Funded by
the European Union

EDUCATIONAL MATERIAL FOR DATA SCIENTISTS, DATA PROTECTION SPECIALISTS, AND DEVELOPERS

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

Tallinn University of Technology

Silvia Lips, PhD

Kristi Joamets, PhD

2025

Funded by
the European Union

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Contributing partners:

GENDER EQUALITY AND
EQUAL TREATMENT COMMISSIONER

Office of the
Equal Opportunities
Ombudsperson

REPUBLIC OF ESTONIA
MINISTRY OF JUSTICE

TAL TECH
DEPARTMENT OF LAW

TABLE OF CONTENTS

1. INTRODUCTION

Estonia is widely known for its advances in the digitalisation of state services. The digitalisation of government services entails the use of algorithms and artificial intelligence (AI) systems in relevant administrative processes and decision-making. Algorithms are designed to ensure that decisions are free of stereotypes, more uniform, objective and accurate, and thus fairer. This, in turn, should reduce the number of appeals against decisions and increase the efficiency and speed of services provided by the state.

At present, no state service in Estonia relies entirely on AI-based decision-making; in all cases, a human is still involved, at least to some extent, in every final decision. However, since decision-making processes consist of several stages, biases may seep into some less visible parts and subtly affect the final decision. Since AI systems can autonomously start generating rules, several factors may lead to discrimination in the final decision. Detecting discrimination in AI-based decisions is further complicated by the lack of transparency of many of such systems, as their developers and, ultimately, users may be unaware of the risks or specific biases. To date, there are no guidelines or legal regulations for identifying and preventing discrimination in AI-based decision-making. It is therefore necessary to raise awareness among both the developers and users of e-services about what AI-based discrimination involves, how it manifests, and how it could be identified and prevented.

The educational material was prepared as part of *EQUITECH Work Package 3* Task 3.4 "Development of training materials for data scientists, data protection professionals, public service providers, and public organisations." Two separate sets of study materials have been developed: one aimed at public servants and the other at data scientists, data protection specialists, and software developers.

This training material is intended for data scientists, data protection specialists, and developers and addresses the complexities and potential risks associated with bias and discrimination in AI systems. Given the target group, it focuses on legal principles and regulations on bias and discrimination in AI-based systems, and to a lesser extent on AI-based systems.

2. KEY CONCEPTS RELATED TO ARTIFICIAL INTELLIGENCE[1]

The AI industry evolves at a remarkable pace, introducing expanding terminology in the process. The main concepts relevant to this field are outlined below. Many terms will be familiar from English usage but may seem novel or unfamiliar for Estonian speakers. The definitions provided here have been formulated on the principle that they must conform to the translations of European Union legislation and Estonian national strategy documents. This helps maintain coherence also with additional materials, as the terms used are unambiguously understandable and relatable.

There is no uniform or universally agreed upon definition of artificial intelligence, as it is a very broad and diverse field that encompasses different theories, technologies, and applications. A definition of AI can shift depending on whether the focus is on a machine's ability to reason, learn, and analyse data, or on it performing tasks that would traditionally require human intelligence.

Artificial intelligence (AI) is a term that refers to machines or systems that are designed to mimic human intelligence. The goal of AI is to create systems that can perform tasks that typically require human intelligence, such as reasoning, solving problems, recognising patterns, and making decisions.

In artificial intelligence, a distinction is made between Narrow AI and broad or General AI.

Narrow AI describes systems designed to solve tasks in a single specific or narrow field. This means that a narrow AI cannot operate outside the task that it is programmed to do. Popular examples of narrow AI are voice assistants such as Siri or Google Assistant in the area of speech recognition, or recommendation engines on Netflix or Amazon.

General AI, by contrast, refers to systems with a capacity to perform any intellectual task that a human can undertake. This means that broad AI has the flexibility and understanding that makes it capable of solving a wide range of problems and learning from new areas. Currently,

[1] When it comes to the terms used in this course material, one should keep in mind that no fixed or universally accepted definition has yet been established. In this context, the terms are therefore explained and applied in a broad, deliberately general sense, intended to facilitate understanding their general nature.

broad AI remains more of a theoretical concept, as no such system is known to have been created yet. Should General AI become a reality, it would be capable of reasoning, learning new skills, and making complex decisions.

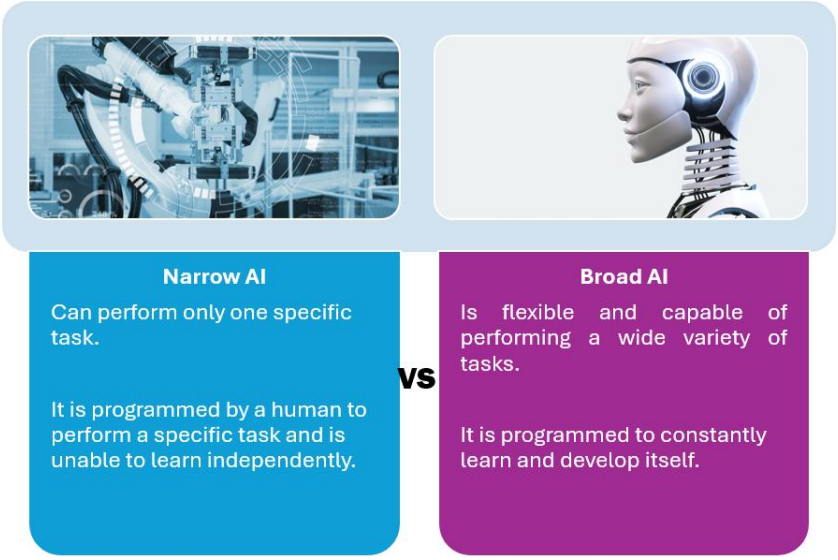

Figure 1. Broad vs. Narrow AI

Returning to the concept of AI, the EU AI Act[2] defines an AI system as follows:

> An "AI system" is a machine-based system that is designed to operate at different levels of autonomy, that can be adaptive once deployed, and that deduces from the input received to achieve direct or indirect goals how to generate outputs, such as forecasts, content, recommendations, or decisions, that may affect the physical or virtual Environment.

The strategy document on the future of Estonian data and AI, the White Paper on Data and Artificial Intelligence 2024–2030, adopts the definition proposed by the Organisation for Economic Co-operation and Development (OECD), according to which an AI system is a machine-based system that, in order to achieve specifically defined or so-called indirect goals, provides outputs based on the given input, such as forecasts, content, recommendations, or decisions, which may have an impact on physical or virtual environments. Once deployed, the different AI systems will have different levels of autonomy and adaptability. Unlike

[2] EU Regulation (EU) 2024/1689 which lays down harmonised rules on artificial intelligence is the first comprehensive legal framework for artificial intelligence worldwide. The rules aim to promote trustworthy AI in Europe.

automation, in which specific tasks are performed according to predefined rules, AI-based systems have the capacity to learn and adapt to new situations, although to varying degrees.[3]

AI systems rely heavily on data. Therefore, these two areas are also closely related and important for the development of data-driven governance and economy. For an AI system to learn and make decisions, it needs data about its surrounding environment.

