EQUITECH

Co-funded by
the European Union

REPORT ON AUDITING DISCRIMINATION AND BIAS WITHIN ESTONIAN PUBLIC SECTOR ADM SYSTEMS

Project 101144709 — EquiTech - Improving response to risks of discrimination, bias and intolerance in automated decision-making systems to promote equality

The Gender Equality and Equal Treatment Commissioner's Office of Estonia

The Office of the Equal Opportunities Ombudsperson of Lithuania

Ministry of Economic Affairs and Communications of Estonia

Ministry of Justice of Estonia

Tallinn University of Technology

25.11.2024

Co-funded by
the European Union

Contributing partners:

GENDER EQUALITY AND
EQUAL TREATMENT COMMISSIONER

Office of the
Equal Opportunities
Ombudsperson

REPUBLIC OF ESTONIA
MINISTRY OF ECONOMIC AFFAIRS
AND COMMUNICATIONS

REPUBLIC OF ESTONIA
MINISTRY OF JUSTICE

TAL TECH
DEPARTMENT OF LAW

Table of Contents

Objective

This report is a result of EQUITECH Work Package 2 Task 2.3. ‘Auditing discrimination and bias within Estonian public sector ADM systems and current assessment practice.’

An audit provides an evaluation of potentially discriminatory and biased assessment practices within AI solutions and potential ADM systems used in the Estonian public sector, along with legal and organisational recommendations for improvement. It maps the risks of bias and discrimination in ADM systems used in the public sector, the assessment practice and suggests method(s) for a risk analysis. This report gives a framework for the next Work Package, ‘Impact Assessment Toolbox development’, in which the impact assessment checklist and guidelines will be developed.

Acronyms and abbreviations

AI – Artificial Intelligence

ADM – Automated Decision-Making

EU – European Union

LLM – Large Language Model

MKM – Ministry of Economic Affairs and Communication

ML – Machine Learning

EUIF – Estonian Unemployment Insurance Fund

GDPR – General Data Protection Regulation

Introduction

Artificial Intelligence (AI) can provide faster and more extensive data analysis than humans can, achieving remarkable accuracy and establishing itself as a reliable tool.[1] It can process large amounts of raw data, which may exceed human analytical capacities, enabling AI to provide recommendations on decisions. The decisions made by AI are shaped by the initial data it receives. If the underlying data is incomplete, non-representative or includes many mistakes, the resulting algorithms can perpetuate bias or discrimination, creating potential for widespread inequality.[2]

This bias refers to systematic and replicable errors in computer systems that result in unequal treatment and discrimination based on legally protected characteristics, such as race, gender, age or others. When assessments consistently overestimate or underestimate scores for certain groups, it creates 'predictive bias.'[3] Unfortunately, these biased results are often ignored due to the mistaken belief that AI processes are inherently 'objective' and 'neutral.'

While most algorithms are indeed designed with a neutral, problem-solving purpose, they can produce biased outcomes if the training data provided to them is skewed or incomplete. Modern algorithms might appear neutral, but it is important to ensure that their use does not disproportionately disadvantage members of vulnerable and protected groups. If not properly managed, algorithms can worsen inequalities and perpetuate discrimination against such groups.

Estonia is well-known for its digital society. As a strategic choice, Estonian e-Governance aims to improve the competitiveness of the country and increase the well-being of its people. The aim is to keep the government working seamlessly 24/7. The state supports this by digital identity, secure data exchange, and high-quality databases.[4] One part of this is the provision of user-friendly public services online (e-services). The use of AI in the Estonian public sector is

[1] Chen, Z. (2023). Artificial Intelligence-Virtual Trainer: Innovative Didactics Aimed at Personalised Training Needs. J Knowl Econ 14, 2007–2025. https://doi.org/10.1007/s13132-022-00985-0.
[2] Bornstein, S., (2018). Antidiscriminatory algorithms. Ala. L. Rev., 70, p. 519.
[3] Raghavan, M., Barocas, S., Kleinberg, J., Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Conitzer V, Hadfield G, Vallor S (eds) Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery.
[4] E-Governance. https://e-estonia.com/solutions/e-governance/e-services-registries/ .

a growing trend, with solutions being implemented across various domains.[5] However, automated decision-making (ADM) systems are still relatively uncommon and have yet to gain significant traction. Meanwhile, AI has been one of the key priorities of Estonian e-governance programs since 2018, when the first expert group, formed by the Ministry of Economic Affairs and Communication (MKM) and the Government Office, completed a report on the potential adoption of AI in Estonia and released the first AI strategy.[6] To date, the development and adoption of AI solutions in Estonia's public sector have gained increasing support from agencies implementing new technologies and seeking to automate routine processes. As a result, Estonia now has over 140 AI use cases in the public sector, and this number continues to grow.[7]

Given the EU's data protection regulations and the principles of national constitutional and administrative law, ADM requires a precise and balanced legal framework. The extent of ADM in public administration should be carefully defined to maintain the rule of law and avoid turning administrative decision-making into mere algorithmic processes. So far, the Estonian legislature has allowed public authorities to make automatic decisions in a few areas, such as tax administration, automated border control, etc., where routine decisions are common.

This report presents a preparatory analysis and mapping of the current state of Estonian e-governance for the next steps of the project, particularly the development of guidelines for assessment and an assessment checklist. It provides an overview of the main challenges posed by ADM systems with regard to bias and discrimination in the public sector, with a focus on e-services. More specifically, the report maps the risks of bias and discrimination in ADM systems used in three e-services provided by the Estonian public sector – Bürokratt, OTT, and ABC Gates – along with their assessment methods and the practices used for risk analysis in these systems.

While fully automated administrative decisions and public services are still limited, there are more systems functioning as support tools which, like ADM systems, may result in discriminatory outcomes. The term 'ADM' in this report is used in a broader sense, encompassing AI systems that are not fully automated but may involve algorithms that make decisions.

[5] See AI use cases. https://www.kratid.ee/en/ai-use-cases
[6] AI. Vision and strategies. Retrieved from: https://www.kratid.ee/en/kratt-visioon
[7] AI. AI use cases. Retrieved from: https://www.kratid.ee/en/ai-use-cases

Based on the findings, the report evaluates current assessment practices for these e-services, provides legal and organisational recommendations for improvement, and suggests methods for developing effective assessment tools.

First, the report explains the key definitions necessary to understand the topic. It then briefly outlines the main risks associated with ADM, particularly regarding bias and discrimination, and describes the evaluation methodology used in this audit. The core of the report focuses on auditing the current state of responsibility and management systems for AI, considering the standards, measures, and processes in place to ensure fairness and transparency, as well as to prevent or detect bias in AI systems. It also examines the knowledge of system users and owners about national strategies in data governance and their practical implementation. Three AI systems are used as examples to map the current situation. This audit provides an overview of the Estonian legal framework for ADM systems, a brief introduction to Estonian e-services, descriptions of the audited e-services, and the audit results.

1. Definitions and fundamentals

1.1. Definitions

To begin research on the state's ADM systems, there is a need to present and summarise key definitions and fundamentals of the target research area of automated decision making. For the purposes of the report, it is important to focus on definitions of AI, data mining, mathematical models, algorithm, machine learning, automated administrative decisions and algorithmic administrative decisions. Differing definitions of the terms can be found, as outlined below. In this report, the most common definitions have been used, considering also the context of our task.

