, there are ethics guidelines developed by several international organizations, governments, academia, and civil society. Second, human rights principles, such as a right to privacy, a right to association, a right to non-discrimination, are an essential source for ethical principles in AI governance.[40]Third, the Sustainable Development Goal (SDGs) context as a framework for human-centeredness. The ethical principles should provide normative constraints on the do's' and 'don'ts' of AI systems. [41] When it comes to the ethics guidelines, over 160 documents were published since 2016, outlining recommendations for the principles of the ethics of AI. Only a few of these guidelines indicate a mechanism for oversight and enforcement.[42] Many of the published documents come from diverse stakeholders, including industry (Google[43], IBM[44], Microsoft[45], Intel), government (Montreal Declaration, Lords Select Committee), international/intergovernmental organizations (OECD[46], European Commission's High-Level Expert Group[47]), and civil society actors (AccessNow, AlgorithmWatch, etc.).[48] A more in-depth

[40] Urs Gasser and Virgilio A.F. Almeida, "A Layered Model for AI Governance," *IEEE Internet Computing* 21, no. 6 (November 2017): 58–62, https://doi.org/10.1109/MIC.2017.4180835.

[41] Jessica Morley et al., "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics*, December 11, 2019, https://doi.org/10.1007/s11948-019-00165-5.

[42] "AI Ethics Guidelines Global Inventory by AlgorithmWatch," AI Ethics Guidelines Global Inventory, accessed August 14, 2020, https://inventory.algorithmwatch.org.

[43] "AI at Google: Our Principles," Google, June 7, 2018, https://blog.google/technology/ai/ai-principles/.

[44] IBM, "Everyday Ethics for Artificial Intelligence" (IBM, 2019).

[45] Microsoft Corporation, "Responsible Bots: 10 Guidelines for Developers of Conversational AI," 2018.

[46] "OECD Legal Instruments," accessed August 14, 2020, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[47] Stephanie Weiser, "Building Trust in Human-Centric AI," Text, FUTURIUM - European Commission, April 3, 2019, https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

[48] "AI Ethics Guidelines Global Inventory by AlgorithmWatch."

analysis of the corpus pf ethical guidelines distinguishes the main principles for ethical AI: beneficence, non-maleficence, autonomy, justice, explicability. These principles constitute the ethical use of AI.

TABLE 1: FIVE ETHICAL PRINCIPLES FOR HUMAN-CENTERED AI[49]

Principle	Meaning
1. Beneficence	Beneficial to, and respectful of, people and the environment
2. Non-Maleficence	Robust and secure; Ensures Privacy and Security;
3. Autonomy	Respectful of human values; The Power to Decide (Whether to Decide)
4. Justice	Fair; Promoting Prosperity; Preserving Solidarity
5. Explicability	Explainable, accountable and understandable

The ethics guidelines mentioned above outline ethical principles that can be divided into two subgroups. First, the ethical principles applicable to AI design and development, and primarily concerned with technological aspects of AI, including data governance and design of ML models. Second, the group of principles that relate to AI governance and impact the formulation of certification standards, legal and regulatory frameworks, and other policy tools. The main ethics principles applied to the technical design of AI are responsibility, explainability, accuracy, auditability, and fairness, which largely overlaps with the central tenets of the use of AI in general outline in Table 1. Embedding these principles in the technological design of AI systems and algorithms could help safeguard against discrimination and ensure accountability.

[49] Floridi, L, Cowls, J, Beltrametti, M, Chatila, R, Chazerand, P, Dignum, V, Luetge, C, Madelin, R, Pagallo, U, Rossi, F, Schafer, B, Valcke, P & Vayena, E 2018, 'AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', Minds and machines, vol. 28, no. 4, pp. 689-707. https://doi.org/10.1007/s11023-018-9482-5