An **algorithm** is a set of instructions or rules that an AI follows to analyse data and make decisions—that is, a sequence of individual steps that are performed in a specific order.

An **AI model** is created when an algorithm is trained to detect patterns and draw inferences from data. The model learns over time by adjusting its internal parameters depending on the data being processed.

The best-known models of AI, categorised according to activity or method, are the following:

Machine learning (ML) is a model of AI that is based on training systems to improve performance through exposure to data rather than direct programming. Thus, machine-learning algorithms learn through experience and are usually trained on large datasets: instead of being told to complete a task, the system detects patterns in the data to make its own decisions. For example, an email spam filter learns from past emails marked by the AI system as "spam" or "non-spam" to predict the likelihood of new emails being spam. Or, for example, the AI model learns to group similar products in an online store. The subtypes of machine learning are:

- *Supervised learning*, in which an AI model is trained on data where inputs are assigned corresponding outputs (e.g., image classification);
- *Unsupervised learning*, in which an AI model is trained on data without providing specific answers (e.g., clustering);
- *Reinforcement learning*, in which an AI model is trained through a system of rewards and punishments. The AI model performs certain activities and receives feedback based on the results of its activities (e.g., training the AI to play games).

[3] Majandus- ja kommunikatsiooniministeerium, Justiitsministeerium ja Haridus- ja Teadusministeerium, Riigikantselei (2024). Tehisintellekti ja andmete valge raamat 2024–2030 (in Estonian). https://www.mkm.ee/media/10154/download?ref=tehisintellekt.co

Deep learning (DL) is a subfield of machine learning that uses artificial neural networks, which simulate the structure and function of the human brain, and proprietary learning algorithms to model complex relationships in data. In deep learning, a series of simple processors are arranged in layers, forming a network through which the system input passes progressively through all the layers. The process can be likened to the way the light entering the retina is processed in the brain.

When it comes to what is called in-depth learning, the process is organised on two levels: an upper and a lower level. At the upper level, the system identifies the most frequently occurring patterns by drawing on numerous examples. These examples are then described again, resulting in much more compact and meaningful descriptions of the examples. At the lower level, new learning takes place but now using these more refined descriptions of examples. For example, in facial recognition, concepts such as nose, eyes, or mouth may appear as patterns at the upper level. Applying and characterising these concepts allows faces at the lower level to be described more precisely—for example, noting a small mouth, a crooked nose, etc. Such layers of the network allow the AI to learn from large datasets and perform complex tasks—for example, recognising objects in photographs, converting speech to text, translating between languages, and generating text.

Examples of deep learning include various facial and speech recognition solutions, as well as machine translation (e.g., Google Translate).

Natural language processing (NLP) is a subfield of science and technology that focuses on understanding and processing human language to enable machines to communicate using human language. NLP is often used in chatbots, translation software, and speech recognition systems. Examples of natural language processing include digital assistants like Google Assistant or Siri, automated word processing, and machine translation.

While AI contains a range of learning models that do not involve learning at all, deep learning learns directly from data. For this reason, traditional AI systems are often easier to implement than deep computing models. In literature, documents, and texts, the term AI is typically used as a general concept, encompassing both machine learning and deep learning. In practice, then, it becomes crucial to understand which kind of a system or model one is dealing with.

AI functions by imitating certain processes of human intelligence, such as reasoning, learning, decision-making, and problem-solving. Using algorithms, data, and models, AI is able to process large amounts of data, detect patterns, and make predictions or decisions, all without the need for explicit programming of each individual task.

3. INTERNATIONAL LEGISLATION ON PREVENTING DISCRIMINATION IN AI-BASED DECISION-MAKING

In European Union (EU) law, discrimination is regulated both by primary law (treaties) and secondary law (regulations and directives). The Treaty on European Union[4] refers to equality on several occasions. Article 2 emphasises the EU's values of human dignity, freedom, democracy, equality, the rule of law, respect for human rights, and equality between women and men. In addition, Article 3(3) establishes as one of the EU's objectives its aim to combat social exclusion and discrimination and to "promote equality between women and men, solidarity between generations and the protection of the rights of the child."

The Treaty on the Functioning of the European Union[5] states that the EU seeks to eliminate inequalities between women and men and to promote equality in all its activities. Article 10 of the Treaty sets out the objective of combating discrimination on the grounds of gender, racial or ethnic origin, religion or belief, disability, age, or sexual orientation, while Articles 19, 153, and 157 articulate concrete measures that the EU will take to reduce inequalities.

Title III of the Charter of Fundamental Rights of the European Union,[6] which must be applied in interpreting EU law, lays down the most important principles of equality and non-discrimination.

To ensure the implementation of these primary legal principles, the EU has developed a number of provisions that are part of so-called secondary law. Examples of this include Directive 2004/113/EC[7] covering protection against discrimination in relation to goods and services, Directive 2006/54/EC[8] on employment, and Directive 2010/41/EU[9] on entrepreneurship.

---

[4] Consolidated version of the Treaty on European Union, OJ C 326, 26.10.2012, pp. 13–390.

[5] Treaty on the Functioning of the European Union, OJ C 326, 26.10.2012, pp. 47–390.

[6] Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, pp. 391–407.

[7] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, OJ L 373, 21.12.2004, pp. 37–43.

[8] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), OJ L 204, 26.7.2006, pp. 23–36.

[9] Directive 2010/41/EU of the European Parliament and of the Council of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity

In the context of non-discrimination, the key piece of legislation is Directive 2000/43/EC,[10] which covers the prohibition of discrimination on the grounds of racial or ethnic origin in the areas of employment (including self-employment), training, education, social protection and social benefits, and access to and supply of goods and services. Directive 2000/78/EC[11] covers discrimination on the basis of religion or belief, disability, age, and sexual orientation, but only in relation to employment and occupation.

In general, the above Directives only regulate specific types of discrimination and apply in specific contexts. While this allows tackling AI-based discrimination in the cases specifically covered by the Directives (such as automated decision-making, or ADM systems, which discriminate directly against people on the basis of ethnic origin when applying for employment), these rules are not comprehensive enough to prevent discrimination in AI-based decisions, as these systems can be complex and opaque.

The EU's central piece of legislation related to data is the **General Data Protection Regulation** (EU) 2016/679[12] (GDPR), reinforced by additional legislation such as Directive (EU) 2016/680 on the processing of data by law enforcement agencies .[13] In addition, the Council of Europe has adopted the **Convention on the Protection of Individuals with Regard to Automatic Processing of Personal Data** (Convention No 108).[14] These instruments are closely linked to

in a self-employed capacity and repealing Council Directive 86/613/EEC, OJ L 180, 15.7.2010, pp. 1–6.