Artificial intelligence can be understood as the ability of a computer system to perform tasks commonly associated with the human mind, such as understanding and observing information, communicating, discussing and learning. These features of artificial intelligence must be considered metaphors in the functional sense, because machine learning is not the same as human learning. Artificial intelligence has many branches: automated decision support, speech recognition and synthesis, image recognition and so on. A robot in our context is an artificial intelligence application – an intelligent system.[8]

Data governance is a structured framework that encompasses the processes, policies, roles, and standards necessary for managing an organisation's data assets effectively throughout their lifecycle. This approach aims to ensure that data is secure, accurate, available, and usable, thereby supporting strategic decision-making and compliance with regulatory requirements. Key elements of data governance are data quality, integrity, accountability and transparency.[9]

Data mining is the process of extracting new knowledge – generalisation, data correlation and repeating patterns – from large amounts of data (big data) using statistical methods.[10]

[8] Herberger, M. (2014). 'Künstliche Intelligenz' und Recht. – NJW 2018, p. 2826; H. Surden. Machine Learning and Law. – Wash. L. Rev. 89, pp. 87, 89.
[9] Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organising data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.101493; Atoum, I., & Keshta, I. (2021). Big data management: Security and privacy concerns. International Journal of Advanced and Applied Sciences, 8, 73–83. https://doi.org/10.21833/ijaas.2021.05.009.
[10] Lehr, D. & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn About Machine Learning. – U. C. Davis L. Rev. 51, pp. 653, 672.

Various statistical methods have previously allowed analysts to build **mathematical models** based on datasets to describe what is happening in nature or society. These can, in turn, help assess and classify new situations and predict the future, such as the weather or criminal recidivism. This becomes particularly effective if the models are built on self-learning (machine learning) algorithms.[11]

An algorithm is an exact set of mathematical or logical instructions; more generally, a step-by-step procedure for solving a given problem. The representation of an algorithm in programming language is a computer program. Algorithms where the performance is entirely human-defined are distinguished from algorithms that change their parameters autonomously in the course of learning; systems that automate the traditional decision-making processes in public administration (expert systems) are based on the former. Artificial intelligence applications in public administration are mostly based on learning algorithms (sometimes also on more sophisticated non-learning algorithms).[12]

Machine learning is the process by which an artificial intelligence system improves its service by acquiring or reorganising new knowledge or skills. It is characterised by using the help of learning algorithms to assess situations or make predictions (e.g. making diagnoses, detecting credit card fraud, predicting crime). There are many machine learning techniques with different characteristics: linear and logistic regression, decision tree, decision forest, artificial neural networks, etc.[13]

By **automated administrative decisions** we mean any administrative decision that is prepared or made using automation. This can be based on simpler or more sophisticated non-learning algorithms (expert systems) as well as on machine learning. For example, land tax statements in Estonia are made entirely according to set rules and require no intelligence on the part of a computer. An algorithmic administrative decision is more narrowly a decision made with the help of artificial intelligence. Automated administrative decisions can be divided into fully and semi-automated ones. The latter are approved by an official. Sometimes the computer decides, based on certain criteria provided, whether it is able to make a final decision, such as granting

[11] Berman. E. (2017), Verhaltenssteuerung durch Algorithmen – Eine Herausforderung für das Recht. – AöR 142, pp. 1, 7–8.
[12] Lemley, M.; Casey, B. (2019). Remedies for Robots. – Univ. of Chicago L. Rev. 86, pp. 1311.
[13] Koit, Cf. M., & Roosmaa, T. (2011). Tehisintellekt, (Artificial Intelligence.) Tartu: Tartu Ülikool, p. 194.

a tax refund claim, or if an official must decide. Sometimes the terms ‘automated’ and ‘algorithmic’ decisions are used synonymously or often combined.

1.2. Challenges caused by ADM systems regarding bias

Algorithms are systematic sets of instructions that support decision-making by processing data inputs. However, bias can be introduced during the selection of variables or the design of the algorithm itself. Data bias occurs when the information used to train or adjust an algorithm reflects societal prejudices, potentially leading to the reinforcement or amplification of these biases.[14] Additionally, the complexity and lack of transparency in algorithms often obscure how decisions are made, contributing to a phenomenon known as ‘automation bias’, where people over-rely on the perceived accuracy and objectivity of algorithmic outputs. As a result, algorithms have the potential to perpetuate and intensify existing biases, which is referred to as ‘algorithmic bias.’[15] Therefore, the process of data governance and risk management becomes crucial; the overall management of the availability, usability, integrity, and security of data, to make sure that data is trustworthy and that it is used in compliance with relevant laws and regulations.

The integration of ADM systems into operations raises new challenges in risk management, particularly when biased data influence the decisions these systems produce. Bias in ADM systems can appear at various stages: input, training, programming, algorithmic bias and feedback loops. Input bias stems from incomplete or historically biased data sources. Training bias occurs during the classification of initial data or when assessing the accuracy of outcomes. Programming bias results from the algorithm’s design or its development over time through user interactions, data integration, or the introduction of new data.[16] Algorithmic biasing is different from programming and might come from the mathematical model itself. Feedback loops are situations in which biased decisions, once fed back into the system, can perpetuate and amplify biases.[17]

[14] Fabris, A., Messina, S., Silvello, G., & Susto, G.A., 2022. Algorithmic fairness datasets: the story so far. Data Mining and Knowledge Discovery, 36(6), pp. 2074-2152.
[15] Metin, E., 2024. Literature review on Automated decision-making (ADM) systems in the public sector, EquiTech project, Tallinn University of Technology.
[16] Johnson, K.N., 2019. Automating the risk of bias. Geo. Wash. L. Rev., 87, p.1214.
[17] Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D’Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. International Journal

Bias can be categorised into two types: first-level and second-level bias. First-level bias refers to the unequal application of standards within a system, where different groups may be treated differently despite using the same criteria. This bias is often associated with the technical aspects of ADM systems and can be addressed through measures that improve transparency and explainability. Second-level bias, on the other hand, is more subtle and arises from the unequal selection of the decision-making criteria themselves, which may inherently favour or disadvantage certain groups. This type of bias is embedded in the system's foundational design, making it harder to detect and emphasising the complexity of ensuring fairness in ADM systems.

Our research in literature regarding bias and discrimination identifies several types of bias in ADM systems, particularly those related to the data used to train AI models and the design of algorithms. Selection bias occurs when the training data do not accurately represent the population, leading to distorted outcomes. Bias from structural inequalities arises when AI systems learn from data reflecting existing societal disparities, reinforcing these biases. Measurement bias occurs when the data collection methods are inherently flawed. Label bias occurs when labels assigned during training do not accurately reflect the real-world contexts or outcomes. Furthermore, flaws in algorithmic design can lead to overgeneralisation or incorrect associations, resulting in discriminatory outcomes. These biases are not only technical in nature but also have important social and legal implications, requiring a comprehensive regulatory approach.[18]

Behavioural biases may appear when human cognitive limitations introduce errors, such as misinterpreting model outputs or allowing personal biases to influence decisions. Economic biases stem from incentive structures that prioritise efficiency or profitability over fairness, which can result in biased outcomes. Technical biases are associated with historical data biases or flaws in algorithm design, leading to systematic discrimination. Operating with these biases

of Information Management, 60, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387; Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212 .

[18] Lee, J.A., 2022. Algorithmic bias and the new Chicago school. Law, Innovation and Technology, 14(1), pp. 95-112.

requires a comprehensive approach that incorporates methods from behavioural sciences, economics, and technical fields.[19]

This short summary illustrates the complexity and diversity of bias. However, this theoretical background should be considered in practices dealing with discrimination, as bias can lead to discrimination.

1.3. Discrimination in ADM systems

Discrimination in ADM systems arises from a variety of factors, primarily associated with the data used, the design of the algorithms, and their operational deployment. The following sections provide a comprehensive analysis of these contributing elements:

Historical and Societal Biases

ADM systems frequently reflect and perpetuate historical and societal biases, such as those related to race, gender, or socioeconomic status.[20] These biases, embedded within the training data, mirror existing social inequalities, which are then transferred into the ADM systems, potentially leading to biased and discriminatory outcomes. For example, if historical data includes discriminatory practices against specific racial or gender groups, ADM systems trained on such data are prone to replicate and even exacerbate these biases, resulting in unjust decisions. The uncritical use of historical data without implementing corrective measures can thus reinforce systemic discrimination, disproportionately affecting marginalised communities.

