Funded by
the European Union

EDUCATIONAL MATERIAL FOR PUBLIC SERVANTS

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 public servants to understand the complexities and potential risks associated with AI system biases and discrimination. Given the target audience, it focuses on the operation of AI-based systems, and to a lesser extent on legal principles.

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.

.

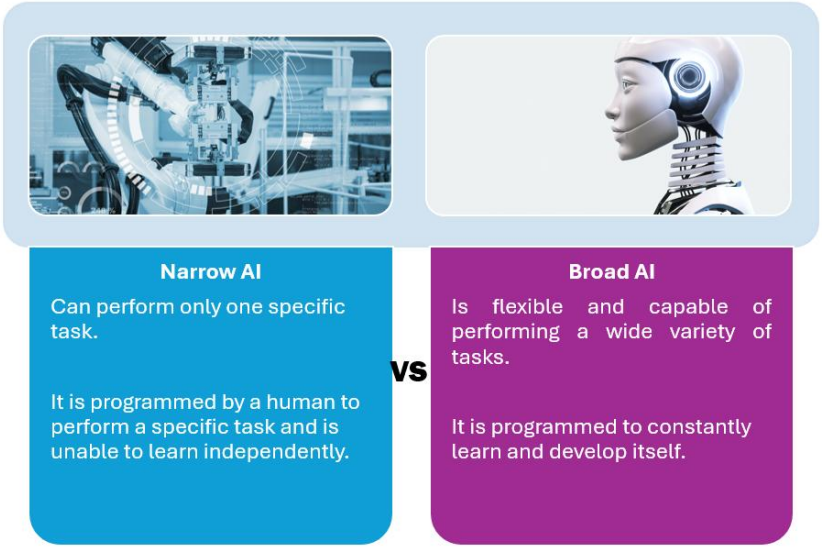

Figure 1. Broad vs. Narrow AI

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

[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.

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 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.

As shown in Figure 2, AI consists of several different layers which will be examined in further detail below.

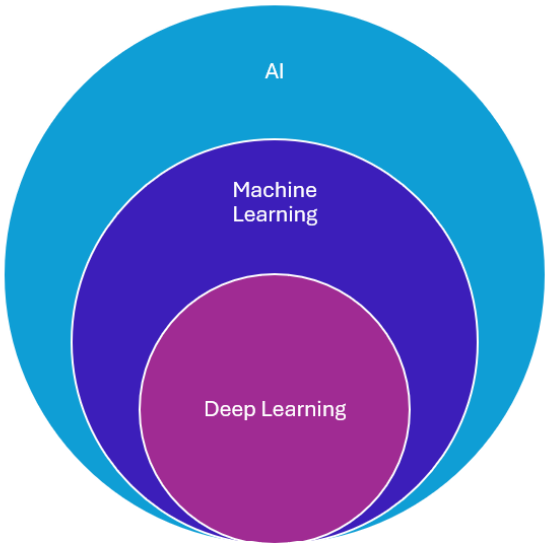

Figure 2. Layers of AI

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

[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

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*, which is 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 an AI to play games).

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.

An algorithm-based decision-making system can either replace or support human decision-making. A decision is considered algorithm-based when it is determined or significantly influenced by one or more algorithms, which may include traditional rule-based algorithms or self-learning algorithms such as machine learning models.

3. AI IN OUR DAILY LIVES

Artificial intelligence has become part of our daily lives, often so seamlessly that we may not even notice its presence. For example, in smartphones, virtual assistants use AI to interpret spoken language, place calls, send messages, or use an image to detect photos, identify people, places, or objects. Social media platforms track and analyse user behaviour, provide personalised advertisements and posts. Chatbots automate customer support services.