[10] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, pp. 22–26.

[11] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, OJ L 303, 2.12.2000, pp. 16–22.

[12] 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, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, pp. 1–88.

[13] Directive (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119, 4.5.2016, pp. 89–131.

[14] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. European Treaty Series, No 108, 28.1.1981.

the right to privacy, which is recognised in the respective legal systems.[15] On August 1, 2024, the **EU Artificial Intelligence Act** (AI Regulation)[16] was adopted, establishing harmonised rules for the development and use of AI across the EU. Taken together, these instruments constitute a comprehensive framework to ensure that AI systems are used in a way that is compatible with fundamental rights, including the right to non-discrimination.[17]

In the case of Estonia, these legal acts are and will also be supplemented by national legislation.[18]

The GDPR presents a robust framework for individual privacy. This provides individuals with clear rights in relation to their personal data, including access, rectification, erasure, and restriction of processing of their personal data. Additionally, the GDPR requires transparency from organisations that collect and use personal data by requiring them to disclose how the data is used and with whom it is shared.

The data protection regulation helps to partially solve the so-called black box problem—the opacity of AI systems. We are aware that the system receives data, analyses it through layers of complex algorithms, and makes a decision (result), for example, on a person's eligibility for a loan or being qualified for a job. Under the GDPR, organisations are required to explain how they use personal data in these automated processes, and the data subjects have the right to request and receive information about it.

In specific cases, organisations processing personal data are required to carry out a data protection impact assessment. This is necessary for the systematic and comprehensive assessment of personal aspects relating to natural persons, based on automated processing,

[15] Zuiderveen Borgesius, F. J. (2020). Strengthening Law Enforcement Against Discrimination Against Algorithms and Artificial Intelligence. International Journal of Human Rights, 24(10), 1578. https://doi.org/10.1080/13642987.2020.1743976

[16] 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), OJ L 2024/1689, 12.7.2024, pp. 1–144.

[17] According to Article 1(1), the AI Act aims, inter alia, to promote the uptake of human-centric and trustworthy AI, while ensuring a high level of protection of health, safety, and fundamental rights from the harmful effects of AI systems.

[18] In Estonia, Personal Data Protection Act. RT I, 4.1.2019, 11. https://www.riigiteataja.ee/en/eli/507112023002/consolide

including profiling, and for the large-scale processing of special types of data.[19] Data protection impact assessments can be used as a preventive measure to identify potential risks of bias and discrimination before they occur. Organisations are forced to consider the impact their AI-based systems may have on individuals, assess potential biases, and outline strategies to mitigate risk.

Given the fact that AI-enabled systems use vast amounts of data, it is important that the GDPR enshrines the principle of data minimisation (Article 5(1)(b)). This principle obliges organisations to collect and use only data strictly necessary for a given purpose. By minimising data, it is possible to avoid amassing large data sets that may carry hidden biases or other information about individuals that the organisation does not actually need. This helps ensure that AI systems focus on the most important factors in decision-making, reducing the risk of discriminatory outcomes. For example, an AI system used to decide on who qualifies for social benefits is less discriminatory in situations where it uses only the data provided rather than the data it also collects from public sources about the individual in question.

It is important to note that Article 22 of the GDPR recognises that "the data subject has the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him/her." However, there are exceptions, such as when the subject gives explicit consent. The rules related to Article 22 of the GDPR allow people to request a re-examination of automated decisions, for example when a bank automatically refuses to grant a loan through the bank's website.[20] This is also important for public bodies whose consent is unlikely. Further legislation will then come into play, which should provide for appropriate measures to protect the rights, freedoms, and legitimate interests of the data subject.[21]

Overall, it can be concluded that strong European data protection law serves as an important, albeit limited, protection against potential discrimination arising from AI systems. The GDPR governs automated decisions involving personal data while providing important safeguards,

[19] European Data Protection Supervisor (2024). Data Protection Impact Assessment (DPIA). Retrieved July 11, 2024, from https://www.edps.europa.eu/data-protection-impact-assessment-dpia_en

[20] Zuiderveen Borgesius, F. J. (2020). Strengthening Law Enforcement Against Discrimination Against Algorithms and Artificial Intelligence. International Journal of Human Rights, 24(10), 1580. https://doi.org/10.1080/13642987.2020.1743976

[21] Williams, R. (2021). Rethinking Administrative Law for Algorithmic Decision-Making. Oxford Journal of Legal Studies, 42(2), 473. https://doi.org/10.1093/ojls/gqab032

requiring an appropriate legal basis, and ensuring that individuals have transparency and some control over their data. Yet, as the GDPR focus is on privacy, it has a limited impact on the establishment of appropriate safeguards against AI-based discrimination. For example, data protection rules only apply to the processing of personal data, whereas AI-based decision-making systems, while potentially discriminatory, do not necessarily involve the processing of personal data. Likewise, data protection rules do not extend to predictive models that use aggregated data rather than data related to specific individuals.[22] Some of these shortcomings, however, are set to be addressed by the AI Regulation.

The **AI Regulation** focuses on high-risk AI systems: these systems will be subject to risk management, data governance, transparency, and human oversight requirements to ensure that they function safely and fairly. According to the Regulation, a wide range of public AI systems (e.g., those used in critical infrastructure), the provision of critical services (e.g., public goods or access to public services), as well as the use of AI in law enforcement and justice, would be considered high-risk.[23]

In addition, the legislation specifies that a small number of AI applications that are deemed to pose an unacceptable risk, are outright prohibited—particularly those that exploit individuals' vulnerabilities related to age, disability, or specific social or economic circumstances.[24] The legislation also addresses specific use cases, such as general-purpose models, generative AI systems, and chatbots, which have their own tailored regulatory requirements, in particular with regard to aspects of transparency. AI systems that do not fall into any of the above categories are low-risk systems and are largely exempt from regulatory regulation (to facilitate innovation in lower-risk areas).

The AI Regulation was developed in line with the ethical principles of AI as laid out in the *Ethics Guidelines for Trustworthy AI*, developed by the High-Level Expert Group on Artificial Intelligence.[25] Among the ethical principles are diversity, non-discrimination, and fairness.

---

[22] See Zuiderveen Borgesius, F. J. (2020). Strengthening Law Enforcement Against Discrimination Against Algorithms and Artificial Intelligence. International Journal of Human Rights, 24(10), 1581. https://doi.org/10.1080/13642987.2020.1743976

[23] See Annex III of the AI Regulation.

[24] Prohibited AI practices are listed in Article 5 of the AI Act.