Proxy Discrimination

Proxy discrimination arises in ADM systems that utilise ostensibly neutral variables that, unbeknownst to developers, correlate with protected characteristics such as race or gender. Although these variables may not directly reference a protected group, their use can still lead to outcomes that disproportionately impact these groups, thus perpetuating discrimination.[21]

[19] Adomavicius, G. and Yang, M., 2022. Integrating behavioural, economic, and technical insights to understand and address algorithmic bias: a human-centric perspective. ACM Transactions on Management Information Systems (TMIS), 13(3), pp.1-27.
[20] Johnson, K.N., 2019. Automating the risk of bias. Geo. Wash. L. Rev., 87, p.1214.
[21] Federal Anti-Discrimination Agency (2023) Research project on the protection against discrimination by algorithms. Retrieved, 18 August 2024, from

This subtle form of bias complicates the identification and mitigation of discriminatory practices, as it operates indirectly, making it challenging to pinpoint the source of inequality within the algorithmic framework.

System Design and Development

The design of ADM systems is often influenced by the assumptions and inherent biases of the developers involved in their creation. These biases, whether conscious or unconscious, can be embedded within the algorithms, leading to discriminatory outcomes. This issue is exacerbated by the lack of diversity within the tech industry, which may result in a limited range of perspectives during the development process, thereby reinforcing existing biases.[22] Homogeneity in the developer pool can narrow the scope of considerations during system design, potentially overlooking the diverse needs and contexts of different user groups, further entrenching systemic bias within ADM systems.

Black-Box Nature of Algorithms

Many ADM systems function as 'black boxes', characterised by a lack of transparency in their decision-making processes. This opacity makes it difficult for users and stakeholders to comprehend how decisions are derived, posing significant challenges in detecting discriminatory practices and assigning accountability. The complex and non-transparent nature of these algorithms often hinders efforts to scrutinise and mitigate potential biases, thereby exacerbating the risk of unintended discriminatory outcomes.

Scaling Effects

Decisions generated by ADM systems have the potential to impact large populations, thereby amplifying the consequences of any inherent biases. This scaling effect means that discriminatory decisions, once embedded in the system, can have widespread repercussions. For example, if an ADM system erroneously labels a specific demographic as high-risk, this could result in increased surveillance or discriminatory treatment towards that group, creating

https://www.antidiskriminierungsstelle.de/SharedDocs/forschungsprojekte/EN/RG_AGG_u_Schutz_v_Diskr_d_Algorithmen_en.html.

[22] Adomavicius, G. and Yang, M., 2022. Integrating behavioural, economic, and technical insights to understand and address algorithmic bias: a human-centric perspective. ACM Transactions on Management Information Systems (TMIS), 13(3), pp. 1-27.

a feedback loop that perpetuates and even intensifies existing inequalities.[23] This magnification of bias underscores the importance of rigorous oversight and transparent design in ADM systems to prevent the exacerbation of social inequities on a large scale.

This chapter has explored key definitions and concerns related to the ADM systems, focusing on AI and its components. AI is described as performing tasks resembling human cognition, such as data analysis, decision-making, and pattern recognition, often through self-learning algorithms. ADM systems automate public administrative processes, using either fixed-rule algorithms or more sophisticated AI, which can lead to both benefits and risks. A key challenge in ADM systems is bias, which may be introduced during data input, algorithm design, or training phases, reflecting societal prejudices that lead to biased decision-making. Bias can occur at multiple levels, from technical design flaws to systemic discrimination embedded in historical data, leading to unfair outcomes, especially for marginalised communities. Thus, bias can lead to discrimination. Discrimination also arises from proxy variables that inadvertently correlate with protected characteristics, such as race or gender, resulting in disproportionate impacts. The opaque, 'black box' nature of many ADM systems further complicates accountability and scrutiny, exacerbating biases and discriminatory practices. Moreover, the scale at which ADM systems operate can amplify the consequences of these biases, reinforcing societal inequalities.

[23] Adomavicius, G. and Yang, M., 2022. Integrating behavioural, economic, and technical insights to understand and address algorithmic bias: a human-centric perspective. ACM Transactions on Management Information Systems (TMIS), 13(3), pp.1-27.

2. Evaluation approach

2.1. Data collection and analysis

The audit process was first conducted through two interrelated actions: a document research analysis and discussion with relevant data science experts from Next Gen Digital State research group, TalTech and MKM. The primary research methodology includes a comprehensive review of literature, encompassing research papers, official reports from MKM, Estonian policies and other relevant sources. To validate the findings and address the concerns raised, three discussions were carried out with senior staff members of MKM and Next Gen Digital State research group and TalTech. The following questions were posted in the first round to the respondents to summarise the current situation in the field:

- What are the AI-based systems you have worked with?
- What are they key challenges in AI systems you have analysed?
- Is there any space for biased decisions in the systems you have worked with?
- Do you know any examples of biased outputs from AI systems?

These discussions aimed to identify current solutions, challenges, and expectations regarding government chatbots and public employment service solutions. The meetings were conducted via Microsoft Teams. The responses from these discussions were analysed to identify themes, which are subsequently reflected in the conclusions and recommendations of this audit report.

In the second round, the following questions were asked of the EUIF, the Estonian Ministry of the Interior, and MKM (who are involved with the three ADM systems used in this project):

1. Which governmental department(s) or position(s) owns or manages the AI application which is related to an ADM system? Can you describe the roles or responsibilities of each department in managing or overseeing the system?
2. Which type/category of data is collected and transferred in the system? Could you elaborate on the sources of this data and any specific criteria used in selecting or processing it?
3. What is the procedure for data collection for the AI application?
4. How do you address bias in practice? What kind of potential risks can be identified?
5. What measures or processes are in place to prevent or detect bias, and how often are these reviewed?

6. How do you make sure that individuals are fairly treated by the ADM process? Are there specific guidelines or frameworks you use to define fairness, and how do you measure their effectiveness?
7. Do you know what characteristics are protected by equality law? How do you make sure that you comply with the law?
8. Do you use any specific assessment or audit tools for evaluating the ADM system?
9. Please describe the appropriate structures, policies and procedures to anticipate and address potential bias? Can you give an example or examples? When bias occurs, what are the next steps and is there a specific procedure in place?
10. Do you have standards of fairness, accountability and transparency (not only for the output of an algorithm but the overall decision-making process)?
11. Do you have equality impact assessments or tools to understand how individuals' protected data is used in the system?
12. Do you use algorithms designed to find bias or mitigate them?
13. Do individuals who receive the decision know if the decision was made with the support of an AI application? Can they dispute it and, if so, how?
14. How do Estonia's national strategies and action plans on data governance, artificial intelligence, and digital society contribute to securing automated decision-making (ADM) systems, particularly in minimising risks of discrimination and ensuring fairness?
15. How do you categorise the technical structure of the AI application that uses an automated decision-making system? Are there any risks of discrimination specific to those categories and if there are, what are those? Does your assessment system consider the technical categories of ADM and ML?

2.2. Limitations

The audit report is subject to several limitations:

- The research focuses exclusively on AI-based solutions used in the Estonian public sector, specifically examining tools and measures implemented by the Estonian government. The solutions chosen for the report were selected based on their potential impact if misused or if they exhibit discriminatory behaviour. During the research, it was discovered that fully ADM solutions are typically not yet used in the Estonian

public sector. Therefore, we adjusted the scope to include AI solutions with higher risks of reproducing discriminatory actions.

- The relatively small sample size, consisting of only three experts, limits the report's ability to comprehensively examine all potential solutions. However, the experts consulted have a strong connection to the implementation of AI solutions in Estonia and a direct influence on the specific solutions analysed in this report.
- Given the rapid evolution of AI, the views and practices described may be subject to rapid changes.
- The opinions expressed by the experts reflect their personal perspectives on the topics discussed and should not be interpreted as the official stance of their respective organisations.

2.3. Ethical issues

The participation and collaboration of individuals and knowledgeable stakeholders are duly acknowledged in this report. At no stage was sensitive data used or published. Expert discussions were conducted with respect and impartiality, ensuring that counterparts were not influenced by preconceived outcomes. In adherence to confidentiality requests, the names of participants will not be disclosed. Furthermore, no personal identifiers were collected in the survey.