Another essential source for ethical guidelines for AI is human rights principles. Civil society organizations and human rights scholars argue that international human rights law, standards, institutions, and instruments can help devise policy and governance solutions to the risk of AI. A wide range of human rights is affected by AI systems, such as the right to privacy, the right to non-discrimination, the right to redress, the right to political participation, the right to health and education, etc.[50] If human rights serve as a foundation for ethical and human-centered AI, it can be used to mitigate the potential risks. Unlike the ethics guidelines mentioned above, the human rights framework provides legal and policy tools for operationalizing and implementing these standards. Another emerging field providing guidelines on ethical and sustainable use of AI is research on "AI for Sustainable Development Goals"(SDGs), whereby the human-centeredness and compliance with ethics are measured by their contribution to the achievement of SDGs.[51] It is expected that in the future SDG-based approach could provide evidence-based ethical guidelines for the use of AI. Ethics guidelines, human rights, and SDG frameworks could be used to evaluate AI systems and their outcomes. However, some scholars argue that these frameworks are not sufficient for addressing many of the systematic risks posed by AI, for they lack structural compliance and enforcement mechanisms. Furthermore, the ethics and human rights frameworks, in particular, tend to assume that AI harms are individual and too centered on individual consumer behaviors, therefore not sufficient for developing solutions to more systemic problems.[52]

[50] Andersen, "HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE."
[51] "Oxford Initiative on AI×SDGs | Saïd Business School," accessed August 15, 2020, https://www.sbs.ox.ac.uk/research/centres-and-initiatives/oxford-initiative-aisdgs.
[52] Julia Black and Andrew D Murray, "Regulating AI and Machine Learning: Setting the Regulatory Agenda" 10, no. 3 (2019): 22.

Regulatory approach

To address systemic risks of the AI, scholars propose a regulatory framework, citing the failure to regulate the Internet at the early stages as proof that the ethics framework alone is not sufficient.[53] In the context of disruptive technologies, scholars distinguish between ex-ante and ex-post regulations. *Ex-ante* regulations require a prior licensing or approval of technology before its commercial marketing and deployment. *Ex post-regulation*, in turn, does not require prior licensing; technology can be commercially marketed and deployed once it is developed. Any associated risks and harms would be addressed through liability mechanisms *post factum.* Historically, early government intervention in nascent technology markets in the form of licensing or approval occurred in those fields that made use of centralized, scarce, or public resources or posed systemic or 'deep regret' risks. For technologies that do not use centralized or public recourses or their risks are deemed to be diffused or could be managed by individual consumers of remediation, early-stage interventions, such as a priori licensing, traditionally has not been required. Thus identifying what kind of risk technology poses determines the type of regulations, and inconsistent rules may stifle innovation as well as yield regrettable outcomes. [54] However, some features of AI development and operation makes it difficult to adopt current ex-ante or ex-post regulations. [55]

- **Discreet –** unlike previous generations of disruptive technologies, AI research and development require little physical infrastructure, more diverse actors can participate in its creation.

[53] Black and Murray.
[54] Black and Murray.
[55] Scherer, "REGULATING ARTIFICIAL INTELLIGENCE SYSTEMS: RISKS, CHALLENGES, COMPETENCIES, AND STRATEGIES."

- **Discrete -** different components of an AI system may be designed without deliberate coordination and organization
- **Diffuse** – many widely geographically dispersed individuals can participate in an AI project by working on separate components.
- **Opaque** - outside observers may not be able to investigate, understand, and detect potentially harmful features of an AI system.[56]

Some of the existing regulatory proposals seek to derive a combination of tools from different forms of regulation and apply them to AI governance. For example, M. Scherer proposes a national regulatory agency that would establish standards for safety certification (safety would include both societal and discriminatory risks) of AI systems. Then, the designers, producers, and retailers of the certified systems would be subject to limited tort liability. At the same time, those developing and selling uncertified systems would be subjected to strict liability rules.[57] Other authors, such as Olivia Erdélyi and Judy Goldsmith, advocate for a new international artificial intelligence organization that would require binding commitments from states.[58]

Challenges for AI governance

Because of the rapidly changing nature of AI technologies, from a regulatory perspective, it represents a "moving target" that poses a threefold policy challenge.[59] The main issues involve 1) Information asymmetries, 2) finding normative consensus, 3) government mismatches.[60] First, the information asymmetry refers to a situation when only a few experts understand the underlying mechanisms behind AI and Automated Decision-making systems, and other stakeholders, such as