[25] Independent High-Level Expert Group on Artificial Intelligence established by the European Commission, Ethics Guidelines for Trustworthy AI (2019), 8.4.2019. Retrieved October 13, 2024, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

According to Recital 27 of the AI Regulation, this means that "AI systems shall be developed and used in a way that involves different actors and promotes equal access, gender equality, and cultural diversity, while avoiding discriminatory effects and unfair biases prohibited by Union or Member State law."[26] Consequently, relevant risks related to fairness and equity are relevant in determining which AI systems should be considered high-risk or prohibited. For example, the AI Regulation precludes most of the remote real-time biometric identification in public use for law enforcement purposes, as technical inaccuracies in AI systems designed for remote biometric identification of natural persons can lead to biased results and discriminatory effects. Such potentially biased outcomes and discriminatory effects are particularly relevant in the case of age, ethnicity, race, gender, or disability.[27] Many AI systems that provide social assessments of natural persons by public or private actors are prohibited as they can lead to discriminatory outcomes and exclusion of certain groups.[28] In the context of high-risk systems, particular attention is paid to certain biometric identification systems and the use of AI in justice and democratic processes. These are defined as high-risk systems and are subject to additional obligations in the AI Act on the grounds that they may lead to biased outcomes and have discriminatory effects.[29]

The requirements for high-risk AI systems contain several provisions that are relevant to mitigate the risk of unequal treatment. First, the law obliges developers of AI systems to ensure that their systems are trained on data that is relevant, sufficiently representative and, to the greatest extent possible, error-free and complete with regard to their intended purpose.[30] The aim is to reduce potential biases in datasets that could lead to discrimination.[31] To support this, the AI Regulation requires the implementation of appropriate data governance and management practices to ensure the high quality of datasets used for training, validation, and testing.[32] In fact, the Regulation provides a legal foundation for processing special categories of personal

[26] Recital 27 of the AI Regulation.

[27] Article 5(1)(h) and Recital 32 of the AI Regulation.

[28] Article 5(1)(c) and Recital 31 of the AI Regulation.

[29] See Recitals 54 and 61 of the AI Regulation.

[30] Article 10(3) of the AI Regulation.

[31] See also Recital 67 of the AI Regulation.

[32] Article 10 of the AI Regulation.

data—as an important matter of public interest[33]—to support the detection and correction of bias.[34] In addition, high-risk systems must be sufficiently accurate and technically reliable. They should be resistant to harmful or otherwise undesirable behaviours, including those arising from limitations in the operation of systems (e.g., errors, inconsistencies, or unexpected situations). Among other things, this means that an AI system is expected to minimise erroneous decisions and unfair outputs.[35]

The AI Regulation also places a strong emphasis on risk assessment and management throughout the system's lifecycle. Providers (and in many cases, deployers) are required to carry out impact assessments of their AI systems, prepare technical documentation, retain logs of the system's performance, and ensure appropriate human oversight to carry out effective post-market supervision. Clear and accessible information about how high-risk AI systems are developed and how they perform throughout their lifecycle is essential for identifying and addressing potential issues related to fairness and equal treatment.

By introducing these requirements, the AI Regulation seeks to proactively address potential discrimination before it occurs, thereby safeguarding against unfair treatment in AI-based decisions and other processes.

When it comes to international legal acts, particular attention should be paid to the **Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law**[36] (hereafter, the Convention). This legally binding convention for Estonia outlines the basic principles for the development and use of AI.

Adopted by the Committee of Ministers of the Council of Europe on 17 May 2024, the Convention aims to ensure that AI systems are fully compliant with human rights, democracy, and the rule of law throughout their lifecycle, with a strong emphasis on equality and non-discrimination. To achieve this, the Convention lays down certain general principles and

[33] As stipulated in Article 9(2)(g) of the GDPR and Article 10(2)(g) of Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC, OJ L 295, 21.11.2018, pp. 39–98.

[34] Article 10(5) of the AI Regulation.

[35] Article 15 and Recitals 74 to 75 of the AI Regulation.

[36] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Council of Europe Treaty Series, No 225, 17.5.2024.

obligations that must be observed during the design, development, and implementation of high-risk AI systems.

Compared to the AI Regulation, the Convention introduces more abstract principles and obligations under which Contracting Parties (that is, signatory countries) must guarantee that the development and provision of AI systems meets the following requirements.

First, AI system providers must implement ongoing risk management throughout the system's lifecycle. This includes assessing the potential negative impact of the system and putting in place measures to either prevent or mitigate those risks. Furthermore, the Convention stipulates that Parties should assess the need to prohibit certain applications of AI systems if they undermine respect for human rights, the functioning of democracy, or the rule of law.

In the context of non-discrimination, Article 10 of the Convention specifically mandates that Parties adopt, or maintain, measures to ensure that the lifecycle activities of AI systems respect the principles of equality, including gender equality, and non-discrimination, as laid down in applicable international and national law. The explanatory memorandum to the Convention stresses the importance of addressing bias at various stages of the AI lifecycle. For example, AI systems should not only refrain from unfavourable treatment of individuals based on protected characteristics but should also actively strive to correct structural and historical inequalities.[37]

EU member countries implement the Convention through the AI Regulation. This means that the Convention is unlikely to have a significant additional social or legal impact on Estonia compared to the impact of the already adopted EU legislation on artificial intelligence. Nonetheless, the Convention will be the first legally binding international instrument in the field of artificial intelligence. Countries that are not members of the Council of Europe, such as the United States, Canada, Japan, Israel, Mexico, Australia, Argentina, Peru, and Uruguay, among others, are also expected to accede to the Convention. The Convention could therefore have a significant impact on promoting greater and more consistent protection of equality rights in the context of AI systems.

[37] 'Explanatory Report to the Council of Europe' Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Council of Europe Treaty Series, No 225, pp. 16–17. Retrieved October 13, 2024, from https://rm.coe.int/1680afae67

4. BIAS AND DISCRIMINATION IN AI-BASED DECISION-MAKING

Processing of data, as described above, may lead to technological systems generating unfair or inaccurate outcomes, which often arise from biases in the data or algorithms.

4.1. What Is Bias?

Bias refers to assumptions in systems or data that can lead to results that are unfair or distorted. For example, if an AI model trained on previous employer data is used in recruitment, it may be biased if women and minority groups have been underrepresented in the past. Such a model may automatically favour men and historically more dominant groups. This bias in AI often reflects societal inequalities, perpetuating discrimination against certain groups based on factors such as race, gender, socioeconomic status, etc.

Specialist literature provides numerous examples.[38] AI systems using facial recognition technologies, for instance, have shown reduced accuracy when identifying individuals with darker skin tones. When the provision of a public service is linked to biometric identification, this discrepancy may lead to discrimination. Similarly, advances in AI could lead to fraud, for example, in the use of deepfake videos. Current emotion assessment systems have proven incapable of correctly reading neurodivergent[39] individuals. In financial services, bias has surfaced in loan approvals: Finnish practice is well-known for refusing to grant a loan based on nationality (Finns are considered poorer than Swedes), place of residence (a person living in a rural area is considered poorer than a person living in a city), or gender. In Germany, a woman in her forties was denied a loan under the assumption, embedded in the AI system, that she was more likely to be divorced and therefore financially less secure. A country of birth can provide a significantly higher insurance price. Also, airport verification systems discriminate against transgender, intersex, non-binary, and gender-non-conforming people because these systems do not know how to handle this data. The Chat GPT model has been trained on Wikipedia, which is widely known as a more male-based platform. Depending on the content, job postings

[38] Bartoletti, I., & Xenidis, R. (2023). Study on the Impact of Artificial Intelligence Systems, Their Potential for Promoting Equality, Including Gender Equality, and the Risks They May Cause in Relation to Non-discrimination. Council of Europe.