In healthcare systems, AI assists doctors in diagnosing diseases by analysing patients' symptoms, medical history, and test results. When it comes to interpreting X-rays, blood tests, and other medical images, AI has become invaluable. Banks use AI-based systems to detect suspicious transactions and prevent fraud, for example, based on transaction patterns. In finance, AI provides customers with personalised investment recommendations or assesses their creditworthiness in loan applications. Many security systems use AI to identify people or detect unusual behaviour, allowing quicker response and prevention of problems. Figure 3 highlights some of the most important areas of use of AI in our everyday lives.

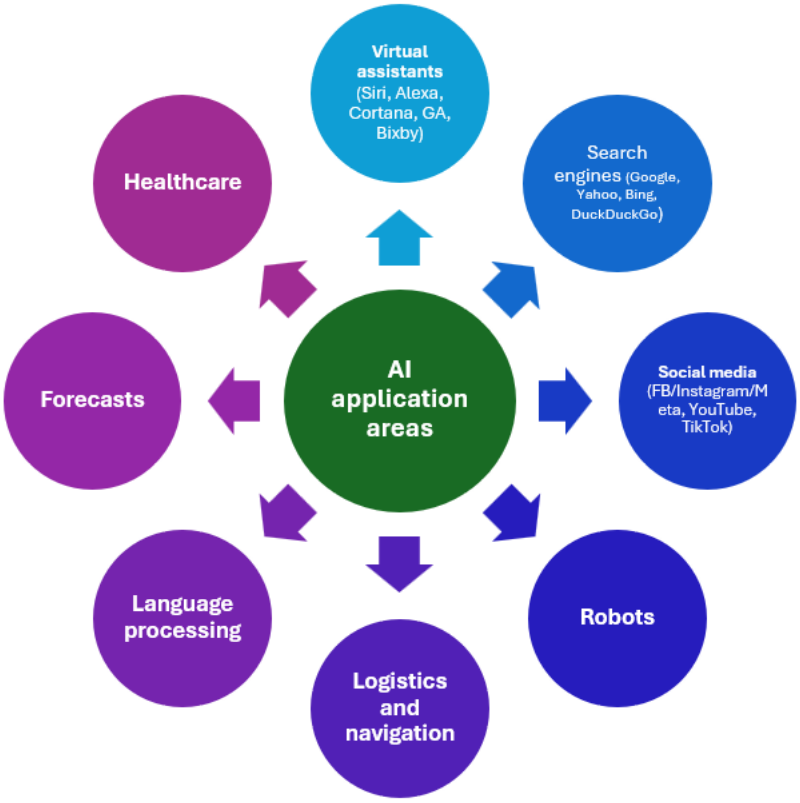

Figure 3. Areas of use of AI

The added value and quality of AI-based systems have grown significantly in recent years. While a few years ago, interactions with a virtual assistant were awkward enough to prompt a smile or make one contact a human customer service representative, today, in many contexts, we may not even be able to distinguish whether something has been generated by AI or is man-made.

To illustrate this development, consider, for example, generative image technologies. Between February 2022 and December 2023, the Midjourney AI model generated increasingly hyper-realistic images of Harry Potter, the famous movie character. The figure below clearly shows a dramatic leap in development over a period of just two years.

Figure 4. Evolution of hyper-realistic images generated by Midjourney[4]

INTERESTING FACT

Current AI solutions for image generation are already so advanced that distinguishing between AI-generated images from real photos can prove difficult.[5] For example, can you tell which of these two images is real?[6]

Those who guessed that the photo of the man was real are correct. If you wish to test your image recognition ability even further, you can explore The Which Face Is Real website at https://www.whichfaceisreal.com/index.php.

[4] Stanford University (2024). Artificial Intelligence Index Report 2024. Human-Centred Artificial Intelligence. https://hai-production.s3.amazonaws.com/files/hai_ai-index-report-2024-smaller2.pdf

[5] Stanford University (2024). Artificial Intelligence Index Report 2024. Human-Centred Artificial Intelligence. https://hai-production.s3.amazonaws.com/files/hai_ai-index-report-2024-smaller2.pdf

[6] Which Face Is Real? (2025). https://www.whichfaceisreal.com/index.php

4. HOW DOES AI WORK?

AI operates by using algorithms to process data, learn patterns, make predictions or decisions, while constantly evolving. It simulates human learning and decision-making processes, but does it at much faster speed and larger scale, which makes it useful for a wide range of applications, including personalised recommendations.