[56] Scherer.
[57] Mathew U. Scherer, "REGULATING ARTIFICIAL INTELLIGENCE SYSTEMS: RISKS, CHALLENGES, COMPETENCIES, AND STRATEGIES.," *Harvard Journal of Law & Technology* 29, no. 2 (2016): 353–400.
[58] Olivia J. Erdélyi and Judy Goldsmith, "Regulating Artificial Intelligence: Proposal for a Global Solution," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18: AAAI/ACM Conference on AI, Ethics, and Society, New Orleans LA USA: ACM, 2018), 95–101, https://doi.org/10.1145/3278721.3278731.
[59] Georgios Kolliarakis and Isabella Hermann, "Towards European Anticipatory Governance for Artificial Intelligence," DGAP Reports (German Council on Foreign Relations, April 2020).
[60] Gasser and Almeida, "A Layered Model for AI Governance."

consumers and policymakers, have significantly much less understanding of underlying mechanisms, implications, and risks. This information asymmetry makes it difficult to assess the benefits and risks of future AI applications. An effective AI governance system needs to incorporate mechanisms through which to increase the public understanding of different forms of AI applications. Second, the issue of finding normative consensus among different interest groups, societies, and states. It refers to the challenge of addressing interest conflict among different stakeholders across various contexts, geographies, institutional and political frameworks, especially where the design and deployment of AI systems involve significant trade-offs, e.g., efficiency versus transparency, etc. Third, AI governance requires complex and multi-tool approach due to the involvement of several policy fields, such as R&D, industrial policies, consumer protection, competition, labor, defense, and foreign affairs. [61] The significant issues that arise here are a government's mismatch between the challenges of the digital age and the existing laws, regulations, and policy tools. [62] Given the degree of autonomy that AI systems demonstrate and the multitude of application possibilities, AI governance requires a blended approach, combining tools ranging from market-oriented solutions to binding command-and-control laws and regulation. Gasser et al. propose a three-layer model approach to AI governance. The proposed layered model allows for different AI models and applications to be considered at different layers using a multitude of tools. [63]

[61] Kolliarakis and Hermann, "Towards European Anticipatory Governance for Artificial Intelligence."
[62] Gasser and Almeida, "A Layered Model for AI Governance."

[63] Gasser and Almeida.

Layered Model for AI Governance

The deployment of AI/ADM systems does not take place in a vacuum. Instead, the new systems interact with already existing socioeconomic, institutional, and political contexts, resulting in multifaceted implications. The complexity, diversity, scale, and level of autonomy demonstrated by AI systems require a new approach to policy, law, and regulation to address these implications. Neither the ethics nor more conventional command-and-control models are sufficient to address these implications, and a more complex, blended approach is needed. Gasser and Almeida (2017) proposed a 3-layer model for AI governance that encompasses both the ethics, soft- and hard- regulatory frameworks for AI governance. The model consists of three interacting layers: 1) technical layer; 2) ethical layer; 3) social and legal layer (Figure 1).

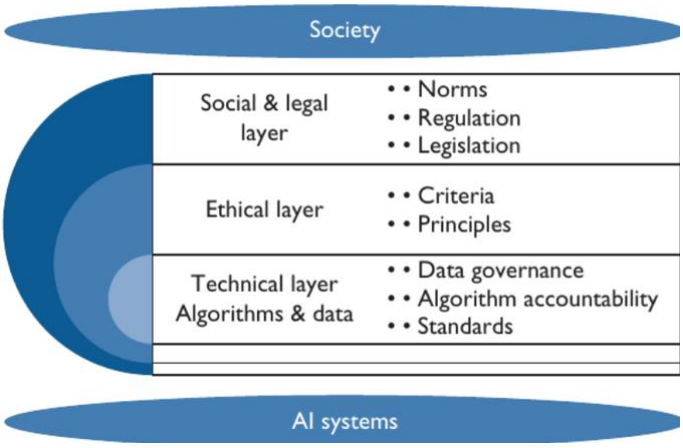


Figure 1: ***Gasser and Almeida's layered model for AI Governance***[64]

A key concept for understanding Gasser and Almeida's 3-layer model of AI governance is modularity. Modularity is a mechanism for managing complex systems, such as AI ecosystems,

[64] Gasser and Almeida.

whereby components of a system are separated and recombined in a variety of flexible ways. In the case of "3-layer-Model', "layering represents a particular form of modularity, in which different parts of the whole system are arranged into parallel hierarchies."[65] This model is particularly helpful as an analytical tool for analyzing whether and how governments adopt various tools at these three levels to mitigate the risks of AI use. In the context of adopting AI in social protection services, it is critical to analyze the context in which these systems are utilized as well as risk mitigation actions that could be taken at each of these three layers.