3. Estonian ADM and AI tools landscape

3.1. Estonian regulatory framework

Estonia, as an EU member state, aligns its national policies closely with the EU's regulatory frameworks, including those governing AI. Rather than developing its own distinct AI regulatory framework, Estonia focuses on implementing and integrating the EU's comprehensive AI regulatory framework together with technology-neutral regulations like the GDPR. The EU Regulatory Framework could be categorised into three groups: Ethical, Policy and Legal. The sources in the 'Ethical' category are *Ethics Guidelines for Trustworthy AI*[24] , *Policy and Investment Recommendations for Trustworthy Artificial Intelligence*[25] and *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence*[26]. The sources in the 'Policy' category are *Communication: Artificial intelligence for Europe*,[27] *European Commission Staff Working Document: Liability for Emerging Digital Technologies*[28], *Communication: Building Trust in Human Centric Artificial Intelligence*[29], *White Paper on Artificial Intelligence: A European approach to excellence and trust*[30], *Impact Assessment of the Regulation on Artificial intelligence*[31] and *Coordinated Plan on Artificial*

[24] INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION. Ethics guidelines for trustworthy AI. 2019a.
[25] INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION. Policy and Investment Recommendations for Trustworthy AI. 2019b.
[26] INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION. Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence. 2020.
[27] EUROPEAN COMMISSION. Artificial Intelligence for Europe. 2018a. COM(2018) 237 final.
[28] EUROPEAN COMMISSION. European Commission Staff Working Document: Liability for Emerging Digital Technologies. 2018b. COM(2018) 237 final.
[29] EUROPEAN COMMISSION. Building trust in human-centric artificial intelligence. 2019. COM(2019) 168 final.
[30] EUROPEAN COMMISSION. White Paper on Artificial Intelligence: A European approach to excellence and trust. 2020. COM(2020) 65 final.
[31]EUROPEAN COMMISSION. Commission Staff Working Document: Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. 2021b. SWD (2021) 84 final.

Intelligence 2021 Review[32]. Finally, the sources in the 'Legal' group are *General Data Protection Regulation of the EU (GDPR)*[33], *Proposal for an AI Liability Directive*[34], *Digital Service Act*[35], *Digital Market Act*[36], *Data Governance Act*[37], *Data Act*[38] and *EU AI Act*[39].

The initial sources in the 'Ethical' category start with the *Ethics Guidelines for Trustworthy AI,* which establish the essential principles for creating, implementing, and utilising trustworthy AI. The *Policy and Investment Recommendations for Trustworthy Artificial Intelligence* provide approaches to promote sustainability, economic development, competitiveness, and inclusivity. The *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence* outline customised recommendations for different sectors, focusing on improving AI applications in distinct areas.

Within the 'Policy' category, *Communication: Artificial Intelligence for Europe* introduces the first strategy, which addresses both the opportunities and challenges of AI. The *European Commission Staff Working Document: Liability for Emerging Digital Technologies* discusses the legal issues that new technologies like AI and the Internet of Things (IoT) present to current

[32] European Comission. Coordinated Plan on Artificial Intelligence 2021 Review. 2021c. COM(2021) 205 final.

[33] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), OJ L119, 4.5.2016, pp. 1–88.

[34] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ, L 265/1, pp. 1-66.

[35] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services (Digital Services Act) and amending Directive 2000/31/EC, OJ, L277/1, pp. 1–102.

[36] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ, L 265/1, pp. 1-66.

[37] Regulation (EU) 2022/863 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), OJ, L 152/1, pp. 1-44.

[38] Regulation (EU) 2023/2854 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), OJ, L.

[39] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance) (2024).

EU laws due to their unique technical and operational features. The *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust* outlines policies that uphold European values and fundamental rights in AI, marking the prominence of the concepts of excellence and trust. The *Coordinated Plan on Artificial Intelligence 2021 Review* expands on the cooperation between the European Commission and member states to develop strategies for investment, legislative measures, and harmonising AI policies. *Impact Assessment of the Regulation on Artificial Intelligence* examines the need for action, the goals, and the impact of various policy alternatives for a European AI framework.

In the 'Legal' category, the *GDPR* is highlighted as the initial regulatory initiative. It sets rules to protect the privacy of individuals in the processing of personal data, including decisions made by automated systems. The *Proposal for an AI Liability Directive* seeks to establish how claims for damages caused by AI systems can be made by revising non-contractual civil liability rules. The *Digital Services Act* is designed to create a safer online space that protects user rights and promotes fair competition among companies. The *Digital Markets Act* aims to enhance competition in Europe's digital sectors by preventing large companies from abusing their market positions and making it easier for new entrants. The *Data Governance Act* aims to boost trust in data sharing, enhance mechanisms for increasing data availability, and address technical challenges to data reuse. The *Data Act* is a broad initiative focusing on overcoming data-related challenges and maximising the opportunities data presents in the EU, emphasising fair access and user rights while ensuring the protection of personal data. Lastly, the *EU AI Act* introduces groundbreaking legislation for AI, providing specific regulations and responsibilities for AI developers, deployers, and users, while also reducing administrative and financial strains, particularly benefiting small and medium-sized enterprises.

All these regulations provide a framework for Estonian national law and policy to deal with AI and ADM systems in public services.

At the national policy level, the Estonian National Action Plan on Artificial Intelligence 2024–2026[40] is the central document outlining the next steps in the digitalisation of public services.

[40] Estonian National Action Plan on Artificial Intelligence 2022–2023. Retrieved 15 October 2024, from https://www.kratid.ee/en/_files/ugd/980182_4434a890f1e64c66b1190b0bd2665dc2.pdf . The third National AI Strategy for 2024–2026 is a continuation of the AI strategy implemented in 2022-2023 (980182_1f685990ca2e462f84c987408a816503.pdf) and 2019–2021 (980182_d2057e15adb24de4924fc2c6f78a7649.pdf).

Published in 2024, the action plan seeks to advance AI adoption across both the public and private sectors in Estonia by increasing the development and implementation of AI solutions and fostering the creation of innovations with export potential. A key focus is on improving digital skills, including cybersecurity and AI knowledge, among the general population and workforce, as well as promoting awareness of AI's capabilities to support effective implementation in organisations. The plan also prioritises growing the pool of AI specialists to meet the needs of the expanding field, advancing academic research related to AI, and enhancing Estonian-language technology to improve access to AI solutions.

Human-centricity is emphasised throughout the plan, aiming to protect people's rights, maintain trust in e-governance, and establish a regulatory framework that supports both protection and innovation. Greater transparency is encouraged through initiatives like providing access to information about algorithms and creating supportive resources, such as sandboxes and a unified portal for AI guidelines, to maintain public trust in AI and e-governance. Data management practices are to remain focused on human interests, ensuring secure and rights-protective AI applications. Furthermore, the plan underscores the importance of making high computing power accessible to academia, AI developers, and the public and private sectors, recognising it as essential for advancing cutting-edge AI technology. Through these efforts, Estonia aims to position itself at the forefront of responsible and innovative AI development.

There are currently three main strategies that cover the activities regarding AI, including ADM systems, in the public sector. The others (in addition to the aforementioned Action Plan on AI) are *Data Action Plan 2024-2025*[41], *Digital Agenda 2030*[42] and *White Paper of Data and AI*[43].