A data-driven model is used in machine learning and it makes predictions or decisions directly based on data rather than rules or predefined logic. In such a model, training and analysis is based on data that contains patterns and relations, and the model uses these patterns to inform subsequent decisions.

4.1. How Does a Data-Driven Model Work?

1. **Data collection**. A data-driven model collects data relevant to the problem that is being solved. The data can be numerical, textual, or visual.
2. **Data analysis and processing**. After the data is collected, it must be cleaned and systematised so that it can be used to train the model. This may involve filling in missing elements, standardising the data, or defining separate categories.
3. **Model selection and training**. Once the data is ready for use, an appropriate AI model is chosen for training. Depending on the task, the model can be based on the abovementioned machine-learning models (supervised or unsupervised learning models or a deep learning-based model).
4. **Learning/Training**. Typically, the dataset is split into two: training data and test data. The chosen model analyses patterns and relationships in the data, moving towards increasingly complex tasks. For example, it might learn to predict whether a customer will buy a product or not based on an analysis of their past purchases and behaviour. The time it takes to train one AI model depends on the chosen model and the amount of data collected.
5. **Model testing and prediction**. The trained model is then tested on new data. If it provides reliable predictions or decisions, the model can be applied in the real world. If not, the model needs to be improved and adjusted.
6. **Model optimisation.** Sometimes the test results of the chosen AI model are inadequate—whether due to poor data quality or the fact that the chosen AI model is too

unrefined to be able to identify the desired associations. Testing may also detect subjective decisions or predictions. In such cases, the AI model needs to be upgraded to achieve the expected result.

7. **Implementation.** If the AI model produces the desired results, it can be put into practice. Implementation involves making it available to the public, integrating it with existing tools, or developing hardware that uses the relevant AI model.
8. **Continuous learning.** The whole process does not end with the implementation of the AI model, because training an AI model is not a one-time activity. To keep the model up to date, it must be regularly trained on fresh data or according to feedback from users.

To put it simply, the first thing that one needs is existing data and the desire to process the data. Then, an AI model suitable for data processing is chosen to train and test it on the data. Testing reveals whether additions or improvements need to be made to the existing model to be implemented in practice. However, it is important that regardless of the introduction of the model, the learning process continues. The following figure describes the main steps in the operation of AI.

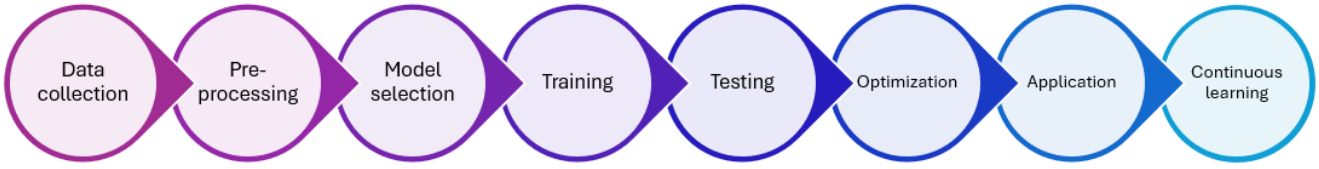

Figure 5. The main stages in the operation of AI

As can be seen from the above, the availability of high-quality data is fundamental to the successful training of an AI model. For this reason, we shall take a closer look at the topic. Within the data-driven model, the terms “Training Dataset” and “Testing Dataset” are used. Figure 6 describes in more detail how datasets suitable for AI training are divided.