Public Sector and AI

Most of the academic literature on AI Governance primarily focuses on government institutions as "regulators" of AI, rather than as the "users" of these technologies.[66] While the has been an increasing uptake of AI technologies by the public sector, especially in employment services, policing, and healthcare,[67] the focus within academic literature is primarily on the governance 'of' AI. Much less attention is dedicated to the governance 'with' AI. An analysis of the academic publications published between 2000 and 2019 shows that only 59 out of 1438 scholarly works focus on AI in the public sector. [68]OECD reports that out of 50 countries with national AI strategies, 36 have included plans for public sector transformation using AI. It further estimates that in a few years, AI applications will free up nearly one-third of civil servants' time,

[65] Gasser and Almeida.

[66] Kuziemski, Maciej, and Gianluca Misuraca. "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings." *Telecommunications policy*, vol. 44,6 101976. 17 Apr. 2020, doi:10.1016/j.telpol.2020.101976
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164913/#bib50

[67] "Hello, World: Artificial Intelligence and Its Use in the Public Sector," OECD Working Papers on Public Governance, vol. 36, OECD Working Papers on Public Governance, November 21, 2019, https://doi.org/10.1787/726fd39d-en.

[68] Misuraca, "Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU."

allowing them to shift from mundane tasks to high-value work. Thus AI is expected to help design better policies, make better decisions, improve communications and citizen engagement, improve the speed and quality of public service delivery. [69] AI Watch reports that AI is mostly used for the following government functions:

1. **Enforcement** - AI is used to identify and prioritize targets
2. **Regulatory research, analysis, and monitoring** - AI is used to collect, monitor, and analyze data to improve decision-making.
3. **Adjudication** - AI systems are used to support decision making regarding benefits and entitlements.
4. **Public service delivery and engagement** - AI is used to support the provision of services and communication.
5. **Internal engagement** - AI is used in internal management, such as human resources, procurement, ICT systems.[70]

Furthermore, the type of AI and its usage varies across various domains of the public sector. According to the AI Watch, general public services, healthcare, environmental protection, and public order and safety are the leading sectors in AI uptake in Europe. The heterogeneity in the types of AI systems and its use implies that the associated risks differ depending both on the kind of technology and the purpose of use. While chatbots and intelligent digital assistants are largely risk-free, predictive analytics, profiling, and automated decision-making can pose substantial risks when used in the public sector.

Type of AI	Mostly used in:	Usage

[69] "Hello, World."
[70] Misuraca, "Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU."

Chatbots, Intelligent Digital Assistants, Virtual Agents, and Recommendation Systems	general public services	Includes virtual assistants or online 'bots' currently used to provide generic advice and recommendations to users (51 cases)
Predictive Analytics, Pattern Recognition Simulation and Data Visualisation'	healthcare environmental protection housing and community	AI systems that learn from large datasets to identify patterns that are used for visualizations, simulations or predictions (37cases)
Computer Vision and Identity Recognition'	environmental protection, general public services, healthcare, public order and safety	AI systems that use some form of image, video or facial recognition to obtain information on the external environment and/or identify peoples/objects (29 cases)
Expert and Rule-based systems, Algorithmic decision Making	Healthcare, public order and safety, general public services	AI systems used to facilitate or fully automate decision-making process (29 cases)

Figure 2: AI systems by Typology and sector in the EU and EEA countries as reported by AI WATCH.[71]

AI Watch highlights the issue that while many systems work while being tested, AI systems applied in the real world and different contexts may pose risks, especially in the public sector. For example, social protection services in the public sector involve non-contributory and contributor social benefit and employment provision. These branches of the public sector employ different types of AI at separate stages of the service-provision chain. Public institutions can choose to deploy AI systems for outreach, communication, eligibility assessment, and monitoring. For example, chatbots are often used for assistance, recommendation-making, outreach, and engagement, while automated decision-making or eligibility assessment, resource allocation. Therefore, when analyzing the risks of AI, it is critical to look at the deployment of those systems in the context, not only at the technical model. In the case of social protection services, various risks and harms can emerge at different levels of the process and be perpetuated by systems that

[71] Misuraca.

would be harmless in other contexts.[72] For example, people with a lack of necessary digital skills might be excluded from obtaining basic information that is provided through digitalized services/ chatbots. In contrast, others could be incorrectly classified as high-risk by automated decision-making systems. The risks may arise not necessarily from the technical model but from the type of use of the system and its acceptance by civil servants and end-users, the efficiency of organizational changes brought by the use of AI.