[39] A neurodivergent person is someone who perceives the world and processes information and whose brain functions in a way that differs from the average or normative neurological pattern, including autism spectrum disorders, attention deficit hyperactivity disorder, dyslexia, etc.

tend to be marketed to one gender, for example, job postings for drivers are shown to men and teaching roles primarily to women. In 2020, an Algorithm Watch experiment showed that Facebook, when instructed to distribute a truck driver's job advertisement in a neutral manner, that is, without targeting a specific group, it still showed the posting to 93% of men and only to 7% of women.

Bias manifests in many forms, but three main types are usually considered:

1. **Statistical bias**, which includes various errors in the collection, use, interpretation, and validation of data.
2. **Human bias**, which can occur both at the individual and at the collective or group level.
3. **Systemic bias**, which includes historical, social, and institutional factors.

Figure 2 illustrates the different forms of bias, their visibility, and complexity.

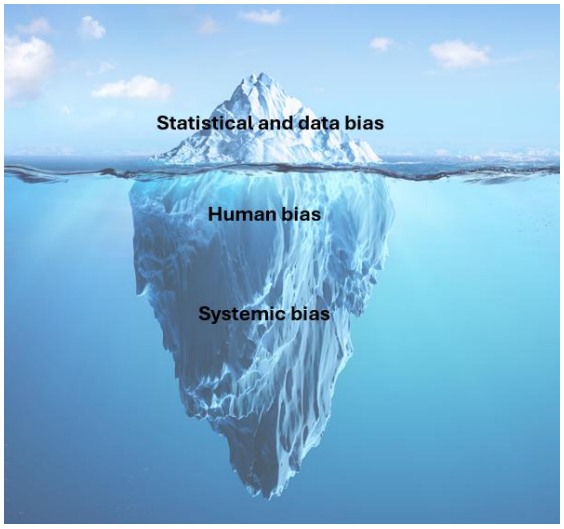

Figure 2. Types of bias[40]

In the following, we will take a closer look at the different types of bias.

Statistical bias or **data bias** occurs when a training dataset fails to accurately represent all population groups or situations. In this case, the model can learn and make predictions that favour one or more groups over others. For example, if there is more data for men and less data for women, the model may be biased towards men.

[40] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. NIST Special Publication 1270. https://doi.org/10.6028/NIST.SP.1270

A well-known case is the AI-based candidate evaluation system, which was developed by Amazon in the mid-2010s. The main reason for discontinuing its use in 2018 was that the system, trained on data from previous hiring decisions, had learned to exclude women from qualified candidates.[41]

Another example is the Philadelphia Southeastern Pennsylvania Transportation Authority, or SEPTA, AI-based surveillance system, which aim was to detect firearms in public transportation. When algorithms learn patterns of criminal behaviour from datasets that reflect trends in crime, policing, or incarceration that disproportionately affect people of colour, these algorithms can predict that people of colour are more likely to be criminals. This, in turn, may encourage the creation of racial profiles and lead to discrimination.[42]

Algorithmic bias arises when an algorithm is built in a way that favours certain outcomes that lead to bias. This may result from selecting certain features or characteristics to determine the behaviour of the model, but these characteristics do not accurately reflect reality.

Also, the use of flawed training data can lead to faulty or unfair results of the algorithms or strengthen the bias inherent in flawed data. Algorithmic bias can be caused by programming errors—for example, when developers weigh factors in algorithmic decision-making based on their own conscious or unconscious bias. An algorithm may use indicators such as income or vocabulary to inadvertently discriminate against people of a certain race or gender.

Facial recognition technologies, used, for example, to identify criminals, represent an example of algorithmic bias. These systems are trained on large datasets of photographs of individuals. Several studies have shown that some of these facial recognition algorithms are more likely to consider dark-skinned people to be criminals. Such algorithmic bias errors derive from the fact that the amount of data used to train the algorithm contains a disproportionate number of photos of light-coloured people, and thus the algorithm performs better on the faces of light-coloured people compared to those of darker-skinned people.

[41] BBC (2018). Amazon Scrapped 'Sexist AI' Tool, October 10. https://www.bbc.com/news/technology-45809919

[42] Torrejón, R. (2024). SEPTA's 'Archaic' Security Cameras Foil Plans for a Program That Would Use AI to Detect Guns. TransitTalent, March 20. https://www.transittalent.com/articles/index.cfm?story=SEPTA_Cameras_Foil_AI_Surveillance_3-20-2024

In 2019, the American Civil Liberties Union (ACLU) published a study in which they tested Amazon's facial recognition technology Rekognition[43] by comparing photographs of members of the US Congress with those of criminals. As a result, 28 members of Congress were misidentified as criminals, including several members of Congress of African American and Hispanic ethnicity. For members with white skin colour, the facial recognition system was almost 100% accurate. Black members of Congress of Hispanic origin had an error rate of more than 40%. This case suggests that algorithms designed for facial recognition do not work equally for all skin colours. Following the study, the ACLU issued a press release calling on Amazon to stop selling its facial recognition technology to law enforcement, citing a serious racial bias issue and potential human rights violations. This case clearly shows how algorithms can be unfair and biased when trained on inconsistent datasets.

Systemic bias arises from the procedures and practices of specific institutions that privilege certain social groups over others. It does not need to be a matter of deliberate bias or discrimination, but of the majority following the established rules or norms. Common examples of systemic bias include racism and gender.

For example, in 2020, the Dutch government used the system designed to detect fraud in claims for child benefits.[44] Unfortunately, the data in the system contained errors and assumptions that led to the false suspicion of thousands of families. Many families were unfairly accused of fraud, which resulted in confusion, reputational damage, and financial harm. The Dutch government acknowledged that there were shortcomings in the system and issued an apology to the affected families. An independent investigation was launched and changes were made to the functioning of the system and the processing of data to avoid possible errors in the future.

Historical bias emerges when using historical data that incorporates past decisions and actions that may contain injustice or prejudices. For example, if certain social groups, such as women, have been disadvantaged in the past, it is reflected in historical data and may also lead to bias in new predictions.