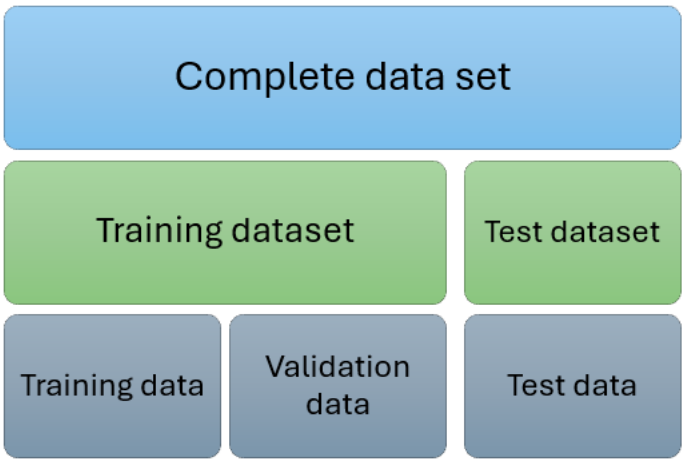

Figure 6. Datasets in AI training

A **training dataset** is a dataset that is used to train a model. Using this dataset, the machine-learning model learns to identify patterns and relationships in the data. The model analyses this data to understand how different traits or factors affect outcomes (or targets). Training data is the most important part of the model's training process, because it forms the foundation of the model's "learning." For example, if we train a model that predicts whether a person is approved for a loan, the training dataset might contain information about the person's age, income, credit history, and previous loans. Based on this data, the model learns how these characteristics are related to the probability of being approved for a loan. For training to be effective, the dataset must be large enough for the model and sufficiently diverse to avoid subjectivity and discrimination.

A **validation dataset** is separated from the training dataset. It allows the model's performance to be evaluated during training and helps adjust parameters such as learning rate. Although the validation dataset is usually small, it most accurately represents the data relevant to the problem to be solved. The use of validation data ensures that the AI model does not just memorise the training data but is also able to learn to generalise on the basis of the data.

A **test dataset** is a dataset that is used to evaluate a model after training against a training dataset. The test dataset contains data that the model has not seen before, and therefore the testing dataset is kept separate from the training dataset. The purpose of the test data is to check how well the model can generalise self-learning by making predictions based on new, previously unseen data. The test dataset helps assess the model's accuracy, its ability to generalise and make predictions that may not be represented in the training data. Returning to our loan prediction example, a test dataset might contain data about other people whose

characteristics are similar to those in the training dataset, but the model has not seen this data before. The testing data is used to assess how well the model performs in predicting loan approval.

As for the ideal proportion of different datasets in the AI teaching process, tentatively 70% of the data could be training data, 15% validation data, and 15% testing data. However, this approach is indicative, and the ratios may vary depending on the training data and the specific model in use.

5. 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.

5.1. 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.[7] 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[8] 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 tend to be marketed to one gender, for example, job postings for drivers are shown to men and

[7] 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.

[8] 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.