There is an increasing concern about the emerging power asymmetry between governments and citizens. While the uptake of AI technologies tends to increase, the public awareness about and the opportunities for citizens to contest recommendations and results of the AI systems used in the public sector might not necessarily rise at the same pace. This asymmetry raises the issue of algorithmic accountability, which is further exacerbated by the lack of tools to measure, monitor, and evaluate the inputs, processes, and outcomes of the AI-based systems.[73]

[72] Asian Development Bank, *AI in Social Protection – Exploring Opportunities and Mitigating Risks* (Asian Development Bank, 2020), https://www.adb.org/publications/ai-social-protection-exploring-opportunities-mitigating-risks.

[73] Maciej Kuziemski and Gianluca Misuraca, "AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings," *Telecommunications Policy* 44, no. 6 (July 2020): 101976, https://doi.org/10.1016/j.telpol.2020.101976.

Research Design and Methodology

Case Study Selection

To answer the question to what extent governments are prepared to mitigate the risks of gender discrimination and social exclusion posed by the use of AI in social protection services, I will use a comparative case study. I will look at countries that use Automated Decision Making (ADM) or AI in social protection systems. Social protection systems entail non-contributory benefit provision, contributory benefit provision, public employment services. The potential case studies to be analyzed are Denmark, Flanders (Belgium), Austria. Case selections may change as I proceed with my research, potentially focusing on one or two due to the word limit for the thesis. The goal of the thesis is to analyze the cases and identify similarities, differences, and patterns between these cases in terms of their capacity to ensure algorithmic accountability and fairness in social service provision systems. The following criteria are used for the case selection:

- A country has been developing or has developed a national/regional AI strategy;[74]
- A country is ranked upper-medium or high on the AI Readiness Index (0-10):[75]
  - Belgium – 6.85, Austria – 7.52, Denmark –8.6
- A country uses ADM in their social protection services.
  - The cases of Denmark and Flanders (Belgium) are particularly interesting, for they are among the first to advance the statistical profiling model by using Machine Learning. In general, ADM and simple statistical profiling have been used in OECD countries since the 1990s. [76]

[74] "AI Policy," Future of Life Institute, accessed August 13, 2020, https://futureoflife.org/ai-policy/.
[75] "Government AI Readiness Index 2019 — Oxford Insights."
[76] Sam Desiere, Kristine Langenbucher, and Ludo Struyven, "Statistical Profiling in Public Employment Services: An International Comparison," OECD Social, Employment and Migration Working Papers, vol. 224, OECD Social, Employment and Migration Working Papers, February 18, 2019, https://doi.org/10.1787/b5e5f16e-en.

Analytical Framework and Steps

I will use an adjusted Gasser and Almeida's '3-layered approach' to AI governance as a guiding tool for developing my ideal framework for risk mitigation framework. The framework entails three layers: technical layer, ethics layer, and policy layer. The technical layer in the original framework is focused on mathematical models and technical data aspects of data governance. Instead, I will analyze the technical aspects from policy and institutional perspective by looking at what relevant instruments, institutions, or laws are in place to ensure algorithmic accountability and data governance. For ethics layer, I will include not only the 'ethics' commitments outlined in national strategies but also the presence and the role of compliance or supervisory institutions, committees, departments. At the policy layer, I will look at the existing legislation or policies and assess whether these frameworks are well-adjusted to capture and address algorithmic gender discrimination in the times of AI. My research entails two main steps:

1. Developing a model framework for mitigating the risk of algorithmic discrimination based on existing policy proposals and recommendations.
2. Use that model to guide my analysis of the selected case studies, with an attempt to obtain a structured, focused comparison of the cases and answer the following questions:
    1. Assess how well governments are prepared to address the risks;
    2. What gaps there are
    3. How can EU-level tools help fill in those gaps