Existing equality laws in Estonia apply to the use of algorithmic systems. The main legal acts regulating discrimination – the Equal Treatment Act[44] and the Gender Equality Act[45] – provide protection on several grounds. However, the problem is that as the Equal Treatment Act binds

[41] Data Field Action Plan 2024–2025. Retrieved 8 November 2024, from https://www.kratid.ee/_files/ugd/7df26f_9796c76b908e492785242d909674a9bf.pdf
[42] Digital Agenda 2030 Retrieved. November 8, 2024, from https://www.mkm.ee/en/e-state-and-connectivity/digital-agenda-2030 .
[43] Andmete ja tehisintellekti valge raamat. Retrieved 8 November 2024, from: https://www.kratid.ee/_files/ugd/7df26f_fda04c9ce24e44f1ba33b17af3030b05.pdf.
[44] Riigikogu. Equal Treatment Act. Retrieved 5 September 2024, from https://www.riigiteataja.ee/en/eli/530102013066/consolide.
[45] Riigikogu. Gender Equality Act. Retrieved 5 September 2024, from https://www.riigiteataja.ee/en/eli/530102013038/consolide.

certain grounds with specific situations then an individual does not get protection in every situation where they should. For example, the grounds 'age' and 'disability' are not protected in case of services, including public services. Even though the Estonian Constitution provides a general protection for any grounds in any situation, in practice, if there is no specific regulation by ordinary law, instances of discrimination cannot be proceeded against or solved.

As shown in the report *Legal Analysis on non-discrimination regulation and regulatory gaps concerning algorithmic systems in Estonia and Lithuania*, the draft Gender Equality and Equal Opportunities Act imposes an obligation on public authorities to ensure that the algorithmic system used by them is designed and developed with due regard to the principle of equal treatment and the objectives of gender equality and equal opportunities following the AI Act's requirement that training, validation, and testing datasets should be relevant, representative, free of errors to the greatest extent possible, and complete in view of the intended purpose. However, the draft proposes a rather abstract principle and a lack of clarity remains as to how far this obligation goes – e.g. the extent to which it requires the incorporation of measures to ensure data quality, impact risk management or auditing. Thus, there is a lack of legal clarity on what authorities are expected to do.

3.2. Relevant AI use-cases

As of 2024, over 140 AI projects have been implemented in the Estonian public sector. These include: risk models of the State Agency of Medicines for agreements on the prices of medicinal products; risk-based selection of claims for VAT refund at the Tax and Customs Board;[46] decision-making support at the Unemployment Insurance Fund for assessing the probability of an unemployed person returning to work; analysis of the customers' calls to the National Social Insurance Board; risk assessment assistance in the Emergency Response Centre; detection of cutting hay with the help of satellite image analysis.[47] Among regular AI solutions, Estonia is also striving for provision of proactive, invisible and human-centric services by relying on the efficient use of data. Proactive government services are public services provided automatically

[46] Sadekov, K. (2021) AI use cases for Government: How Estonia is Leading the Way, MindTitan. Retrieved July 16, 2024, from https://mindtitan.com/resources/industry-use-cases/ai-use-cases-in-government.

[47] Nortal. (2022). OTT – an AI-powered success story in the public sector. Retrieved 16 July 2024 from https://nortal.com/insights/ott-an-ai-powered-success-story-in-the-public-sector/.

by multiple authorities, allowing people to fulfil their responsibilities and access their rights related to a specific event or situation without having to request each service individually.[48]

In determining which of the use-cases to audit for this report, we identified the following criteria:

- System's importance/significance: To facilitate a deeper analysis of AI solutions with risks comparable to those of ADM systems, particularly regarding potential discrimination, the selection criteria included the influence of the target system for the general public (particularly whether the system impacts people's rights or obligations);
- Data processing: Large-scale collection of everyday data from systems' customers. These data may be collected and processed by the selected system either as user input queries or as users' personal data;
- Live mode: The system should be fully operational;
- Availability: The results on system's functioning should be publicly available;
- Higher risk of discrimination: The system operates in a field with a substantial likelihood of discrimination due to its direct impact on individuals' lives, direct interactions with people, and the nature of the data it uses.

Based on the proposed criteria, three systems were chosen: Bürokratt, OTT and ABC Gates.

The virtual assistant Bürokratt is an interoperable network of chatbots on the websites of public authorities that allows people to obtain information from these authorities through a chat window. It provides individuals, or users, with the opportunity to access direct public and information services using virtual assistants. The target group of Bürokratt is the entire population of Estonia, provided that they can communicate in Estonian.

EUIF, which is the Public Employment Service in Estonia, connects candidates with job opportunities and related training modules. OTT is an ML-based data analysis tool that predicts the probability of registered unemployed individuals becoming employed again and identifies factors affecting this probability.

[48] Proactive Government Services. Retrieved 8 November 2024 from https://www.mkm.ee/en/e-state-and-connectivity/digital-services/proactive-government-services .

ABC Gates, or automated border control gates, based on biometrics, speed up border crossings and provide border guards with an additional tool for identifying people and verifying their right of entry. The identification algorithms used in ABC Gates minimise the risk of people crossing the border with false documents and shorten the average time required to cross the border.[49]

[49] AI use cases. Retrieved 8 November 2024 from https://www.kratid.ee/en/ai-use-cases .

4. Analysed AI use-cases

4.1. Bürokratt

Institution: Ministry of Economic Affairs and Communication
Technology: Chatbot
Area: Economy
Project start year: 2017
Project status: In use

Bürokratt is an interoperable network of public sector chatbots that serves as a unified gateway for Estonian citizens and residents, providing effortless access to digital services and information from different bureaus, agencies and institutions.

As a part of the national strategy, Estonia determined the necessity to ensure development of AI capabilities across the public sector, private sector and general public. One of the concepts towards which the Estonian government decided they would build was government services accessible via virtual assistant. As a path to the development of this vision of a virtual-assistant-enabled, proactive relationship between civilians and government, the Ministry of Economic Affairs and Communication (MKM) proposed a proof of concept for a chatbot called Bürokratt that would be a one-stop shop for those who wanted to ask questions of different government ministries.

Bürokratt was developed on the freeware RASA6 software and is based on various language technologies and natural language processing components that allow further development of the chatbot through machine learning.

The principle that people give their data only to one government body, combined with the data protection law passed by Estonia, means that data has to be stored in that place where it is captured. This data protection law means that it is not legally feasible for the ministries in question to create a data hub through which mass processing or queries can be conducted. Because of this, a network of interoperable, localised chatbots associated with individual government entities and databases, needed to be built rather than a single government chatbot. Thus, Bürokratt directs received user queries to the specific chatbot of the respective institution, and further communication with this institution remains private. The service provider describes

the recognition of its service (keywords, sentences and rules for referral to the institution and requests) in the Bürokratt user interface.

4.2. OTT

Institution: Estonian Unemployment Insurance Fund (EUIF)
Partners: CITIS, Nortal, Resta
Technology: Forecast model
Area: Economy
Project start year: 2018
Project status: In use

Using a machine learning model trained on the data of unemployed persons from the past five years, OTT, the decision support system of the EUIF, summarises the situation of a specific person by predicting the probability of them finding employment during the year and the probability of them experiencing unemployment again, highlighting the factors affecting these probabilities.

OTT is an ML-based data analysis tool that predicts the probability of registered unemployed individuals becoming employed again, identifies factors affecting this probability, and assists consultants in offering tailored solutions to clients. It works as a decision support system, which is not a classical ADM system as such, but rather a tool for the EUIF case worker. This automation allows counsellors to dedicate more time to discussing individual problems with clients and evaluating the suitability of job offers recommended by the tool. As a result, the service provided to end-users is efficient and of high quality.

The data utilised in existing AI-enabled solutions within the EUIF includes various personal attributes, such as gender, education level, past employment history, duration of unemployment, eligibility for benefits, and health restrictions. It also incorporates data about the overall labour market situation, such as the types of available jobs in different regions and the number of unemployed individuals. From this, it generates different probabilities, such as the possibility of finding a new job and the likelihood of becoming unemployed again, and introduces different factors that change these indicators.

4.3. ABC Gates

Institution: Police and Border Guard Board/Ministry of Interior

Partners: SMIT

Technology: Computer vision

Area: Security

Project start year: 2020

Project status: In use

Automated border control gates based on biometrics speed up border crossings and provide border guards with an additional tool for identifying people and verifying their right of entry.