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

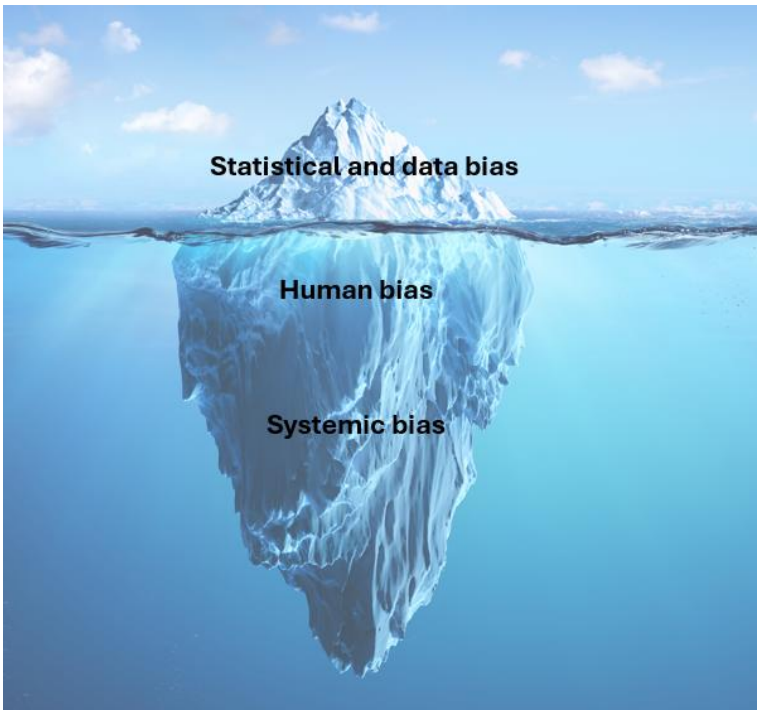

Figure 7. Different types of bias[9]

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

[9] 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

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.

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.[10]

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.[11]

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

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

[11] 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

of light-coloured people, and thus the algorithm performs better on the faces of light-coloured people compared to those of darker-skinned people.

In 2019, the American Civil Liberties Union (ACLU) published a study in which they tested Amazon's facial recognition technology Rekognition[12] 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.[13] 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,

[12] 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

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

have been disadvantaged in the past, it is reflected in historical data and may also lead to bias in new predictions.

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.[14]

5.2. What Is Discrimination?

Having discussed the various forms of bias, it is also important to understand the nature of 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[15] and. more importantly, by the Gender Equality Act.[16]

According to these laws, as shown in Figure 8, discrimination can be divided into two categories: direct discrimination and indirect discrimination.

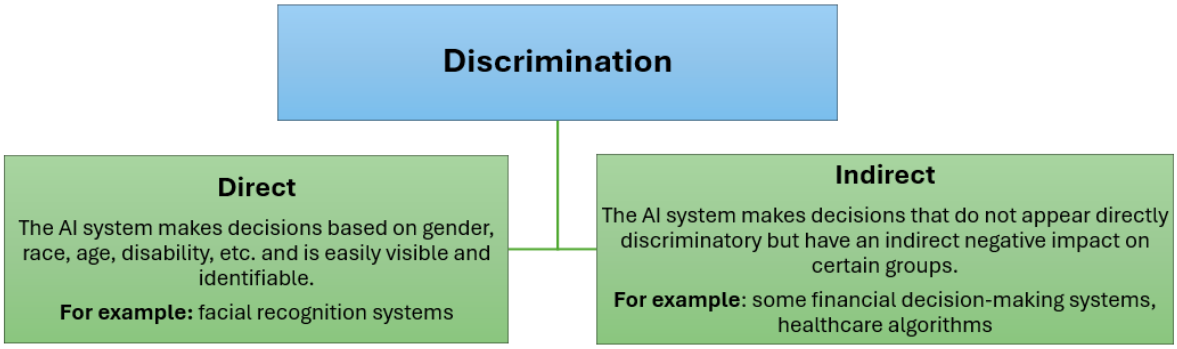

Figure 8. Discrimination and its different forms

[14] 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
[15] Equal Treatment Act. RT I 2008, 56, 315, 1.1.2009. https://www.riigiteataja.ee/akt/122102021011
[16] Gender Equality Act. RT I 2004, 27, 181, 1.5.2004. https://www.riigiteataja.ee/akt/130062023072

In the case of direct discrimination,[17] 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 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 examples above 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 that 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

[17] 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 wider (see § 2(3) of the Equal Treatment Act).

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 result. 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 on 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.

6. 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.

7. HOW TO PREVENT BIAS AND DISCRIMINATION IN ALGORITHMIC 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 different 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	Level 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 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[18]

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.

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

8. WHAT IS BÜROKRATT?

Bürokratt[19] 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 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.

8.1. 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).

8.2. 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.

[19] More information about the Bürokratt initiative is available in Estonian at https://www.kratid.ee/burokratt and https://www.ria.ee/riigi-infosusteem/personaalriik/burokratt#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.

8.3. 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 users 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.

9. 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