Data Collection

The first step in the data collection process includes conducting systematic desk research. The main deliverables of this step include a more elaborate literature review and analysis of existing academic papers, documents, reports, official institutional publications to derive the 'ideal' framework for safeguards at technical, ethics, and policy layers. The second step in the data collection process Use expert interviews to 1) fill in the gaps in information that cannot be found

online; 2) to gain better insight from experts for developing the 'model' framework; 3) to gain a better and more in-depth understanding of the case studies, to clarify why specific actions were or were not adopted. For this study, experts to be interviewed can be academics, policymakers with expertise or experience in Ai governance, AI use in the public sector; people who work on projects that led to the adoption of these systems; people who work in social protection agencies in selected countries. Potentially relevant sources would be AlgorithmWatch, AI Watch, European Social and Economic Committee, and relevant civil servants from selected case studies. Interviews would entail semi-structured open-ended questions to gain a better understanding of why or why not specific policies were adopted or not adopted.

Work Plan

Activities	Deliverable	Timeframe and deadlines	2020 Octob	Nove	Decer	2021 Janu	Febr	Marc	April	May	June	July
Finalizing case selection and preparing relevant	**1) Background overview of selected case studie**	**October 30th**										
Desk Research	2) Identify and contact experts for interviews; develop potential questions	October 30th										
Developing theoretical framework for case study analysis	**1)Theoretical framework for case study analysis**	**November 30th**			x							
	2)more extensive literature review											
Primary data gathering	**Interview Transcripts**	**January - February 2021**				x						
Conducting expert interviews (online)		February 30th										
Data Analysis and Theory testing	**synthesis of data in the context of case study**	**March - May 2021**								x		
Using the framework to analyse the studies		April 30th										
triangulating findings from secondoary sources with interview findings		May 15th										
Thesis Writing	**the first draft of complete tehsis**	**March -July 2021**									x	
*Abtract		June 10th										
*Introduction		June 10th										
*Problem Specification		March 30th										
* Research Design		March 30th										
*Research Results and analysis		May 20th										
*Conclusion		May 30th										
*References		May 30th										
proofreading the first draft		May 30th										
submiting the 1st draft for feedback		May 30th and June 15th										
Thesis Sumbission	**complete final thesis**	July 1st										x

Bibliography

Google. “AI at Google: Our Principles,” June 7, 2018. https://blog.google/technology/ai/ai-principles/.

AI Ethics Guidelines Global Inventory. “AI Ethics Guidelines Global Inventory by AlgorithmWatch.” Accessed August 14, 2020. https://inventory.algorithmwatch.org.

Future of Life Institute. “AI Policy.” Accessed August 13, 2020. https://futureoflife.org/ai-policy/.

Andersen, Lindsey. “HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE.” Access Now, 2018.

Bank, Asian Development. *AI in Social Protection – Exploring Opportunities and Mitigating Risks*. Asian Development Bank, 2020. https://www.adb.org/publications/ai-social-protection-exploring-opportunities-mitigating-risks.

Barocas, Solon, and Helen Nissenbaum. “Big Data’s End Run around Procedural Privacy Protections.” *Communications of the ACM* 57, no. 11 (October 27, 2014): 31–33. https://doi.org/10.1145/2668897.

Barton, Nicol Turner-Lee, Paul Resnick, and Genie. “Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms.” *Brookings* (blog), May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

“Basic Questions.” Accessed August 14, 2020. http://www-formal.stanford.edu/jmc/whatisai/node1.html.

Black, Julia, and Andrew D Murray. “Regulating AI and Machine Learning: Setting the Regulatory Agenda” 10, no. 3 (2019): 22.

Burrell, Jenna. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society* 3, no. 1 (January 5, 2016): 205395171562251. https://doi.org/10.1177/2053951715622512.

Criado-Perez, Caroline. *Invisible Women: Data Bias in a World Designed for Men.* New York: Abrams Press, 2019.

Crider, Cori. “Mapping RegulatoryProposals for Artificial Intelligence in Europe.” Access Now, 2018.

Desiere, Sam, Kristine Langenbucher, and Ludo Struyven. “Statistical Profiling in Public Employment Services: An International Comparison.” OECD Social, Employment and Migration

Working Papers. Vol. 224. OECD Social, Employment and Migration Working Papers, February 18, 2019. https://doi.org/10.1787/b5e5f16e-en.

Erdélyi, Olivia J., and Judy Goldsmith. "Regulating Artificial Intelligence: Proposal for a Global Solution." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. New Orleans LA USA: ACM, 2018. https://doi.org/10.1145/3278721.3278731.