The ABC Gate is an automatic border control system that has been in use since 2020 at Tallinn Airport and some border points. With the help of this gate, border control is carried out based on machine-readable document authentication, identifying that the person is the legal user of the document, performing queries in various databases based on his documents, and automatically determining the person's permission to cross the border according to predefined rules. The system mostly consists of a self-service system and an ABC Gate. This kind of border control should make border crossing safer and faster. Since the working principle of ABC Gates is based on biometric technologies, only holders of biometric passports can pass the automatic border control. A person's biometric information is used, such as a digital photograph of the person's face or a fingerprint or iris scan stored on a chip embedded in the passport. The border control system identifies a person by the characteristics of his face and fingerprints. Systems that read biometric data use various technologies to scan and, based on algorithms, collect information about a person's fingerprints, facial features, and then, performing personal identification, compare these data with already known data in databases. ABC Gates cannot be used by those crossing the border with bicycles, wheelchairs, baby carriages, large hand luggage or accompanied by another person; those persons must use normal border control to cross the border. Headgear, glasses, mask or other items that prevent facial recognition must be removed before border control.

5. Audit findings

The aim of the audit was to map the current state of assessment practices for Estonian e-services, provide legal and organisational recommendations for improvement, and suggest methods for developing effective assessment tools. To achieve this, we collected data and information from experts in various positions, as detailed in the section 'Data Collection and Analysis.' We mapped the general framework and governance system of e-services, before focusing specifically on bias and discrimination within the ADM system. We found that the experts' knowledge was somewhat specialised: they were more confident in addressing the issues they deal with regularly, but their understanding of the broader e-services system and its approaches varied. Additionally, questions regarding bias and discrimination appeared to be a new consideration for them. We collected a lot of information, but in a different context. While these results are generally related to our aim, they are not always directly connected. It was challenging to decide which themes to focus on, as we had to consider whether the specific information could contribute to the subsequent tasks of the project. As a result, our findings may appear somewhat unsystematic, but selected findings are very important for us to continue the project by supporting subsequent work packages and tasks.

Based on the sources used, interviews, and questionnaire, the following findings were made:

5.1. Bürokratt

Bürokratt is a network of chatbots. The opportunities for creating AI based services and software/platform is provided to different government agencies, authorities and local municipalities by the Estonian information System Authority. Each client owns and is responsible for the services within their own organisation. As of today, the AI chatbot is being used for answering people's general questions. The bots are trained on the open data (client's own website's information and articles, legal documents etc.). The data is validated by the bot owner. Each Bürokratt owner may have their own data collection procedures, tailored to their specific needs, rules, and regulations, which must, of course, comply with data protection and collection requirements set out by the GDPR, as well as other relevant EU and national laws. Bürokratt uses datasets and databases in various formats for its operations, the majority of which do not contain personal data. Databases created for purposes such as data analysis and machine learning do not include data that would allow for personalisation.

However, from the interviews, we also learned the opinion that a key characteristic of Bürokratt is the unpredictability and constant change of datasets, which makes it impossible to define them unambiguously in terms of names, quantities, and, most importantly, data structures. For example, LLMs can work successfully on different (linguistic) datasets, but to make the entire service cost-effective and highly available, so-called proxy databases must be created for non-machine-learning use.

Regarding the addressing of bias in practice and/or identifying potential risks, each Bürokratt bot owner handles bias according to their specific field. This may be a general risk assessment or a more thorough procedure. Assessing the measures or procedures to prevent or detect bias, the bot mainly gives pretrained answers to people's questions and does not make automated decisions. The individuals have always clear indication if they are communicating with a human or a machine. The identified issues (incorrect interpretation of user requests, misinterpretation of user input words, dialects, slang, and difficulties with processing users' voice commands) do not result in biased output from the chatbot.

Lately, Bürokratt has been updated and also has the potential to use LLMs. When using the LLMs, a monitoring system will be set up to ensure the accuracy of the answers. When implementing the LLM in services, the prevention of bias must be considered as early as during the service-design process. Bürokratt will also have the ability to use privacy-enhancing technologies, such as consent for services and data tracking as one of the measures to make the system more transparent, ethical and trustworthy.

Currently, Bürokratt does not have a common guideline or framework for defining fairness: each bot owner must use the bot according to the general law. However, the risk of bias is very low today. Should bias occur, it is the bot owner's responsibility to remove the related elements from the system. However, a transparent algorithmic standard will be used in the future.

There are no specific assessment or audit tools for evaluating an ADM system and there are no equality impact assessments or tools to understand how individuals' protected data is used in the system. No algorithms have been designed to find bias and mitigate it.

The experts also suggested training the chatbot in relation to state services (Bürokratt or their own bot developed by the state institution) with the chat data of as many institutions as possible (impersonal, but at the same time with important demographic data characterising the customer

from the point of view of service provision) in order to ensure the quality of the source datasets and increase the quality and relevance of the chatbot's output.

Experts also emphasised that state service providers should have seamless access to services requiring data from multiple institutions through a limited number of interfaces. Users should be directed to the appropriate service based on existing interfaces and data collections. However, the current inability to utilise data from various institutions prevents ADM processes.

As Bürokratt was initially planned to be in Estonian (in 2025 a second version of Bürokratt will be able to communicate in both English and Ukrainian) and should include as many different services as possible. Thus, the target group of Bürokratt is the entire population of Estonia, provided that they are able to communicate in Estonian. However, it is advisable (e.g. using translation tools) to consider the English version as soon as possible to help foreigners living in Estonia, who face significantly more obstacles than a local, Estonian-speaking person in using state services and reaching the necessary information because they are not familiar with the functioning of the country and the principles of the digital environment. The interviews also revealed that a native speaker of Estonian who has lived abroad for a long time needs support more than usual.

5.2. OTT

The OTT's product owner is the EUIF Department of Services for Job Seekers, though other departments have also been involved in the development process of the system. OTT is an ML-based data analysis tool that predicts the probability of registered unemployed individuals becoming employed again, identifies factors affecting this probability, and assists consultants in offering tailored solutions to clients. It works as a decision support system, which is not a classical ADM system as such, but rather a tool for the EUIF case worker. This automation allows counsellors to dedicate more time to discussing individual problems with clients and evaluating the suitability of job offers recommended by the tool.

Data used in OTT comes from the EUIF and from other registers. No additional data is collected but data is used that was collected by EUIF from previously provided services.

There are no specific guidelines or frameworks for detecting bias or discrimination, nor are there any specific assessment or audit tools to identify bias or discrimination, or any equality impact assessment tools. Additionally, there is no design in place for algorithms to detect or

mitigate biases in OTT. However, the OTT product owner is currently participating in a pilot project focused on an AI transparency tool, which, upon completion, will provide key information to the general public about how the system is used and its impact. In the future, the algorithmic transparency standard will be implemented.

No case of discrimination has yet been found. The reasoning of the owner for this is that the final decision is made by a human, ensuring the necessary oversight to mitigate any possible biases of OTT.

There is no mechanism for informing a client that the decision is partly supported by AI, and no specific rules for disputing the decision based on that. This means that the general administrative procedure rules are applied but these can be complicated, as discussed in the document, 'Legal analyses on non-discrimination regulation and regulatory gaps concerning algorithmic systems in Estonia and Lithuania.'

Experts raised ethical concerns regarding the implementation of OTT during the interviews. First, they argued that it is problematic that the results are only visible to counsellors and not to the clients being analysed. Although the EUIF has considered sharing the results with the individuals concerned, they have decided against it due to the uncertain impact on those individuals.

Another ethical consideration highlighted in experts' interviews is the dilemma of machine versus human decision-making. Some comparisons have shown that OTT is more precise than EUIF counsellors, raising the question of whether decisions should be made by a machine or if each individual should be evaluated separately. One pilot planned in the EUIF is the use-case aimed at employers' skills intelligence. The field of skills intelligence and identification of skills gaps is a new area where many public employment services are trying to implement AI tools.[50] To identify the skills gap of a person in the EUIF, data about skills is gathered from job advertisements, available training programmes and the person's past employment history, and recommendations will be made to the client based on this data.