[43] Queram, K. E. (2019). Face-Recognition Tool Misidentified State Lawmakers as Criminals: ACLU. Defense One, August 17. https://www.defenseone.com/technology/2019/08/face-recognition-tool-misidentified-state-lawmakers-criminals-aclu/159190

[44] Wikipedia (n.d.). Dutch Childcare Benefits Scandal. Retrieved October 13, 2024, from https://en.wikipedia.org/wiki/Dutch_childcare_benefits_scandal?utm_source

Although such a classification may not always be practical, it helps us understand the origins of biases. Ultimately, whether through the use of limited or incomplete data or the algorithm designer's own bias, AI-based systems risk giving rise to discrimination based on gender, race, skin colour, or other characteristics.[45]

[45] Chen, Z. (2023). Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices. Humanities and Social Sciences Communications, 10, Article 567. https://doi.org/10.1057/s41599-023-02079-x

4.2. What Is Discrimination?

In the most general sense, discrimination is a situation where a person or a group of persons are treated less favourably than other persons or groups on a basis of a characteristic protected under the relevant law.

In Estonia, in addition to the general prohibition of discrimination provided for in § 12 of the Constitution, discrimination is regulated by the Equal Treatment Act[46] and, more importantly, by the Gender Equality Act.[47]

According to these laws, as shown in Figure 3, discrimination can be divided into two categories: direct discrimination and indirect discrimination.

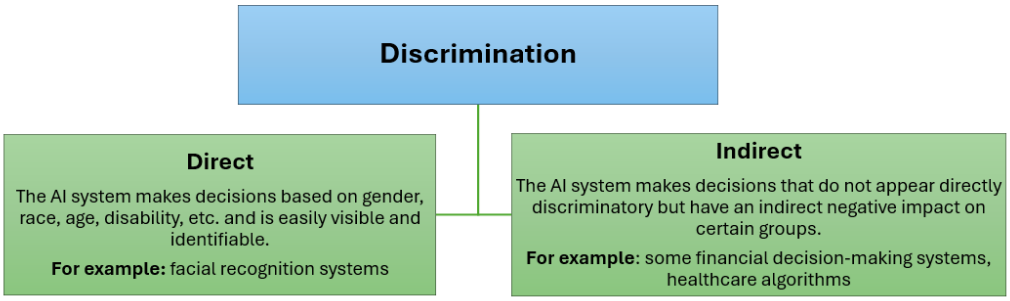

Figure 3. Discrimination and its different forms

In the case of direct discrimination,[48] one person is treated less favourably than another person has been, is, or may be treated in a similar situation based on a protected characteristic provided by law.

Indirect discrimination occurs when an apparently neutral provision, criterion, or practice places a person in a disadvantageous position compared to others based on a characteristic

[46] Equal Treatment Act. RT I 2008, 56, 315, 1.1.2009. https://www.riigiteataja.ee/akt/122102021011

[47] Gender Equality Act. RT I 2004, 27, 181, 1.5.2004. https://www.riigiteataja.ee/akt/130062023072

[48] In the Equal Treatment Act, these characteristics are nationality, race, skin colour, religion and belief, age, disability, sexual orientation and in the Gender Equality Act—gender. However, in an employment relationship the range of characteristics is slightly broader (see § 2(3) of the Equal Treatment Act).

specified in these Acts, unless the provision, criterion, or practice pursues an objective legitimate aim and the means of achieving that aim is appropriate and necessary.

When it comes to discrimination in AI-based decision-making, both types of discrimination are possible, as the above examples demonstrate. Yet, it is found that indirect discrimination tends to be prevalent in the context of AI. This is because developers avoid entering protected identifiers for training data in AI systems.

A related concept that is often mentioned is "proxy discrimination." Arguably the most common form of discrimination in AI systems, proxy discrimination emerges when discrimination is related to a characteristic which is not directly mentioned in the law, but which often leads to discrimination. Though similar to indirect discrimination, it is usually considered a distinct phenomenon in the context of AI. For example, a person's zip code (indicates poverty in the region), occupation (indicates financial capacity), first name (indicates age), employment history in terms of length or interruptions (women tend to have a shorter period of employment or gaps in work due to parental leave or other family responsibilities) can push an algorithm to form a discriminatory conclusion.

Bias can occur already in the early stages of data collection and preparation. All systems that rely on data analysis only receive the data on which they have been trained to draw their own conclusions. However, bias can also occur in the process of training the model, depending on the algorithms and training processes chosen. For example, when only specific characteristics or traits are selected that determine predictions or decisions, these characteristics may affect the outcomes. Also, if these characteristics are not equal for all groups, this can lead to a biased conclusion. Consider a loan approval model that only takes into account income and academic achievement: such a model may favour individuals with higher incomes and exclude those with a lower level of education. If unreliable data is used in data training, models may learn that some characteristics (such as gender or skin colour) are relevant to decision-making, even if this is not the case. However, such a situation may lead to a discriminatory decision. Bias can also manifest in testing if there is biased selection of testing data or if the testing data does not include data from all groups. Also, bias can emerge during the implementation phase of the system, when the model is applied in real situations and users start making decisions based on it, but the operation principles of the model itself are not fully transparent.

Bias, and therefore discrimination, in AI-based decision-making is difficult to detect, as these systems are often too complex for users to understand how conclusions are reached. Also, an AI-based model does not explain why it made this decision. When countless different decisions feed into making a decision, even a single biased decision can lead to a discriminatory result. It is also worth noting that focusing solely on the data input and output what data is entered and what data comes out ignores the reason or wider context of why the data was originally entered. Individuals who use AI-based systems in their decision-making trust AI too much and rarely question whether an AI system could, in fact, be wrong.

It is also important to know that bias is not only created by AI alone. Persons who design and use the AI are also important, as they guide the selection of data and the training process. It can be concluded that the emergence of bias in an AI-based system is not merely a technical issue, but a sociotechnical one, as the data is drawn from the real world and if there is bias, this bias also ends up in AI-based decisions.

The quality of an AI-based decision depends on the availability of good quality data. In this context, data quality entails both the quantity and quality of data. Furthermore, trustworthy and responsible AI does not only mean that the AI system is unbiased, fair and ethical, but also that the system does what it was designed to do.

State bodies have different competences in dealing with cases of discrimination. For example, the Gender Equality and Equal Treatment Commissioner can process cases on the basis of the characteristics provided for in the Gender Equality Act and the Equal Treatment Act. In other situations, the cases are processed by the Chancellor of Justice, though typically with certain limitations, such as reconciling the parties in the conflict. In addition, the Chancellor of Justice has the right, within her competence, to make assessments in the context of general discrimination provided for in § 12 of the Constitution. Notably, § 12 of the Constitution provides an open list of protected characteristics, while the lists of characteristics of the Equal Treatment Act and the Gender Equality Acts are closed. It is also important to note that the Commissioner’s competence is limited to assessing whether discrimination may have occurred in a given situation, and that discrimination disputes are handled by courts and labour dispute committees.