Espinoza, Javier, and Madhumita Murgia. "The Four Problems with Europe's Vision of AI." *Financial Times*, February 26, 2020. https://www.ft.com/content/6759046a-57bf-11ea-a528-dd0f971febbc.

"ETHICALLY ALIGNED DESIGN A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems." IEEE, n.d.

"Fairness, Accountability, and Transparency in Machine Learning." Accessed August 14, 2020. https://www.fatml.org/.

"Five Years of Tech Diversity Reports—and Little Progress | WIRED." Accessed August 14, 2020. https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/.

Gabriel, Iason. "Artificial Intelligence, Values and Alignment." *ArXiv:2001.09768 [Cs]*, January 13, 2020. http://arxiv.org/abs/2001.09768.

Gasser, Urs, and Virgilio A.F. Almeida. "A Layered Model for AI Governance." *IEEE Internet Computing* 21, no. 6 (November 2017): 58–62. https://doi.org/10.1109/MIC.2017.4180835.

Oxford Insights. "Government AI Readiness Index 2019 — Oxford Insights." Accessed August 13, 2020. https://www.oxfordinsights.com/ai-readiness2019.

"Hello, World: Artificial Intelligence and Its Use in the Public Sector." OECD Working Papers on Public Governance. Vol. 36. OECD Working Papers on Public Governance, November 21, 2019. https://doi.org/10.1787/726fd39d-en.

IBM. "Everyday Ethics for Artificial Intelligence." IBM, 2019.

Kolliarakis, Georgios, and Isabella Hermann. "Towards European Anticipatory Governance for Artificial Intelligence." DGAP Reports. German Council on Foreign Relations, April 2020.

Kuziemski, Maciej, and Gianluca Misuraca. "AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings." *Telecommunications Policy* 44, no. 6 (July 2020): 101976. https://doi.org/10.1016/j.telpol.2020.101976.

Microsoft Corporation. "Responsible Bots: 10 Guidelines for Developers of Conversational AI," 2018.

Misuraca, Gianluca. "Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU." Science for Policy. Luxembourg: AI Watch, 2020.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.” *Science and Engineering Ethics*, December 11, 2019. https://doi.org/10.1007/s11948-019-00165-5.

“OECD Legal Instruments.” Accessed August 14, 2020. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Osoba, Osonde, and William Welser. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, Calif: RAND Corporation, 2017.

“Oxford Initiative on AI×SDGs | Saïd Business School.” Accessed August 15, 2020. https://www.sbs.ox.ac.uk/research/centres-and-initiatives/oxford-initiative-aisdgs.

Pasquale, Frank, and Danielle Citron. “The Scored Society: Due Process for Automated Predictions.” *Washington Law Review* 89 (2014): 1.

Raphaële Xenidis, and Linda Senden. “EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination.” In *General Principles of EU Law and the EU Digital Order*, edited by Ulf Bernitz et al. Kluwer Law International, 2020.

VentureBeat. “Researchers Discover Evidence of Gender Bias in Major Computer Vision APIs,” August 6, 2020. https://venturebeat.com/2020/08/06/researchers-discover-evidence-of-gender-bias-in-major-computer-vision-apis/.

Russell, Stuart J., Peter Norvig, and Ernest Davis. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Upper Saddle River: Prentice Hall, 2010.

Scherer, Matthew U. “REGULATING ARTIFICIAL INTELLIGENCE SYSTEMS: RISKS, CHALLENGES, COMPETENCIES, AND STRATEGIES.” *Harvard Journal of Law & Technology* 29, no. 2 (2016): 353–400.

Stix, Charlotte. “A Survey of the European Union’s Artificial Intelligence Ecosystem.” Element AI. Cambridge: Leverhulme Centre for the Future of Intelligence, University of Cambridge, March 2019.

Stuart Russell. “Q&A: The Future of Artificial Intelligence",” 2016. http://people.eecs.berkeley.edu/~russell/temp/q-and-a.html.

Weiser, Stephanie. “Building Trust in Human-Centric AI.” Text. FUTURIUM - European Commission, April 3, 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

World Economic Forum Global Future Council on Human Rights 2016-18. “How to Prevent Discriminatory Outcomes in Machine Learning.” World Economic Forum, March 2018.