[50] Estonian Unemployment Insurance Fund. (2023) Majandusaasta aruanne 2022' Retrieved July 16, 2024, from https://www.tootukassa.ee/web/sites/default/files/2023-04/Eesti%20T%C3%B6%C3%B6tukassa%20aastaaruanne%202022.pdf

5.3. ABC Gates

The owner of the Automated Border Control System (ABC Gates) is the Police and Border Guard Board within the Ministry of the Interior. According to the information received from the Ministry of the Interior, the AI component is used only in face recognition software and face matching is done on a one-to-one principle. This means comparison of a live picture with a picture stored in a chip. The process of mapping high-risk and general-purpose AI solutions is still in the early stages at the Ministry of the Interior; hence, they could not answer questions forwarded to them. The risk management system and data quality assessment methodology are still under development.

Based on this information received from the Ministry of the Interior, it can be assumed that, for detecting biases and discrimination, there are no specific guidelines or frameworks to define fairness, no specific assessment or audit tools for detecting bias or discrimination, and no equality impact assessment tool. Also, there are no algorithms designed to find bias or mitigate it.

There is no mechanism for informing a client that the decision is partly supported by AI, and no specific rules for disputing the decision on that basis.

Based on the data collected and analysed, we can conclude the following: it is clear who owns the ADM system, and in principle, it is known what data is used in the system, but it seems unclear exactly what data is used. Bias and discrimination are new topics, and the experts had never considered them before. Experts answered that no discrimination could occur in their system. When discussing ADM, it was understood as the final decision made in the process, without recognising that algorithms could also make decisions during earlier stages of the process, which could lead to discrimination. No measures or processes were in place to prevent or detect bias, and there were no specific guidelines or frameworks to ensure fairness and transparency in the AI system. There was also a lack of knowledge regarding equality law, specifically as it relates to ADM systems. Furthermore, no specific equality impact assessment exists. We argue that, despite the fact that there are several policies and regulations in relation to AI, a clear system of knowledge about AI systems, including ADM, responsibilities and obligations are uncertain.

6. Conclusions and recommendations

Trust in automated decisions within society signifies a readiness to accept such decisions. That (trust) should mean that individuals can be confident that automated decisions are not biased or discriminatory, can be challenged, and are accompanied by sufficient explanations about their outcomes. These decisions are seen as reliable tools of public administration and an integral part of the rule of law, a unifying principle of the EU.

Estonia should aim to accelerate the development of AI by starting with foundational competencies. By swiftly advancing in AI, Estonia can distinguish itself and gain a competitive advantage. This initiative is further supported by Estonia's existing robust e-government infrastructure.

To mitigate this issue of appearance of algorithmic bias, it is recommended to implement technical measures, such as using unbiased dataset frameworks and enhancing algorithmic transparency, along with management measures like internal corporate ethical governance and external oversight.

Based on the research carried out, several challenges are often associated with chatbots, which can adversely affect user experience and satisfaction during interactions.[51] The key obstacle which should be considered is the fact that users' input can have different meanings depending on the conversational context. Some interactions are highly context-dependent, presenting a particular challenge for generative AI chatbots. These chatbots often struggle to accurately interpret user intent in such conversations. Maintaining an understanding of the context throughout the conversation is a core challenge for these systems.

Besides that, users often exhibit diverse ways of speaking and expressing themselves, encompassing various language styles, spelling, and grammar. Chatbots must be capable of effectively interpreting this diversity. Failures in identifying user intent can result from the variability in user input and limitations within the training data.

Finally, chatbots are susceptible to perpetuating biases inherent in the training data of the LLMs on which they are based. Such biases can lead to unintentional discrimination or the

[51] Dreyling, R., Koppel, T., Tammet, T., & Pappel, I. (2024). Challenges of Generative AI Chatbots in Public Services - An Integrative Review. Retrieved July 16, 2024, from https://ssrn.com/abstract=4850714

dissemination of misinformation due to unbalanced training datasets. Developing a large language model necessitates vast amounts of textual data, yet these foundational models are inherently limited in scope, failing to encompass all business domains.

User confidence and trust in AI-enabled chatbots are largely dependent on addressing several key challenges. For instance, if a chatbot provides outdated information, it can undermine user trust not only in the chatbot itself but also in the government service it represents. Users expect government-provided information to be reliable, comprehensive, and up-to-date.

Another ethical issue raised in the interviews was whether to disclose OTT results to clients. Initially, there was a plan to share results with both EUIF counsellors and clients. However, due to ethical concerns, data protection legislation and the potential for demotivating outcomes, this plan was withdrawn.[52] But what if the client wants to see the results of OTT's prediction and councellor's decision based on that? And then what if this is contested in the court? Every step of administrative authority that led to a certain decision must be traceable: this includes explaining OTT´s functionality and outcome to the court, to decide whether the decision or action of the administrative authority meets all conditions and is legitimate. Again, there is an issue of the cost-effectiveness of public administration vs. compliance with legal norms and finding the balance.

The audit showed that there is a need for specific legal norms regarding discrimination in ADM systems. These rules must cover the whole ADM process, not just the final decision. Also, national law should not follow only the obligations provided by the AI Act but be wider, because some algorithmic decisions used in the public sector would conflict with an obligation to follow these requirements.

There must be clear cooperation between the organisations dealing with supervision and investigation and this must be established by the law. On the other hand, organisations which develop and use AI systems must understand their legal obligations in avoiding and detecting bias and discrimination in ADM systems. They must have certain risk management requirements, and organisations must understand why such requirements are useful and how to apply them. These requirements can also be stipulated in the law – either in the equality laws

[52] Estonian Unemployment Insurance Fund. Data Protection Terms. https://www.tootukassa.ee/en/data-protection-terms/data-protection-estonian-unemployment-insurance-fund/which-cases-do-we

or in the Administrative Procedure Act. Decision-makers in the organisations should demand sufficient explainability of how algorithms work, otherwise they cannot counteract any bias.

Policies about AI, either at EU or national level, should be specified for officials who need to use them in their decisions. However, clearer national leadership is needed. This audit showed that the picture is fragmented, and it is not clear who is responsible. There should be clearer roles and responsibilities for national policing bodies.

It is important to consider that mitigation of bias and discrimination is not purely a technical issue but requires considerations of the wider policy, organisational and legal contexts.

The issue is not simply whether an algorithm is biased but whether the overall decision-making processes are biased: algorithms cannot be looked at in isolation, and it is not enough to look at the final result of the decision, but the whole process needs to be considered.

If possible, specific algorithms should be designed to find bias and mitigate it.

There must be a clear capacity among public officials, developers and others to understand bias and discrimination in the context of ADM systems.

In summary, the audit showed that

- **There is a need for an impact assessment checklist, and guidelines (a set of minimum clear standards on a state level and more specific ones at an AI system level) on how to use this checklist;**
- **There is a need for the guidelines to detect bias and discrimination at all stages of ADM – in its design, implementation and in the development of an ADM system;**
- **There is a need for knowledge about discrimination and bias in general, and more specifically related to ADM systems.**

Based on the scientific literature reviewed, research papers, official reports, legal acts, policies and Estonian use cases, the following three self-assessment methods could be considered in

constructing self-assessment tools: Z-Inspection,[53] Trustworthiness Assurance Assessment for High-Risk AI-Based Systems[54], and SHAPES Project Pilots' Self-assessment for Trustworthy AI[55].

Z-Inspection

Z-Inspection is a comprehensive framework based on 'applied ethics', designed to assess the trustworthiness of AI systems, focusing on ethical, technical, and social implications.[56] It involves a structured process where a multidisciplinary team evaluates AI systems based on pre-defined criteria.

This methodology is especially suited for assessing AI systems in fields where ethical implications are significant, such as healthcare, finance, and public services. According to the authors of the article, Z-Inspection is the first process to assess trustworthy AI in practice.[57] Its multidisciplinary approach ensures a thorough evaluation, making it ideal for systems with high societal impact.