In the context of AI-based discrimination, the boundaries between direct and indirect discrimination are blurred, which makes it more difficult to detect discrimination as well as to

delimit and process the related liability. Since the responsibility of handling discrimination cases is divided between different bodies depending on the protected characteristics and the competence of these bodies, there is still little clarity on how AI-based discrimination cases should be processed in practice. Article 77 of the EU AI Regulation provides for the supervision of mutual assistance, market surveillance, and artificial intelligence systems, according to which a Member State must designate the authorities responsible for overseeing compliance with the requirements in the Regulation. What this should look like in practice, including cooperation with other state authorities, primarily the Gender Equality and Equal Treatment Commissioner, remains unclear. Furthermore, there is the issue of intersectional discrimination, where discrimination is based on several different characteristics, whereas these characteristics fall under the competence of different authorities.

5. WHO IS RESPONSIBLE?

In legal terms, it is always important to establish who is responsible when a violation of the law occurs. Who, for example, should be held accountable if it turns out that discrimination has arisen from an AI-based decision? The opacity and complexity of AI systems makes it generally very difficult to attribute responsibility and, therefore, create appropriate liability rules. The decisions of these systems are technically based on the interaction of different components of hardware and software and may be constantly changing. Rarely are these systems developed by the users themselves; more often they are commissioned. This leads to a further question: which stage of data processing should the responsibility be tied to? In other words, at what point exactly in the AI-based decision-making does the element that causes discrimination emerge?

Therefore, the levels of responsibility in relation to AI systems can be broadly distinguished as follows:

- **Creators and developers of AI systems** are responsible for ensuring that the AI system is free from bias, fair, and transparent.
  Example: If a facial recognition algorithm is biased towards dark-skinned people, the developers are responsible for correcting this bias.
- **Implementers of AI systems** (companies, institutions, and organisations) are responsible for the lawful and non-discriminatory use of AI systems.
  Example: If a credit-scoring algorithm discriminates against certain ethnic groups, the lender is obliged to re-evaluate and improve the system.
- **Legislators** are responsible for establishing ethical and fair regulations and implementing appropriate oversight measures.
  Example: The AI Regulation establishes clear rules and requirements for high-risk AI systems, including obligations of transparency and accountability for AI decisions.
- **End-users** who make decisions based on AI-based systems (judges, HR managers, and loan decision-makers) are responsible for interpreting and implementing AI decisions in as objective manner as possible.

6. PREVENTING BIAS AND DISCRIMINATION IN AI-BASED DECISIONS

The first step in identifying and addressing AI bias is conscious management of its implementation. It is important to ensure that the organisation has the capacity to manage and oversee AI activities. The following measures are central to reducing the risk of bias and discrimination in AI-based decisions:

1. **Ensuring data balance and representation:** Training data should be representative of all groups. If representation of certain groups is inadequate, the system must be designed to avoid reliance on such features or ensure that the data is balanced.
2. **Using evidence-based methods and ensuring transparency:** The AI systems used must be verified and tested to detect and prevent bias. Tools should be used that can determine whether an AI system discriminates against a group or part of it.
3. **Applying ethical guidelines and regulations:** The development of AI technology must adhere to ethical principles and applicable legal regulations to ensure that systems do not make discriminatory or biased decisions.
4. **Monitoring and providing feedback on AI systems:** AI systems and models require continual oversight and updating to prevent them from generating new biased and discriminatory results. Oversight plays an important role in this process.

With these principles in mind, it is necessary to understand who holds responsibility and what must be done to prevent bias and discrimination.

The state must raise awareness among its residents and decision-makers who rely on algorithm-based tools that AI poses risk of discrimination. People need to be informed about what it means and how to recognise and prevent such discrimination. In the Estonian context, this responsibility would be supported by legal provisions, among other things. For example, there should be a legal obligation to inform individuals against whom an algorithmic decision is made that AI has been used in the decision and the extent of its influence on the decision. Also, individuals should be made aware of appeal procedures and who to turn to if they suspect that discrimination against them has taken place. Developers must follow guidelines established or recommended by the state throughout the creation and development process of the respective application, to identify and avoid bias at various stages of individual AI-based decision-making.

Both the GDPR and the AI Regulation have established certain principles for human control of AI systems. Article 22 of the GDPR refers to “human intervention” in automated decision-making, while Article 14 of the AI Regulation provides for “human oversight.” Such human oversight can take various forms, in particular system control (which involves design, implementation, testing, and auditing of the system) and individual control (which involves oversight of a specific decision).

In addition to what has been explained above, the Regulation also lays down several general obligations for high-risk systems under Articles 9, 10, 11, 13, and 15. However, high-risk schemes remain narrowly defined. In practice, there are still no clear agreements for associating “high risk” with public services, on the one hand, and possible discrimination by AI, on the other.

The table below shows the risk levels under the EU AI Regulation, their descriptions, and the main requirements.

Table 1. Risk levels of the EU AI Regulation

Level of risk	Description	Requirements	Examples
Minimal	AI systems that present minimal or no risk to people's rights, health, or safety.	The AI Act does not introduce specific rules for AI that is considered minimal or no risk.	Applications such as AI-enabled video games or spam filters, product recommendation engines.
Limited	AI systems that pose a low risk to fundamental rights or safety but still require some transparency.	Users must be informed that they are interacting with an AI system (e.g., chatbots) and when content is AI-generated. Deepfakes and AI-generated content intended to inform the public must be clearly labelled as such.	Chatbots (e.g., customer support assistants on websites), virtual agents (e.g., booking assistants).
High	AI systems that pose a significant risk to people's health, safety, or fundamental rights.	These systems are subject to strict obligations before they can be introduced on the market. The requirements include risk management and mitigation, quality data to prevent bias, activity logging for traceability, detailed documentation, clear user information, human oversight, robustness, security, and accuracy.	AI for recruitment and hiring decisions, credit scoring and loan approval systems, biometric identification and verification systems (e.g., facial recognition).
Unacceptable	AI systems that pose a clear threat to people's safety, rights, or fundamental freedoms.	The use of such systems is prohibited under the EU AI Regulation.	Mass implementation of facial recognition systems by law enforcement, social scoring systems, and manipulative AI.

On the one hand, it is argued that the AI systems covered in Annex III of the Regulation concern public services and are therefore subject to the same principles as public services. On the other hand, doubts have been raised as to whether this is really the case. In the context of discrimination, the question arises as to which rules apply to medium- or low-risk systems. Proceeding from the general principle that any AI system that makes or supports decision-making related to the rights of a natural person always falls into the category of high-risk systems, some authors have turned to the definition of "profiling" in Article 4(4) of the GDPR

and interpret it in a way that public services ought to be excluded from the list of high-risk systems[49]

Yet, the AI Regulation imposes important obligations on organisations and bodies that must ensure the "human oversight" provided for in the Regulation. Broadly speaking, this entails the following requirements:

- Those responsible need to understand how a high-risk AI system operates and be able to react accordingly when a problem occurs;
- They need to be aware that an AI system can lead to "automated biases," especially when it comes to decisions made against natural persons;
- They must be prepared to quickly discontinue the use of an AI system in case of doubt.