Z-Inspection works by following a structured, three-phase process.[58] The first phase, the **Set Up**, involves forming a multidisciplinary team, setting the scope and boundaries of the inspection, and creating a clear protocol for the process. Next, during the **Assess**[59] phase, the team analyses the AI system's usage, identifies potential ethical, technical, and legal issues, and

[53] Zicari, RV, Brodersen, J, Brusseau, J, Düdder, B, Eichhorn, T, Ivanov, T, Kararigas, G, Kringen, P, McCullough, M, Möslein, F, Mushtaq, N, Roig, G, Stürtz, N, Tolle, K, Tithi, JJ, van Halem, I & Westerlund, M 2021, 'Z-Inspection®: A Process to Assess Trustworthy AI', *IEEE Transactions on Technology and Society*, vol. 2, no. 2, pp. 83–97, https://doi.org/10.1109/TTS.2021.3066209.

[54] Stettinger, G, Weissensteiner, P & Khastgir, S 2024, 'Trustworthiness Assurance Assessment for High-Risk AI-Based Systems', *IEEE Access*, vol. 12, pp. 22718–22745, https://doi.org/10.1109/ACCESS.2024.3364387.

[55] Rajamäki, J, Lebre Rocha, PA, Perenius, M & Gioulekas, F 2022, 'SHAPES Project Pilots' Self-assessment for Trustworthy AI', *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp. 1–7, https://doi.org/10.1109/DESSERT58054.2022.10018790.

[56] Zicari, RV, Brodersen, J, Brusseau, J, Düdder, B, Eichhorn, T, Ivanov, T, Kararigas, G, Kringen, P, McCullough, M, Möslein, F, Mushtaq, N, Roig, G, Stürtz, N, Tolle, K, Tithi, JJ, van Halem, I & Westerlund, M 2021, 'Z-Inspection®: A Process to Assess Trustworthy AI', *IEEE Transactions on Technology and Society*, vol. 2, no. 2, p. 83

[57] *Ibid.*, p. 83.

[58] *Ibid.*, p. 85.

[59] *Ibid.*, p. 86.

verifies these against trustworthy AI requirements, using socio-technical scenarios to map specific issues. In the final **Resolve**[60] phase, the team addresses identified issues and tensions, provides recommendations, and sets up mechanisms for ongoing ethical AI maintenance. Each phase incorporates stakeholder input and strives for transparency, aiming to customise the assessment for specific use cases.

Z-Inspection is valuable because it emphasises ethical principles and the system's impact on people and society.[61] This is essential in applications where stakeholder trust is critical and where the implications of AI use are far-reaching.

Trustworthiness Assurance Assessment for High-Risk AI-Based Systems

This assessment framework focuses on evaluating the trustworthiness of AI systems deemed high-risk.[62] It looks at various dimensions of trustworthiness, such as robustness, fairness, and accountability, specifically tailored for systems that could significantly affect individuals' rights or safety.

The Trustworthiness Assurance Assessment for High-Risk AI-Based Systems follows a structured process focused on ensuring that AI systems meet the trustworthiness standards necessary for deployment. The assessment begins with defining the Operational Design Domain (ODD)[63] and Behaviour Competency (BC)[64], which set the scope and requirements for safe AI operation within specified conditions. Next, it identifies dysfunctional cases and tests the AI against predefined scenarios to assess its resilience to potential failures. The assessment then applies Key Performance Indicators (KPIs) and metrics to evaluate the AI's adherence to safety, reliability, and compliance standards across its lifecycle. Finally, it formulates a trustworthiness argument that consolidates these evaluations, ensuring that the AI meets trustworthiness criteria before deployment and remains compliant through continuous monitoring.

---

[60] *Ibid.*, p. 89.
[61] *Ibid.*, p. 84.
[62] Stettinger, G, Weissensteiner, P & Khastgir, S 2024, 'Trustworthiness Assurance Assessment for High-Risk AI-Based Systems', *IEEE Access*, vol. 12, pp. 22718–22745, https://doi.org/10.1109/ACCESS.2024.3364387, p. 22718.
[63] 'ODD is a concept for specifying the operating conditions of an automated system, often used in the field of Autonomous Vehicles (AV)., Ibid, 22724.
[64] *Ibid.*, 22724.

This assessment can be used for AI tools classified as high-risk under frameworks like the EU AI Act. The framework is designed to ensure that high-risk AI systems meet essential criteria to operate safely and reliably. This understanding aligns with the EU AI Act's classification of high-risk AI applications, which mandates strict oversight to ensure safety, reliability, and ethical compliance. Its structured approach helps organisations comply with regulatory requirements, making it particularly useful in regulated industries.

SHAPES Project Pilots' Self-assessment for Trustworthy AI

The SHAPES Project[65] focuses on AI systems in healthcare, particularly those used by older adults or people with disabilities.[66] Its self-assessment tool enables organisations to evaluate trustworthiness factors like privacy, transparency, and usability from a user-centred perspective.

The SHAPES Project Pilots' Self-assessment for Trustworthy AI works by applying a structured evaluation methodology across its pilot themes.[67] First, each pilot theme, such as those focused on in-home care or cross-border health data exchange, undergoes a self-assessment using an interactive version of the Assessment List for Trustworthy AI (ALTAI),[68] which aligns with the European Commission's Trustworthy AI guidelines.[69] During the self-assessment, the pilot teams evaluate their AI systems against criteria like transparency, technical robustness, and fairness, documenting and rating each area based on its adherence to ethical standards.[70] This multi-case-study approach allows the SHAPES project to identify consistent strengths, such as transparency, and areas for improvement, such as technical robustness. The results from

[65] SHAPES, 2020. About SHAPES. [online] Available at: https://shapes2020.eu/about-shapes/ [Accessed 13 October 2024].
[66] Rajamäki, J, Lebre Rocha, PA, Perenius, M & Gioulekas, F 2022, 'SHAPES Project Pilots' Self-assessment for Trustworthy AI', *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, p. 1.
[67] *Ibid*., 2.
[68] European AI Alliance, 2022. Welcome to the ALTAI Portal! [online] Available at: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal [Accessed 13 October 2024].
[69] European Commission's High-Level Expert Group on Artificial Intelligence, 2019. Ethics Guidelines for Trustworthy AI. [online] Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed 13 October 2024].
[70] Rajamäki, J, Lebre Rocha, PA, Perenius, M & Gioulekas, F 2022, 'SHAPES Project Pilots' Self-assessment for Trustworthy AI', *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, p. 5.

individual cases are then cross-analysed, which informs future recommendations for improving AI trustworthiness within the project's broader goals.

This methodology is effective in environments where end-user trust and understanding are important, particularly in healthcare and social care settings. The SHAPES self-assessment emphasises user-centred evaluation, making it suitable for systems that directly affect users' well-being. By focusing on elements like privacy and transparency, it helps ensure that AI systems are aligned with users' needs and expectations, which is especially important in care contexts.

In conclusion, the three self-assessment methods—Z-Inspection®, Trustworthiness Assurance Assessment for High-Risk AI-Based Systems, and the SHAPES Project Pilots' Self-assessment for Trustworthy AI—are recommended for their comprehensive approaches to evaluating AI systems' trustworthiness. Z-Inspection® employs a multidisciplinary framework to assess ethical, technical, and social implications, making it suitable for high-impact sectors like healthcare and public services. The Trustworthiness Assurance Assessment focuses on high-risk AI applications, ensuring they meet essential criteria for safety, reliability, and compliance, aligning with regulatory standards such as the EU AI Act. The SHAPES Project's self-assessment emphasises user-centred evaluation, particularly in healthcare settings, addressing factors like privacy, transparency, and usability to ensure AI systems meet end-user needs. Collectively, these methodologies provide structured processes to identify and mitigate potential biases and ethical concerns, fostering the development of responsible and trustworthy AI systems.