The EquiTech project will develop an Impact Assessment Toolbox and Guidelines. The Guidelines are a document that provides a step-by-step guide on how to use the Impact Assessment Checklist, explaining the evaluation process and ways to avoid discrimination when developing and using AI-based solutions.

[49] Defender of Rights (2024). Report: Algorithms, AI Systems and Public Services: What Rights Do Users Have?, 20.

7. PREVENTING BIAS AND DISCRIMINATION ON THE EXAMPLE OF E-SERVICE BÜROKRATT?

7.1. What Is Bürokratt?

Bürokratt[50] is a network of chatbots that operates on the websites of public sector institutions and is designed to help citizens obtain information and access various public and informational services through virtual assistants. The Estonian e-service allows users to receive information from institutions and access basic services via a chat window. In essence, it functions as an interoperable network of public sector AI systems (called *kratt*s in Estonian), integrated with state information systems. From the users' perspective, it serves as a single channel for direct access to public and information services. The initiative was launched by the Ministry of Justice and Digital Affairs, and the Information System Authority is responsible for its technical implementation.

7.2. The AI Models Used in Bürokratt

The network relies mainly on text-based technologies, including:

- Machine learning to interpret users' questions and generate accurate responses;
- Anonymisation tools that remove personal data from texts;
- Automatic translation for providing services to non-native speakers;
- Text classification, which helps to route queries to the correct institution (e.g., the pilot solution Siimuke).

7.3. How Does Bürokratt Prevent Bias?

Eliminating bias when developing technical solutions is critical to ensuring the fairness and reliability of the system for all users. To avoid bias, several strategies can be employed:

- The data used to train the Bürokratt machine-learning models are carefully selected to be diverse and representative of all the needs of user groups.

[50] More information about the Bürokratt initiative is available in Estonian at https://www.kratid.ee/burokratt and https://www.ria.ee/riigi-infosusteem/personaalriik/buro-kratt#teadvustamine. See also the introductory video on Bürokratt at https://youtu.be/SoiLEDXZvi4?list=TLGGxGKZUh0EAl8yMDA0MjAyNQ

- The Bürokratt machine-learning models are constantly evolving and continuously improved through feedback.
- Bürokratt uses a variety of methods to analyse and identify potential signs of bias in its machine-learning models.
- Throughout the development process of Bürokratt, the transparency and verifiability the algorithms are prioritised.

7.4. The Future of Bürokratt

Looking ahead, the plan is to develop Bürokratt towards a system where communication with the state becomes fully human-language-based, accessible and seamless, regardless of the device, communication channel, or language used. Bürokratt should begin to understand spoken language and respond using synthetic speech, allowing communication through voice commands instead of writing and listening. Also, the availability of services is planned to be improved for foreign-language residents by adopting automatic translation, which allows communication in Estonian, English, Russian, and Ukrainian. In the future, Bürokratt will also allow user requests to be automatically forwarded to the relevant state authority, sparing citizens the need to know where to turn. In addition, the possibilities of sign-language recognition and machine-assisted interpretation will be explored to make the services accessible to people with hearing impairment.

8. FURTHER READING

1. Bostrom, N. (2016). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
2. Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Allen Lane.
3. White Paper on Data and Artificial Intelligence 2024–2030. https://www.mkm.ee/media/10154/download?ref=tehisintellekt.co
4. Estonian Open Data Information Gateway. https://avaandmed.eesti.ee/
5. Bartoletti, I., & Xenidis, R. (2023). Study on the Impact of Artificial Intelligence Systems, Their Potential for Promoting Equality, Including Gender Equality, and the Risks They May Cause in Relation to Non-Discrimination. Council of Europe.
6. Wulf, J. (2022). Automated Decision-Making Systems and Discrimination. Understanding Causes, Recognizing Cases, Supporting Those Affected. A Guidebook for Anti-Discrimination Counselling. https://algorithmwatch.org/en/authocheck/
7. Orwat, C. (2020). Risks of Discrimination through the Use of Algorithms. Institute for Technology Assessment and Systems Analysis (ITAS). Karlsruhe Institute of Technology (KIT). https://www.antidiskriminierungsstelle.de/EN/homepage/-_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1
8. Centre for Data Ethics and Innovation (2020). Review into Bias in Algorithmic Decision-Making. https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf
9. Loi, M., Mätzener, A., Müller, A., & Spielkamp, M. (2021). Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities. Algorithm Watch. https://algorithmwatch.ch/en/wp-content/uploads/2022-/09/2021_AW_Decision_Public_Sector_EN_v5.pdf
10. Bullock, J. B., Chen, Y.-C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. M., & Zhang, B. (2024). The Oxford Handbook of AI Governance. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.001.0001
11. Paul, R., Carmel, E., & Cobbe, J. (Eds.). (2024). Handbook on Public Policy and Artificial Intelligence. Edward Elgar Publishing. https://doi.org/10.4337/9781803922171

12. Acemoglu, D., & Johnson, S. (2024). Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. PublicAffairs.
13. Coeckelbergh, M. (2022). The Political Philosophy of AI: An Introduction. Polity.
14. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group.
15. Madan, R., & Ashok, M. (2023). AI Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda. Government Information Quarterly, 40(1), Article 101774. https://doi.org/10.1016/j.giq.2022.101774
16. Mergel, I., Dickinson, H., Stenvall, J., & Gasco, M. (2023). Implementing AI in the Public Sector. Public Management Review, 1–14. https://doi.org/10.1080/14719037.2023.2231950
17. Agarwal, P. K. (2018). Public Administration Challenges in the World of AI and Bots. Public Administration Review, 78(6), 917–921. https://doi.org/10.1111/puar.12979
18. Radu, R. (2021). Steering the Governance of Artificial Intelligence: National Strategies in Perspective. Policy and Society, 40(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728
19. Taeihagh, A. (2021). Governance of Artificial Intelligence. Policy and Society, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377
20. Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the Use of Artificial Intelligence in Public Governance: A Systematic Literature Review and a Research Agenda. Government Information Quarterly, 38(3), 101577. https://doi.org/10.1016/j.giq.2021.101577
21. Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019), Hello, World: Artificial Intelligence and Its Use in the Public Sector. OECD Working Papers on Public Governance, No 36. OECD Publishing. https://doi.org/10.1787/726fd39d-en
22. UK Government (2019). A Guide to Using Artificial Intelligence in the Public Sector. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector
23. Rodrigues, R. (2020). Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities. Journal of Responsible Technology, 4, Article 100005. https://doi.org/10.1016/j.jrt.2020.100005

24. Chen, Y.-C., Ahn, M. J., & Wang, Y.-F. (2023). Artificial Intelligence and Public Values: Value Impacts and Governance in the Public Sector. Sustainability, 15(6), Article 6. https://doi.org/10.3390/su15064796
25. Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. International Journal of Public Administration, 42(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103